Machine-learning assisted development of a knowledge-based system in dairy farming

by

Diederik Pietersma

Department of Animal Science
McGill University, Montreal

May, 2001

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy

© Diederik Pietersma 2001

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-78751-6

Canada

Abstract

The goal of this research was to explore the use of machine learning to assist in the development of knowledge-based systems (KBS) in dairy farming. A framework was first developed which described the various types of management and control activities in dairy farming and the types of information flows among these activities. This framework provided a basis for the creation of computerized information systems and helped to identify the analysis of group-average lactation curves as a promising area of application. A case-acquisition and decision-support system was developed to assist a domain specialist in generating example cases for machine learning. The specialist classified data from 33 herds enrolled with the Québec dairy herd analysis service, resulting in 1428 lactations and 7684 tests of individual cows, classified as outlier or non-outlier, and 99 interpretations of group-average lactation curves. To enable the performance analysis of classifiers, generated with machine learning from these small data sets, a method was established involving cross-validation runs, relative operating characteristic curves, and analysis of variance. In experiments to filter lactations and tests, classification performance was significantly affected by preprocessing of examples, creation of additional attributes, choice of machine-learning algorithm, and algorithm configuration. For the filtering of individual tests, naïve-Bayes classification showed significantly better performance than decision-tree induction. However, the specialist considered the decision trees as more transparent than the knowledge generated with naïve Bayes. The creation of a series of three classifiers with increased sensitivity at the expense of reduced specificity per classification task, allows users of a final KBS to choose the desired tendency of classifying new cases as abnormal. For the main interpretation tasks, satisfactory performance was achieved. For the filtering tasks, performance was fairly poor since a reasonable sensitivity was associated with many false positives. The specialist evaluated the learned knowledge and suggested small modifications to improve several classifiers. Machine-learning assisted knowledge acquisition proved to be a feasible approach to support the development of a KBS in dairy farming. This approach is expected to be especially useful for areas where specialists have difficulty expressing decision rules, such as tactical-level management and analysis of graphical information.

Résumé

L'objectif de cette recherche était d'explorer l'utilisation de l'apprentissage-machine pour assister le développement de logiciels cognitifs en production laitière. La première étape consistait en le développement d'un cadre conceptuel qui décrivait les diverses activités de contrôle et de régie en production laitière, ainsi que les types d'information qui y sont échangés. Le cadre devait servir de base pour la création subséquente de systèmes d'information informatisés. Il a facilité l'identification de l'analyse de courbes de lactation par groupes de vaches comme étant un domaine d'application d'intérêt. Un système d'acquisition de cas et d'aide à la décision a été développé pour assister un spécialiste du domaine à générer des exemples de cas pour l'apprentissage-machine. Un spécialiste a analysé des données provenant de 33 troupeaux inscrits au contrôle laitier québécois, résultant en 1428 lactations et 7684 données au jour du test classifiées comme normales ou hors-limites, et 99 interprétations de courbes de lactation moyennes de groupes de vaches. Pour faciliter l'analyse de performance de classificateurs générés par apprentissage-machine à partir de ces groupes restreints des données, une méthode a été établie qui impliquait la validation croisée, les courbes de caractéristiques opérationnelles relatives, et l'analyse de variance. Lors d'expériences visant le filtrage de lactations et d'épreuves individuelles, le prétraitement des données, la création d'attributs additionnels, le choix d'un algorithme d'aprentissage-machine, et la configuration des algorithmes avaient un effet significatif sur la performance de classification. Pour le filtrage de données individuelles au jour du test, une classification Bayes naïve démontrait une performance significativement meilleure que l'induction d'arbres de décision. Cependant, le spécialiste du domaine considérait les arbres de décision plus transparents que les connaissances générées par l'approche Bayes naïve. La création d'une série de trois classificateurs avec sensibilité accrue au dépens d'une specificité réduite permettait à l'usager d'un système cognitif final de choisir l'intensité désirée de classifier de nouveaux cas comme étant anormaux. Pour les tâches principales d'interprétation, une bonne performance de classification a été atteinte. Pour les tâches de filtrage, la performance était assez pauvre puisqu'une sensibilité raisonnable était associée avec de nombreux faux-positifs. Le spécialiste a évalué les connaissances apprises et a suggéré quelques modifications pour améliorer plusieurs classificateurs. L'acquisition de connaissances assistée par apprentissage-machine a été demontrée comme étant une approche faisable pour soutenir le développement de systèmes cognitifs pour la production laitière. Cette approche sera spécialement appropriée pour des domaines où les spécialistes éprouvent de la difficulté à exprimer les règles de décision, tel que lors de gestion tactique ou lors de l'analyse de représentations graphiques.

Samenvatting

Het doel van deze studie was om het gebruik van machinaal leren voor het ontwikkelen van kennissystemen in de melkveehouderij te onderzoeken. Eerst werd een raamwerk ontwikkeld om de verschillende management- en controleactiviteiten en de informatiestromen tussen deze activiteiten te beschrijven. Dit raamwerk functioneerde als basis voor de ontwikkeling van geautomatiseerde informatiesystemen en was behulpzaam bij het identificeren van de analyse van groepsgemiddelde lactatiecurven als een veelbelovend toepassingsgebied. Een beslissingsondersteunend systeem werd ontwikkeld om een melkveehouderijspecialist te helpen bij het analyseren en classificeren van voorbeelden voor machinaal leren. De specialist classificeerde gegevens van 33 bedrijven die geregistreerd waren bij de melkcontroledienst in Quebec. Dit resulteerde in 1428 lactaties en 7684 melktesten van individuele koeien, geclassificeerd als normaal of uitschieter, en 99 groepsgemiddelde lactatiecurven, waarvan een aantal aspecten geïnterpreteerd waren. Een methode werd ontwikkeld voor een kwantitatieve analyse van de prestatie van classificatiesystemen die via machinaal leren van kleine gegevens bestanden zijn gegenereerd. Deze methode maakte gebruik van kruisvalidatie testen, "relative operating characteristic" curven en variantie-analyse. De voorbewerking van de gegevens, het construeren van additionele attributen, de keuze van een algoritme voor machinaal leren en de configuratie van het algoritme hadden een significante invloed op het classificatieresultaat in experimenten voor het filteren van lactaties en testgegevens. Voor het filteren van individuele testgegevens leidde classificatie via de naïef-Bayes aanpak tot significant betere resultaten dan de inductie van beslissingsbomen. De specialist vond de beslissingsbomen echter gemakkelijker te begrijpen dan de kennis gegenereerd met naïef-Bayes. De creatie van een serie van drie classificatiesystemen met toenemende sensitiviteit ten koste van gereduceerde specificiteit voor elke classificatie taak, maakt het voor gebruikers van het uiteindelijke kennissysteem mogelijk de gewenste gevoeligheid te kiezen om nieuwe gegevens als abnormaal te classificeren. Voor de belangrijkste interpretatietaken werd een goed classificatieresultaat bereikt. Voor het filteren van lactaties en test gegevens was het resultaat vrij slecht omdat een redelijke sensitiviteit was geassocieerd met een grote hoeveelheid normale gevallen geclassificeerd

als abnormaal. De specialist evalueerde de machinaal geleerde kennis en stelde kleine veranderingen voor om een aantal classificatiesystemen te verbeteren. Kennisverwerving met behulp van machinaal leren bleek goed toepasbaar voor de ontwikkeling van een kennissysteem in de melkveehouderij. Deze aanpak is waarschijnlijk het meest geschikt voor probleemgebieden waar het voor een specialist moeilijk is om beslissingsregels aan te geven, zoals tactisch management en het analyseren van grafische informatie.

Acknowledgements

I would like to thank the many people that have supported me in this research. First of all, I wish to thank my thesis supervisor Dr. Kevin Wade for his guidance and support. Secondly, I wish to thank Dr. René Lacroix for the many suggestions and countless hours of consultation. I also like to thank Dr. Daniel Lefebvre and Dr. Elliot Block, the two domain specialists involved in this research, for taking the time to express their expertise in the many consultation sessions and through the analysis and classification of example cases. I greatly appreciate the help of Dr. Roger Cue with the statistical analyses used in this project.

This research was made possible through the financial support of the Natural Sciences and Engineering Research Council of Canada and the Fonds pour la formation de chercheurs et l'aide à la recherche (Gouvernement du Québec). Data used in this research were provided by the Québec dairy herd analysis service (Programme d'analyse des troupeaux laitiers du Québec).

I wish to thank the staff of the Department of Animal Science for their friendly support. Special thanks to my colleagues at the Information Systems Group quarters: Mark Strasser, Ainsley Archer, Madhu Nagaraja, Lael Parrott, John Fletcher, and Annie St-Onge. Your friendship was much appreciated.

I wish to thank my family for their love and support, especially my parents, Riekje Ligterink and Miente Pietersma, who have always encouraged me to pursue my studies. Finally, I wish to thank Anne, my companion for life, and Larissa and Sofia, our daughters, for their patience, encouragement, love, and support, that made the completion of this thesis so much easier.

Contributions to Knowledge

This research resulted in the following original contributions to knowledge:

1. A framework to support the creation of computerized information systems for use in dairy farming, which deals with both management and process-control activities. This framework can be used as a basis for analyzing existing information systems and to support the development of new systems in dairy farming (Chapter 3).

2. An approach to facilitate the acquisition of example cases classified by a domain specialist through the iterative development of a case-acquisition and decision-support system. This approach involves the development of a series of prototypes to enable the specialist to explore new ways of viewing and analyzing the data and to elicit the preferred method of data analysis (Chapter 4).

3. A performance index defined as the mean true positive rate for a specified range of false positive rate values of the relative operating characteristic curve. This index facilitates comparison of the classification performance in machine-learning experiments that involve data with highly unbalanced class distributions. This index makes use of domain expertise to limit the performance analysis to the range of false positive rate values considered reasonable (Chapter 5).

4. An approach to evaluate the results of machine-learning experiments using k-fold cross-validation and analysis of variance with a repeated measures design. With this approach, the factors of interest are analyzed using a mixed statistical model that accounts for the correlation among the repeated measurements on the randomly chosen folds (Chapter 5).

5. A naïve-Bayes algorithm employing a misclassification cost-sensitive approach to attribute selection. This algorithm allows users to direct the attribute selection process to achieve classifiers that focus on correctly classifying a particular class (or classes) of outcome (Chapter 6).

6. An approach to enable the application of two-class performance analysis techniques to classification tasks involving three of more classes. With this approach, classes other than "Normal" are considered as "Abnormal" during performance analysis (Chapter 7).

7. Application of machine-learning assisted knowledge acquisition in the domain of dairy farming. Classifiers generated with machine learning were implemented in the case-acquisition and decision-support system to form a prototype knowledge-based system for the analysis of group-average lactation curves (Chapters 5 through 8).

Contributions of Authors

This manuscript-based thesis was prepared following the December, 2000 revision of the *Guidelines for Thesis Preparation* of the Faculty of Graduate Studies and Research, McGill University. These guidelines include the following statements:

"As an alternative to the traditional thesis format, the dissertation can consist of a collection of papers of which the student is an author or co-author. These papers must have a cohesive, unitary character making them a report of a single program of research… The thesis must be more than a collection of manuscripts. All components must be integrated into a cohesive unit with a logical progression from one chapter to the next. In order to ensure that the thesis has continuity, connecting texts that provide logical bridges between the different papers are mandatory… In general, when co-authored papers are included in a thesis the candidate must have made a substantial contribution to all papers included in the thesis. *In addition, the candidate is required to make an explicit statement in the thesis as to who contributed to such work and to what extent.* This statement should appear in a single section entitled "Contributions of Authors" as a preface to the thesis."

The contributions of the various authors to the chapters representing papers are detailed in the following paragraphs.

The candidate was responsible for: a) part of the ideas represented by the framework described in Chapter 3; b) most of the activities involved in the development of the case-acquisition and decision-support system described in Chapter 4, which included: preprocessing of the milk-recording data of the 33 example herds obtained from the Québec dairy herd analysis service, writing of the computer code for the software, and organizing prototype evaluation sessions with two domain specialists; c) developing an approach to performance analysis for machine-learning experiments using small data sets described in Chapter 5; d) designing and conducting machine-learning experiments to generate knowledge-based components and analysis of the results, which were described in Chapters 5, 6, and 7; and e) preparation of the papers.

Dr. Kevin M. Wade, associate professor at the Department of Animal Science of McGill University, provided supervisory guidance throughout this research and editorial input for the entire thesis.

Dr. René Lacroix, assistant professor at the Department of Animal Science of McGill University, was responsible for part of the ideas represented by the framework described in Chapter 3 and provided input for the entire thesis through extensive consultation.

Dr. Daniel Lefebvre, Department of R&D of the Programme d'analyse des troupeaux laitiers du Québec in Ste. Anne de Bellevue, Québec, contributed as a dairy nutrition specialist to the development of the decision-support system described in Chapter 4. Furthermore, Dr. Lefebvre analyzed and classified example cases and evaluated the results of learning for the machine-learning experiments described in Chapters 5, 6, and 7.

Dr. Elliot Block, formerly professor at the Department of Animal Science of McGill University and presently with Church & Dwight Co. in Princeton, New Jersey, contributed as a dairy nutrition specialist to the development of the decision-support system described in Chapter 4.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Research problem

Decision making on dairy farms has become increasingly complex in recent times and, in addition, dairy producers and their advisors have to deal with an increasing volume of data. For example, dairy herd improvement agencies provide dairy producers with a large number of test-day values related to milk, fat, and protein yields, and optionally, somatic cell count and milk urea nitrogen. These data may improve on-farm decision making, but only if interpreted properly. Computerized information systems have been developed to support dairy producers in dealing with the increased volume of data and complexity of decision making, and include management-information systems (Crosse, 1991; Spahr et al., 1993; Tomaszewski, 1993) and decision-support systems (Allore et al., 1995; DeLorenzo et al., 1992; Grinspan et al., 1994). However, such systems need to be fully integrated with each other and with on-farm installed sensors and robotic units to ensure a coordinated execution of all dairy-farm management and control activities. Various frameworks have been developed to support the long-term development of computerized information systems in dairy farming (Brand et al., 1995; De Hoop, 1988; Devir et al., 1993). However, none of these frameworks covered both management and process control activities. Such a framework could be used to determine potential areas for the development of new decision-support systems and may help to ensure that such systems can function as components within an overall dairy information system.

Decision-support systems can be enhanced with knowledge-based components to automate parts of the decision-making activities. The resulting knowledge-based systems (KBS) may reduce the time required for data analyses and provide dairy producers and their advisors with expert interpretation. The analysis of group-average lactation curves, generated from milk-recording data, is an example of a promising area for KBS development (Whittaker et al., 1989). Group-average lactation-curve analysis involves comparison of group-average lactation curves with standard curves and analysis of additional explanatory data, and may lead to the detection of potential management deficiencies.

Traditionally, the knowledge for KBS has been acquired through interviewing domain specialists and from other sources such as documentation (Dhar and Stein, 1997; Durkin, 1994). However, the acquisition of knowledge through interviews has proven to be time-consuming and difficult. Alternatively, acquisition of knowledge can be partially automated with machine learning (Dhar and Stein, 1997; Langley and Simon, 1995). With this approach, a domain specialist first classifies example cases of the problem at hand. Then, a machine-learning technique, such as decision-tree induction, is used to learn how to classify new cases from these examples. Machine learning may speed up the knowledge-acquisition process (Dhar and Stein, 1997) and may also result in a more accurate representation of the specialist's performance (Michalski and Chilausky, 1980; Ben-David and Mandel, 1995). However, a review of the literature revealed only a few examples of machine-learning assisted knowledge acquisition in the agricultural domain. These included the application of rule induction to develop an expert system for soybean disease diagnosis (Michalski and Chilausky, 1980) and the use of decision-tree induction to support the creation of a KBS for tomato crop management in greenhouses (Mangina et al., 1999). In dairy farming, machine-learning techniques have been applied in the context of modeling and prediction (Kim and Heald, 1999; Lacroix et al.,1995; Mitchell et al., 1996; Nielen et al., 1995; Yang et al., 1999) and for knowledge discovery from large data bases (McQueen et al., 1995; Lokhorst et al., 1999). However, no accounts were found of the use of machine learning to support the acquisition of knowledge from domain specialists.

Several different approaches to machine learning exist, including inductive learning, probability-based methods, genetic algorithms, artificial neural networks, and instance-based learning (Langley, 1996; Mitchell, 1997; Witten and Frank, 2000). Although some of these approaches may lead to better classification performance than others for a given application, the understandability of the learned knowledge may be even more important when these techniques are used in the context of knowledge acquisition. A knowledge representation that is easy to understand allows a domain specialist to evaluate the plausibility of the results of machine learning. Decision-tree induction is generally considered to yield knowledge representations that are easier to understand than many other machine-learning approaches (Dhar and Stein, 1997; Kononenko et al., 1998; McQueen et al., 1995), while the naïve-Bayes classifier - a probability-based approach - might also be a reasonable alternative (Kononenko et al., 1998).

Although machine learning may solve some of the problems associated with traditional knowledge acquisition, new challenges arise. These include decomposition of the overall problem into classification tasks, acquisition of an adequate number of example cases of sufficient quality, creation of potentially predictive attributes, selection and configuration of an appropriate machine-learning algorithm, and interpretation of the learned knowledge (Adriaans, 1997; Langley and Simon, 1995; Verdenius et al. 1997). In the context of machine-learning assisted knowledge acquisition, the most important difficulties might relate to the acquisition of a sufficient number of example cases that have been classified by the domain specialist (Kubat et al., 1998) and the analysis of the performance of classifiers generated from small data sets (Weiss and Kulikowski, 1991; Witten and Frank, 2000).

1.2 Goal and objectives

The main goal of this research was to explore the use of machine learning to support the development of knowledge-based systems in dairy farming. Specifically, the objectives were:

1. to establish a framework for the development of computerized information systems in dairy farming in general, providing a basis for the creation of specific knowledge-based systems;

2. to identify a promising area for machine-learning assisted development of a knowledge-based system;

3. to develop a procedure to support the acquisition of example cases for machine learning from domain specialists, and apply this procedure to obtain example cases for the chosen application area;

4. to develop a method to support the analysis of the performance of classifiers, generated through machine learning from small data sets;

5. to explore the classification performance and understandability of classifiers generated with the decision-tree induction and the naïve-Bayes approach to machine learning;

6. to develop knowledge-based components through the use of machine learning for the identified application area and evaluate their classification performance; and

7. to evaluate the usefulness and limitations of machine-learning assisted knowledge-based system development for dairy farming in general.

2 Literature review

2.1 Decision making in dairy farming

In the past decades, dairy farming has become increasingly complex due to factors such as increased size of operation, higher levels of milk production, demands for improved quality by consumers, and more governmental regulations. Furthermore, dairy producers and their advisors have access to an increasing volume of data collected on the farm as well as from external sources. On-farm collected data may result from observations made by the dairy producer and also from sensors that observe the status and behavior of the cows (Frost et al., 1997; Spahr, 1993; Tomaszewski, 1993). External sources of data include dairy herd improvement agencies, breed associations, artificial insemination units, feed companies, and veterinarians. In addition to the use of sensors to partially automate observation tasks in dairy farming, some of the physical farm activities, such as feeding and milking, can now be partially or completely automated (Lévesque et al., 1994; Rossing and Hogewerf, 1997; Spahr and Puckett, 1986). Although these automated feeding and milking systems reduce the amount of required physical labor, they require additional decision making to update their set points and monitor their functioning.

This wealth of data may improve on-farm decision making, but only if interpreted and utilized appropriately. To support the capturing, storage, and treatment of on-farm collected data, so-called dairy management-information systems have been developed (Crosse, 1991; Spahr et al., 1993; Tomaszewski, 1993). Decision-support systems have been created to help dairy producers in dealing with the increased complexity of decision making, covering such areas as breeding (DeLorenzo et al., 1992; Esslemont and Williams, 1992), health (Allore et al., 1995; Enevoldsen et al., 1995), and nutrition (Grinspan et al., 1994). Some of these systems dealt with short-term decision making, e.g. to support the detection of estrus (Mitchell et al., 1996), while other systems were focussed on decision making with long term effects, e.g. to support the planning of production strategies related to feeding, quota management, and calving patterns (Mainland, 1994).

However, computerized information systems, including management-information systems and decision-support systems, need to be fully integrated with each other and also with sensors and robotic systems for feeding and milking to ensure a coordinated execution of dairy farming activities. The long-term development of such systems should, therefore, be guided by a framework describing the various on-farm management and control activities and the information flows among them (De Hoop, 1988). Several frameworks have been developed to support the creation of computerized information systems for dairy farming (Brand et al., 1995; De Hoop, 1988; Devir et al., 1993). The framework described by De Hoop (1988) and Brand et al. (1995) focussed on dairy farm management activities. Devir et al. (1993) adapted the management framework proposed by De Hoop (1988) to incorporate the management and control involved with automatic milking systems, but did not consider other short-term control activities. Thus, there is a need for a complete framework that deals with both management and process-control activities in dairy farming.

2.2 Analysis of milk-recording data

Dairy herd improvement agencies are an important source of information to support decision making in dairy farming. On dairy farms that are enrolled with a dairy herd improvement program, the milk yield of the lactating cows is measured on a test day and milk samples are taken to determine the percentages of milk fat and milk protein and, optionally, the somatic cell count and level of milk urea nitrogen (Skidmore et al., 1996). Additional variables recorded with such a program may include reproduction and replacement events and ration-related data. This milk-recording data may support on-farm management in areas such as nutrition, health, and breeding (Bailey et al., 1998; Lefebvre et al., 1995; Skidmore et al., 1996).

The milk yield and milk components have a large day-to-day variation within cows (Svennersten-Sjaunja et al., 1997). This means that the performance indices used for milk-recording data interpretation should be based on the average of a sufficiently large number of observations to ensure that detected deviations are due to management instead of normal variability of the data. Therefore, techniques to support the monitoring of milk-recording data collected monthly, tend to focus on data averaged for a group of cows that belong to, for example, a particular parity and stage of lactation. Examples of such techniques include the analysis of group-average lactation curves, mature equivalent milk yields, and milk protein to fat ratios (Bailey et al., 1998; Lefebvre et al., 1995; Skidmore et al., 1996). Due to the large variability in performance within and among cows, proper interpretation of milk-recording data tends to be more difficult for small-sized dairy herds, where group-averaged performance indices are based on only a few observations (Lefebvre et al., 1995).

The process of interpreting milk-recording data not only requires domain expertise but is also repetitive and time-consuming, involving the analysis of a large amount of data, potentially every time new test-day data become available. Several computerized systems have been developed to support this process. The systems described by Fourdraine et al. (1992b) and Jones (1992) provide the user with summaries of the performance of the cows and graphs with individual cow lactation curves and group-averaged lactation curves, respectively. Although these systems automate part of the preprocessing of the data, they leave the task of interpretation to the dairy producers and their advisors. Knowledge-based systems (KBS) may be developed to also automate parts of the interpretation process. Such systems may be appropriate to support or automate several types of activities in dairy farming, including monitoring, diagnosis, and planning (Doluschitz, 1990; Spahr et al., 1988). Several KBS have been developed to support the interpretation of milk-recording data. These include systems to detect potential problems related to mastitis based on somatic cell count data (Allore et al., 1995; Heald et al., 1995), and a KBS to detect management problems based on milk-recording data and additional on-farm collected data available with a dairy management-information system (Pellerin et al., 1994).

At the Texas A&M University, a KBS was developed to support the detection of nutritional problems through the analysis of group-average lactation curves (Fourdraine et al., 1992a; Whittaker et al., 1989). A similar system may be useful to dairy producers in Canada, but should be focussed on the specific Canadian dairy-farming conditions and milk-recording system. Specifically, such a KBS should be able to make use of the data representations of the Canadian dairy herd improvement agencies and deal with the relatively small size of the dairy herds in many Canadian provinces.

2.3 Knowledge-based systems and their development

A KBS can be defined as a computerized system that uses knowledge to solve problems in a particular domain (Gonzales and Dankel, 1993; Plant and Stone, 1991). Knowledge-based systems tend to solve problems using heuristic knowledge, the rules of thumb that specialists use in their reasoning, instead of the algorithmic knowledge used by conventional computing systems. Knowledge-based systems most often use "IF…THEN" or condition-action rules to express and store their knowledge, in which case, they are therefore also-called rule-based systems (Dhar and Stein, 1997).

Similar to conventional software, KBS can be developed in a structured way, following a life-cycle model (Durkin, 1994; Gonzales and Dankel, 1993). For example, the waterfall model of software development represents a stepwise approach from conception through development to implementation. Alternatively, prototyping involves several iterations of identification of requirements, design and development of a prototype, implementation and use, and evaluation. Knowledge-based systems tend to be developed for problems that are relatively complex and poorly structured. Prototypes can be used as a communication vehicle to crystallize the user requirements and to determine the knowledge that must be obtained from the specialist for the KBS to function according to those requirements (Durkin, 1994; Gonzales and Dankel, 1993). Thus, for the development of KBS, prototyping is preferred over the waterfall model.

Traditionally, knowledge acquisition has involved interviewing domain specialists to elicit their knowledge followed by the organization and transfer of this knowledge into a representation that can be used in a KBS (Durkin, 1994; Gonzales and Dankel, 1993). However, this process has proven to be difficult, time-consuming and costly, and has been referred to as the knowledge-acquisition bottleneck (Feigenbaum, 1979). Specialists often have difficulty expressing how they reason and make their decisions and, in addition, it is not easy to structure and encode the knowledge expressed through interviews into a representation that can be used as part of a KBS. A number of knowledge-acquisition tools have been developed to address these problems, ranging from tools to support the interviewing process to machine-learning techniques (Gonzales and Dankel, 1993; Julien et al., 1992). The first generation of knowledge acquisition tools was designed to support the system developer to construct and maintain the KBS, and consisted of functions for bookkeeping and consistency checking. More recent knowledge acquisition tools tend to focus on the specialist rather than the system developer and make use of knowledge-elicitation techniques (Gonzales and Dankel, 1993; Julien et al., 1992).

The acquisition of knowledge from specialists can be partially automated with machine-learning techniques (Dhar and Stein, 1997; Gonzales and Dankel, 1993; Julien et al., 1992). This approach makes use of example cases that have been classified by the domain specialist rather than asking the specialist to express how he or she solves a specific problem. Machine-learning algorithms are able to automatically generate a description of the knowledge embedded in these examples, and use this description to classify new problems. Using machine learning may not only speed up the knowledge acquisition process but may also lead to a more accurate representation of the specialist's performance (Ben-David and Mandel, 1995; Michalski and Chilausky, 1980). Since human specialists often have difficulty expressing how they reason, examples of specialist decisions may represent more reliable information on the specialist's knowledge than his or her own descriptions (Michalski and Chilausky, 1980).

2.4 Machine Learning

Machine learning is a diverse field of research, held together with the common goal of developing computational methods to improve the performance of some tasks (Langley and Simon, 1995). Definitions of machine learning generally focus on two issues: 1) learning involves the generalization of new knowledge from experience, which includes examples and background knowledge, and 2) this knowledge allows for the performance of new tasks or for old tasks to be performed better (Briscoe and Caelli, 1996; Carbonell, 1989; Langley, 1996). Most machine-learning approaches make use of the principle of induction and each one of them requires a description language to describe training examples and the results of learning. Several approaches to machine learning, or machine-learning paradigms, have been developed, each of which can be applied in different contexts of learning.

2.4.1 Induction and description language

Induction involves the generalization of new knowledge descriptions from a large number of examples (Briscoe and Caelli, 1996; Langley, 1996). However, the fundamental problem of induction is the lack of guarantee that the learned knowledge will work for all new cases. Thus, the induced knowledge can only be seen as promising hypotheses that should be tested empirically or examined by human specialists (Guan and Gertner, 1991; Langley, 1996).

Machine-learning systems usually learn from a training set, which consists of examples that are classified by an external source such as a human specialist or some objective measurement. The attributes that represent the input for the classification are called conditional attributes or simply attributes, while the output may be referred to as a decision or classification attribute. The possible values of the classification attribute are called classes or concepts. Most machine-learning techniques expect the training data to be presented in a simple attribute-value or flat-table format (McQueen et al., 1995). Examples of learned knowledge representations include a list of condition-action rules, a decision tree, and a network of nodes and weighted connections (Langley, 1996).

2.4.2 Machine learning paradigms

Within the diverse field of machine learning, several paradigms can be recognized, including inductive learning, probability-based methods, genetic algorithms, artificial neural networks, and instance-based learning (Carbonell, 1989; Langley, 1996; Mitchell, 1997; Witten and Frank, 2000).

Inductive-learning algorithms attempt to induce new knowledge in the form of condition-action rules or decision trees from a set of training examples (Carbonell, 1989; Langley, 1996). The learning algorithms of this paradigm usually perform a heuristic search through the space of possible rule sets and decision trees based on, for example, the degree of impurity or disorder in the data (Bratko et al., 1996; Dhar and Stein, 1997; Langley, 1996). Rule-induction algorithms try to find rules for each class that cover all positive examples, but none of the negative examples (Briscoe and Caelli, 1996; Feng and Michie, 1994). Decision-tree algorithms learn in a top-down fashion using a so-called "divide and conquer" approach (Briscoe and Caelli, 1996; Feng and Michie, 1994; Witten and Frank, 2000). With this approach, attributes are recursively used to split the data into subsets or leaves until each subset only contains examples of one class or until a stopping criterion is reached. At each decision node, the algorithm considers each attribute and attribute value to split the data into sub-sets, and the split that leads to the maximum reduction in impurity of the data is chosen. The resulting decision tree consists of a hierarchical structure going from a root decision node via other decision nodes to the final leaf nodes, each of which indicate the predicted class. A main advantage of inductive learning over other approaches to machine learning is that the knowledge representation is relatively easy to understand for human beings, allowing for evaluation of the learned knowledge (Kononenko et al., 1998; McQueen et al., 1995).

Probability-based learning methods classify a new case using Bayes theorem to calculate the most probable class given the attribute values of the case and knowledge about the prior probability of each class (Mitchell, 1997). A successful approach to probabilistic learning is the so-called naïve-Bayes algorithm, which makes the simplifying assumption that the attribute values are conditionally independent given the class (Mitchell, 1997; Witten and Frank, 2000). The naïve-Bayes approach results in lists of conditional probability values for each attribute and class, which can be analyzed regarding their plausibility (Kononenko et al., 1998).

Genetic algorithms are search procedures based on natural selection, recombination, and mutation (Goldberg, 1994). They can be used for machine learning by considering condition-action rules as binary strings which are manipulated to search for the best set of rules within the entire space of possible rules (Guan and Gertner, 1991; Langley, 1996).

Artificial neural network systems have an architecture based on neuron-like processing units with threshold values and connection weights. These basic components are grouped in several layers and connected to each other, allowing for complex behavior due to their interactions. Learning typically occurs through the adjustment of the connection weights using trial and error runs on training data (Langley, 1996; Rumelhart et al., 1994).

Instance-based learning methods are different from the other machine-learning paradigms since they delay generalization from the example cases until the class of a new case needs to be predicted (Aha, 1992). Instance-based learning is, therefore, also referred to as lazy learning (Mitchell, 1997). During classification, the most similar case or cases are retrieved from storage and their class or classes are used to predict the class of the new case. Case-based reasoning is an instance-based approach that tends to use complex descriptions of cases, rely on complex case-retrieval mechanisms, and involve adjustment of the class or solution of the problem to account for the differences between the new and retrieved cases (Allen, 1994; Kolodner, 1993; Dhar and Stein; 1997).

2.4.3 Context of machine-learning application

Machine-learning techniques can be used for learning in different problem situations or contexts including: 1) modeling and prediction; 2) data mining; and 3) acquisition of knowledge from domain specialists.

Machine learning can be used to develop models to predict or classify some state or behavior (Guan and Gertner, 1991). The advantage of machine learning over other approaches is that it does not require complete quantitative knowledge of the underlying processes as with simulation studies, and that it is not restricted to certain parameter distributions and numerical data as is the case with many statistical approaches (Weiss and Kulikowski, 1991). In the field of dairy farming, the artificial neural networks approach was used by Lacroix et al. (1995) to predict the 305-day production of milk, fat, and protein of dairy cows based on monthly test-day data. To predict the occurrence of mastitis, artificial neural networks have been used (Heald et al., 2000; Nielen et al., 1995; Yang et al., 1999) as well as inductive learning (Kim and Heald, 1999). Marchand (1995) used inductive learning, instance-based learning, and artificial neural networks to predict the future performance of young dairy bulls at the time of their acquisition by an artificial insemination center. And finally, Mitchell et al. (1996) used decision-tree induction to predict estrus in dairy cows based on variations in the daily milk yield and the ranking at which cows arrive in the milking parlor.

Machine-learning techniques can also be used to discover unknown relationships in large data sets, which is often referred to as data mining or knowledge discovery in databases (Fayyad, 1996; Witten and Frank, 2000). In this context, machine learning is used to develop new hypotheses from the data, which subsequently can be tested with traditional statistical techniques (McQueen et al., 1995). In the field of dairy farming, McQueen et al. (1995) used machine-learning techniques to discover the most important factors involved in culling decisions made by dairy producers; Lokhorst et al. (1999) reported on data mining using management-information system data from a group of dairy farms.

One of the main reasons for the initial research efforts in developing machine-learning algorithms has been the knowledge-acquisition bottleneck associated with the creation of KBS (Gillies, 1996; Michie et al., 1994). In this context, machine-learning techniques learn from examples supplied by domain specialists. The classic example of this use of machine learning is the development of an expert system for soybean disease diagnosis with rule induction by Michalski and Chilausky (1980). More recently, decision-tree induction was used to support the creation of a KBS for tomato crop management in greenhouses (Mangina et al., 1999). Many non-agricultural applications of machine learning to support the acquisition of knowledge from domain specialists were reported by Langley and Simon (1995) and in Provost and Kohavi (1998). Although in the field of dairy farming, several KBS have been created based on knowledge from domain specialists, no examples were found in the literature regarding the use of machine learning to support the development of those systems.

2.5 Applying machine learning in practice

Although the use of machine learning to support the acquisition of knowledge for KBS development may solve some of the problems associated with interview-based knowledge acquisition, several new challenges arise. Adriaans (1997) described three bottlenecks of the application of machine-learning techniques to real-world problems: 1) acquisition of an adequate number of example cases of sufficient quality, 2) selection of an appropriate machine-learning algorithm, and 3) interpretation of the learned knowledge. Additional challenges include decomposition of the overall problem into classification tasks (Langley and Simon, 1995; Verdenius et al. 1997), creation of potentially predictive attributes (Langley and Simon, 1995), and configuration of the chosen algorithm (Verdenius et al. 1997).

Several process models have been described to support the successful application of machine learning to real-world problems (Brodley and Smyth, 1997;Langley and Simon, 1995; Verdenius et al. 1997). Aspects of the overall process that are especially important for machine-learning assisted knowledge acquisition include the acquisition of example cases, performance analysis with small data sets, and the evaluation of the learned knowledge.

2.5.1 Process models

Langley and Simon (1995) divided the overall process of using machine learning to develop real-world applications into five main stages: 1) reformulation of the problem into simple classification tasks, 2) deriving potentially predictive attributes from the available data, 3) collection of training data, 4) evaluation of the performance on test data and evaluation of the learned knowledge by specialists, and 5) implementation of the learned knowledge in the field. They emphasized the importance of the first two steps to facilitate the use of relatively simple but robust induction algorithms.

Brodley and Smyth (1997) described the overall process in four main steps: 1) problem analysis and formulation, 2) model and algorithm selection, 3) analysis and diagnosis of test results, and 4) deployment in an operational environment. The first step was detailed into factors related to the application being developed, factors related to the data, and human factors. The second step involved the matching of the problem-dependent factors with domain-independent characteristics of classification models and algorithms to find an appropriate learning algorithm. The last two steps, testing and deployment, were described as leading to iterations of the overall process, since it is not possible to predict how well a particular algorithm will perform when applied to a specific problem.

Verdenius et al. (1997) identified three levels of activities involved in the process of using machine learning to developed real-world applications: 1) application-level activities to decompose the problem into tasks, 2) analysis-level activities to determine the appropriate machine-learning technique based on the type of data , and 3) technique-level activities to configure the chosen algorithm. For each analysis level, a knowledge base could be developed to provide the user with heuristic support.

2.5.2 Acquisition of example cases

Successful application of machine learning requires the availability of a substantial number of labeled example cases. Historical records of example cases analyzed and classified by a domain specialist may not be available. In such situations, the domain specialist will need to classify cases specifically for the KBS being developed. Since the classification of example cases by a domain specialist is expensive, the acquired data set of labeled example cases for machine learning is likely to be small (Kubat et al., 1998).

2.5.3 Performance analysis

The standard procedure to evaluate the performance of knowledge generated through machine learning is training and testing on separate data sets (Weiss and Kulikowski, 1991). This is necessary because machine-learning algorithms tend to overspecialize to the training data, leading to a much better apparent performance (determined through testing of the generated classifiers on the training data) than the true performance with new data. However, with small data sets, the amount of data often limits the achievable classification performance (Cohen, 1995). Thus, in such situations, one would like to use all the available data to generate a final classifier for implementation in the field (Witten and Frank, 2000). For small data sets, stratified ten-fold cross-validation has often been recommended to estimate the true performance on new data of a final classifier induced from the entire data set (Breiman et al., 1984; Mitchell, 1997; Weiss and Kulikowski, 1991; Witten and Frank, 2000). With ten-fold cross-validation, the training set is divided into ten approximately equally-sized mutually exclusive subsets or folds with approximately the same class distribution as the original data set. Each fold is used once for testing of the classifier generated from the combined data of the remaining nine folds. In addition to estimating the performance of a particular classifier, it may be important to determine whether differences among classifiers generated by different algorithms or algorithm configurations, are due to chance or likely to hold for new data. In the machine-learning literature, various statistical techniques to determine differences among machine-learning algorithms have been reported, including paired t-tests and analysis of variance (Bradley, 1997; Dietterich, 1998; Mitchell, 1997). However, these studies focussed on the comparison of machine-learning algorithms instead of comparing classifiers generated from a particular data set.

Analysis of the types of mistakes made by a classifier may also be an important aspect of performance analysis. In many applications, some errors are more important than others. It may, for example, be more important to correctly classify positive or abnormal cases than to correctly classify negative or normal cases (Swets, 1988; Weiss and Kulikowski, 1991). To analyze the types of error made by a classification system, performance indices such as sensitivity and specificity can be used. The sensitivity is defined as the correctly predicted positives as a proportion of the actual positives and the specificity is defined as the correctly predicted negatives as a proportion of the actual negatives (Weiss and Kulikowski, 1991). Machine-learning algorithms can often be tuned to focus on either sensitivity or specificity, which should be taken into account during performance analysis (Provost et al., 1998). The trade-off between sensitivity and specificity that can be achieved with a particular machine-learning algorithm can be visualized with so-called relative operating characteristic (ROC) curves (Swets, 1988). Such curves represent the sensitivity of the classifiers of a classification scheme plotted against one minus the specificity. To facilitate comparison among classification schemes, the area under the ROC curve was proposed by Swets (1988) and used in several machine-learning studies (see e.g., Bradley, 1997; Yang et al., 1999). However, for a particular application, such as data filtering, the entire range of specificity values associated with the area under the ROC curve may not be applicable. With a very low prevalence of positive cases, a low specificity would give too many false positives for the classifier to be of practical use.

2.5.4 Evaluation of learned knowledge

In addition to the quantitative evaluation, it may be useful to let the domain specialist examine the learned knowledge regarding its validity for the problem at hand (Langley and Simon, 1995). The domain specialists may suggest revisions to the representation of the problem or indicate areas of the domain not covered with the learned knowledge. This evaluation requires, however, an understandable representation of learned knowledge.

Inductive-learning approaches, such as decision-tree induction, tend to lead to knowledge representations that are easier to understand by domain specialists and end-users of the system than other machine-learning paradigms (Dhar and Stein, 1997; McQueen et al., 1995). Thus, inductive learning has often been the favored machine-learning approach to support knowledge acquisition. However, inductive-learning techniques have difficulty handling complex problems that have many interactions between the variables (Dhar and Stein, 1997). These situations may lead to many rules or complex decision trees with reduced understandability. Proper decomposition of the overall problem into simple classification tasks is therefore especially important with the use of inductive learning techniques (Langley and Simon, 1995).

2.6 Conclusions

Computerized information systems may be useful to support decision-making activities in dairy farming. However, these activities have a high degree of interdependence and require the exchange of a large amount of information. Thus, a framework describing the types of decision making activities and information flows in dairy farming may be useful to facilitate the long-term development of such information systems.

The analysis of group-average lactation curves has been identified as a useful tool to support decision-making activities related to dairy nutrition. A KBS could support dairy producers and their advisors with the time-consuming and complex task of analyzing group-average lactation curves and related milk-recording data. However, no description was found in the literature of such a KBS developed for the specific dairy farming conditions and milk-recording system in Québec, Canada.

The bottleneck in the development of KBS has been the acquisition of knowledge from domain specialists. Knowledge acquisition through interviews followed by manual transfer of the expressed knowledge into rules has proven to be very time-consuming and costly. Machine-learning techniques, which learn from examples, may be an attractive alternative to the acquisition of knowledge from domain specialists.

Several different machine-learning approaches exist, and they can be used in different contexts of learning. Machine-learning techniques such as inductive learning and artificial neural networks have been used in the field of dairy farming for modeling and prediction, and the discovery of new knowledge from large databases. However, examples of machine-learning assisted knowledge acquisition for KBS development in dairy farming were not found in the literature.

Although machine-learning assisted KBS development may solve some of the problems associated with traditional knowledge acquisition, challenges associated with this new approach need to be addressed. These include decomposition of the overall problem into classification tasks, acquisition of a sufficient number of example cases classified by the domain specialist, creation of potentially predictive attributes, selection and configuration of an appropriate machine-learning algorithm, analysis of the performance of classifiers trained from small data sets, and evaluation of the machine-learned knowledge.

Preface to Chapter 3

Dairy producers and their advisors are facing an increasing amount of information, from sources on and off the farm, and an increasing complexity of decision making. Computerized information systems, such as knowledge-based systems, may be useful tools to help producers to deal with this situation, but should be fully integrated with each other to ensure a coordinated execution of all dairy farm management and control activities. A framework describing the types of decision making activities involved in dairy farming would support the creation of computerized systems that are integrated and support the exchange of information. In addition, such a framework may help to identify and prioritize areas that are expected to benefit from computerized support, to identify the information flows involved among new and existing systems, and to facilitate the reuse of system components.

This chapter describes a framework for the long-term development of computer systems in dairy farming. This framework defines the types of management and control activities involved, how these activities can be performed, and the types of information flows among the management and control activities. This framework forms the basis for the development of computerized information systems in general and, specifically, for the creation of a knowledge-based system in this study.

This chapter was published in the Journal of Dairy Science (Pietersma, D., R. Lacroix, and K. M. Wade. 1998. A framework for the development of computerized management and control systems for use in dairy farming. J. Dairy. Sci. 81(11):2962-2972).

3 A Framework for the Development of Computerized Management and Control Systems in Dairy Farming

Abstract

Computerized information systems can potentially help the dairy producer to deal with the increased complexity of decision making and availability of information in dairy farming. These systems should, however, be fully integrated to ensure a coordinated execution of dairy farming activities. A framework was, therefore, developed to support the creation of computerized management and control systems in dairy farming. Within this framework, a management and control system was defined as a network consisting of the management and control activities and the flows of information that are involved in dairy farming. The management and control activities consist of a cycle of decision making, implementation, and assessment. These activities were classified according to level (strategic, tactical, operational, and regulatory) and sphere (breeding, health, nutrition, environment, milk production, fixed assets, labor, and finance). These activities can be performed by human beings or automated systems and on or off the farm. A large amount of information exchange exists among these management and control activities, and between the overall management and control system, and the physical farm environment and external agents. The interdependence among decisions at the various levels and spheres necessitates computerized management and control systems that are integrated and that allow for easy exchange of information. The developed framework should facilitate the creation of such systems and could also act as a reference base for the analysis and improvement of existing dairy farm information systems.

3.1 Introduction

Decision making on dairy farms has become more complex because of the intensification of dairy farming and factors such as an increase in knowledge about animal management, higher quality demands by consumers, and more governmental regulations. In addition, an increasing volume of data is becoming available from sensors that observe the status and behavior of dairy cows (Frost et al., 1997; Spahr, 1993;

Tomaszewski, 1993) and from external sources including the DHIA, breed associations, AI units, and feed companies. These data may improve on-farm decision making, but only if interpreted and utilized appropriately. Also, some of the physical farm activities, such as feeding and milking, can now be partially or completely automated (Lévesque et al., 1994; Rossing and Hogewerf, 1997; Spahr and Puckett, 1986). Although these automated systems reduce the amount of required physical labor, they nevertheless require additional decisions to be made to update their set points and monitor their functioning.

Computerized information systems, which comprise components such as decision-support systems (DSS) and record-keeping systems, can partially automate the interpretation of data and information and support the dairy producer in dealing with the increased complexity of decision making. However, these information systems, which can be developed for various areas in dairy farming, such as breeding, nutrition, or finance, need to be fully integrated with each other to ensure a coordinated execution of dairy farming activities. The long-term development of computerized dairy farm information systems should, therefore, be guided by a framework describing the various on-farm management and control activities (MCA) and the information flows among them (De Hoop, 1988). Such a framework allows for the modular development of these systems within the global information model of the dairy farm, and can also act as a reference base to compare and analyze currently existing computerized information systems for dairy farms.

Several frameworks have been developed to support the creation of computerized information systems for dairy farming (Brand et al., 1995; De Hoop, 1988; Devir et al., 1993) and for agriculture in general (Gauthier and Kok, 1989; Kok and Lacroix, 1993; Wagner and Kuhlmann, 1991). The framework described by De Hoop (1988) and Brand et al. (1995) focussed on dairy farm management activities. Alternatively, the framework described by Kok and Lacroix (1993) tended to be oriented more toward process control (i.e., related to the regulation of physical processes). However, the management and process-control types of activities on a dairy farm need to be integrated with each other to ensure a coordinated functioning of all farm activities and to allow for optimization of the whole farming operation. The frameworks developed by Gauthier and Kok (1989) and by Wagner and Kuhlmann (1991) included both management and process-control activities, but were concerned with agricultural systems in general and did not deal with the specific MCA involved in dairy farming. Devir et al. (1993) adapted the management framework proposed by De Hoop (1988) to incorporate the management and control involved with automatic milking systems, but did not consider other short-term control activities. Thus, there is a need for a complete framework that deals with both management and process-control activities in dairy farming.

This paper describes a framework for the long-term development of computerized information systems in dairy farming. It focuses on the virtual part of dairy farming (i.e., the information processing activities and information flows on the farm), while the physical implementation of decisions made (e.g., how the actual feeding of the cows takes place) is considered outside the scope of this analysis. The first three sections of the paper describe the information processing activities and information flows that take place on dairy farms in general. The MCA model is described in the first section followed by a classification of dairy farming MCA in the second section. The information flows that drive these information processing activities and that connect them with one another are described in the third section. The fourth section focuses on how and where these MCA can be performed in specific farming situations. The final section deals with aspects concerning the actual creation of computerized information systems in dairy farming, based on the proposed framework.

3.2 Dairy Farm MCA

A framework for the development of computerized information systems in dairy farming should provide a global description of both the information processing activities and the information flows that are required for the proper management and control of the farm (Brand et al., 1995). In this framework, the information processing activities are referred to as MCA. The MCA are interconnected and, together, form the management and control system (MCS). The MCS is thus a network of MCA within which information is exchanged and processed by various MCA. In this framework, the term "information" is interpreted in a broad sense, including signals, data, information, and knowledge. In the real world, many of these MCA are performed by human beings (i.e., dairy producer, farm employees, or external advisors), but some may be automated through implementation in computing devices.

A model describing dairy farm MCA needs to accommodate both management and process-control activities. Farm management can be described in terms of planning, implementation, and control (Boehlje and Eidman, 1984; Kay, 1986). In such a management model (see the frameworks described by Brand et al. (1995) and Devir et al. (1993)), planning consists of selecting a course of action from among various alternatives to accomplish goals; implementation consists of acquiring the necessary resources and putting the chosen plan into action; and control consists of record keeping, evaluating the performance, and taking corrective action if necessary. These functions together form a management cycle in which the control function is followed by improved planning, based on new information (Kay, 1986). Alternatively, with classic process control, a decision-making unit decides upon an action, depending on the difference between the measured and desired state, or the behavior of the system that is being controlled. The action is then carried out upon the system and, finally, the effects are measured for comparison with goals or set points and fed back to the decision-making unit, resulting in a continuously operating control loop (Leigh, 1992). Thus, both management and process control imply some kind of decision-making, implementation of a decision, and feedback of the results. The main difference is that management activities tend to be more complex and ill-defined than process-control activities, usually requiring human participation, while process-control types of activities can often be formalized and fully automated.

In this framework, the management and process-control models were combined into a universal concept of MCA consisting of three functions: decision making, implementation, and assessment (Figure 3.1). In our model, decision making can have varying degrees of complexity, ranging from relatively simple (e.g., temperature control) to quite complex (e.g., whole-farm strategic planning). Decision making can be more than just making a final choice and can involve all phases of the decision-making model defined by Simon (1960) including the detection and diagnosis of problems or opportunities, the development and analysis of alternatives, and the selection of a course of action. Although the implementation of a decision can be a physical act (e.g., detachment of the milking cluster during milking), it can also be a virtual activity, in which case it leads to a new cycle of decision making, implementation, and assessment by another MCA. Assessment involves measurement of the effects of the implemented decision as well as record keeping and feedback of the measured performance. In this definition of an MCA, decision making is restricted to one part of the MCA. Implementation involves decision making only if it implies the activation of another MCA. Assessment does not involve decision making. Detected deviations between goal and measured performance need to be analyzed during a new session of decision making, as described by Huirne (1990).

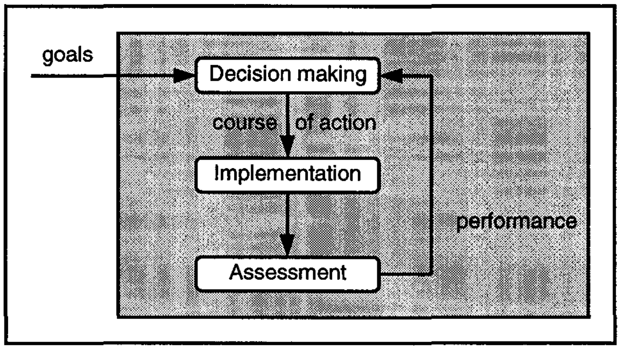


Figure 3.1 Management and control activity model.

3.3 Classification of MCA

The nature of MCA in dairy farming varies considerably: they can be classified according to the level at which they are performed and the sphere of the farm of which they are part. This classification supports the global picture of how the dairy farming activities are related to one another, which is necessary to ensure that computerized dairy farm information systems are developed in a coherent and integrated fashion.

3.3.1 Level of MCA

The level at which MCA are performed reflects the horizon at which factors are taken into account within the MCA and the hierarchy at which they operate. De Hoop (1988) and Brand et al. (1995) considered three levels according to a planning horizon: strategic, tactical, and operational. Alternatively, Gauthier and Kok (1989) considered three levels of control: strategic, tactical, and regulatory. In this framework, the previous notions are combined to consider four levels of management and control: strategic, tactical, operational, and regulatory. At the strategic level, MCA tend to be broad in nature (including the whole farm structure) and focus on the long term (multiple years). Tactical MCA are performed within the scope of the strategic plan to obtain optimal results within the given farm structure. Decisions are focused on the medium term (year or season) and tend to be made more frequently than at the strategic level. Operational MCA are influenced by the actual day-to-day situation on the farm, are related to the implementation of the tactical plans, and focus on the short term (weeks, days, or hours). The regulatory level concerns MCA at a very short term (minutes or seconds), that tend to be continuous, and take place in real time. Although this classification indicates four distinct levels, in reality, MCA are performed within a continuous range from the strategic to the regulatory level. It may, therefore, be better to classify many MCA somewhere between two levels. Daily milking and feeding activities could, for example, be classified at a level higher than regulatory but lower than operational.

The relation between the three MCA functions and the levels of activity can be viewed as nested within each other (Figure 3.2). On the one hand, the implementation of

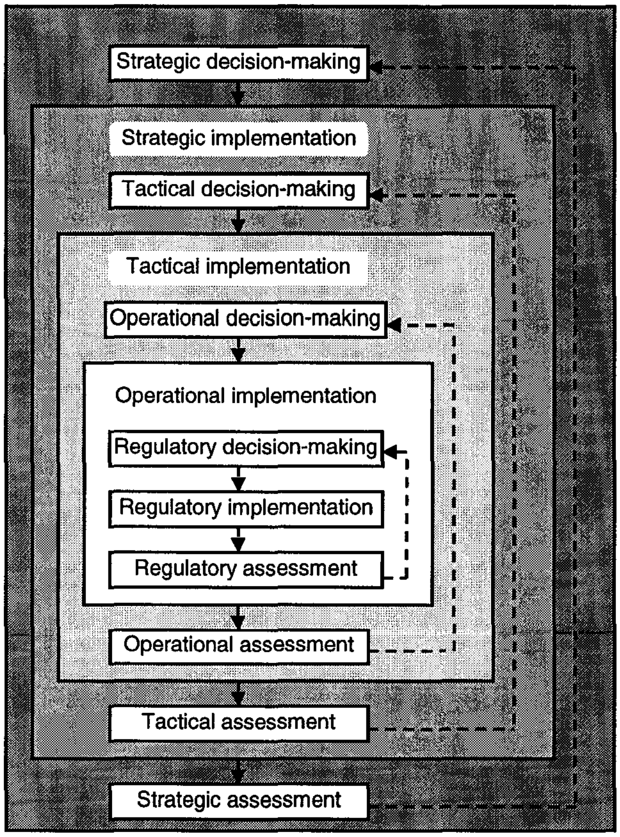


Figure 3.2 The hierarchy of management and control activity (MCA) functions nested within each other.

higher level decisions tends to involve decision making, implementation, and assessment at lower levels. For example, strategic plans need to be implemented at the tactical, operational, and regulatory levels, as in the case of a new breeding goal which requires the choice of sires with traits that conform to this goal or a change in culling criteria. On the other hand, decision making at lower levels needs to be performed within the plans formulated at a higher level. The implementation of strategic plans, therefore, tends to be mainly virtual, because it usually involves additional decision-making activities at the tactical and lower levels; toward the regulatory level, implementation tends to acquire more of a physical nature.

3.3.2 Sphere of MCA

In addition to the level at which they are performed, MCA can also be classified according to the sphere of farming of which they are part. This decomposition into spheres allows one to focus on different areas of expertise within dairy farming. Hogeveen et al. (1991) described a broad classification of dairy-management support into three modules: health, production, and finance. In this framework, a more detailed classification was used, comparable with the dairy farming functions described by Brand et al. (1995). The following eight spheres of activity were recognized: breeding, health, nutrition, environment, milk production, fixed assets, labor, and finance.

The first four spheres (breeding, health, nutrition, and environment) are directly related to the treatment of the current production units (dairy cattle) and future production units (calves and heifers). In the sphere of breeding, decisions are made regarding reproduction (e.g., estrus detection, artificial insemination, pregnancy checking, and calving), replacement (e.g., rearing of replacement heifers and the purchasing and culling of animals), and mating (e.g., program objectives, choice of sires, and age at breeding). The production of embryos, calves, heifers, and cows as genetic material is also classified as part of the breeding sphere. In the health sphere, actions need to be taken concerning prevention, early identification, and treatment of diseases. The nutrition sphere consists of activities such as ration formulation, feed analyses, feed storage, and feeding. The use of fresh grass, either through grazing or summer-feeding, is considered part of this sphere, while growing pasture or other feed crops belongs to separate spheres, outside the scope of this framework. Decisions in the environment sphere relate to the living conditions for the animal, including its welfare. Topics include housing, climate, and wastewater and manure handling. The milk production sphere includes such obvious activities as milking the cows, milk testing, and milk storage, and also covers such areas as analyses of lactation curves for the individual cows and at the herd level and, if necessary, quota management. The labor sphere includes the hiring of labor and the planning and scheduling of the work. The fixed assets sphere involves activities such as acquisition and maintenance of farm buildings, installations, and equipment. The finance sphere involves the management of funds, both for fixed assets and working capital, and includes cash flow decisions, bookkeeping activities for financial, farm-economic, and tax purposes, and the acquisition and repayment of funds. Similar to the classification into levels, clear boundaries between spheres cannot always be drawn. Some activities could be classified as part of multiple spheres (e.g., body condition scoring is related to both health and nutrition).

Table 3.1 Examples of dairy farm management and control activities classified by level and sphere.

Sphere of activity	Level of activity			
	Strategic	Tactical	Operational	Regulatory
Breeding	Development of long term breeding goals	Planning of calving pattern Selection of sires for herd	Selection of sire per cow Culling and buying of animals	Measurement of cow activity
Health	Development of disease prevention strategies	Development of treatment procedures	Diagnosis and treatment of disease	Measurement of body temperature
Nutrition	Choice of feeding system	Seasonal ration formulation based on available feeds	Ration formulation per cow Purchase of feeds	Allocation and transportation of feed to cow
Environment	Choice of ventilation or manure system	Choice of bedding material	Adjustment of climate set points	Climate control
Milk production	Development of long term milk production goals	Development of milking procedures	Identification of cows with abnormal milk	Milking cluster detachment Milk yield measurement
Labor	Hiring of permanent labor	Hiring of seasonal labor	Scheduling of labor	Timing of tasks
Fixed assets	Investment in housing and equipment	Development of maintenance schedules	Maintenance of fixed assets	Control of vacuum level in milking system
Finance	Long term financial planning	Acquisition, investment, and repayment of funds	Cash flow management Bookkeeping	Automatic payment

Table 3.1 gives examples of MCA in different spheres and levels. It should be noted that strategic decision making (e.g., expansion of the herd) often involves the whole farm, in which case it cannot be confined to one sphere of activity. Figure 3.3 gives an example (ration calculation) of the relationships among MCA at different levels and spheres. In this figure, the operational level MCA for ration formulation depends on other MCA within the nutritional sphere, both at the operational and tactical level, and leads to lower level implementation. In addition to the interaction within spheres, a substantial amount of interaction and information exchange exists among the spheres. Decisions concerning one sphere may have important effects on some aspects of others. For example, most decisions involve financial aspects, and decisions in the nutrition sphere influence and depend on the milk production and health of the cows, as shown in Figure 3.3. This high level of inter-relationships among spheres needs to be accounted for in the development of computerized information systems.

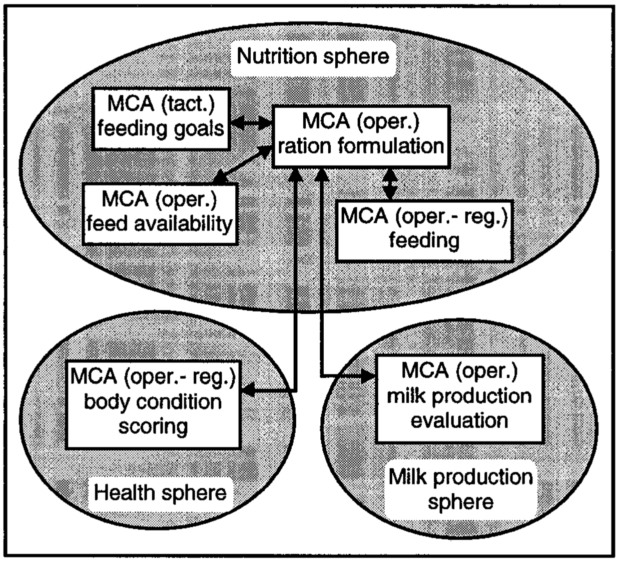

Nutrition sphere
MCA (tact.) feeding goals
MCA (oper.) ration formulation
MCA (oper.) feed availability
MCA (oper.- reg.) feeding
MCA (oper.- reg.) body condition scoring
Health sphere
MCA (oper.) milk production evaluation
Milk production sphere

Figure 3.3 An example of the interactions among the management and control activity (MCA) for ration formulation in the nutrition sphere and management and control activities in the same and other spheres at the regulatory (reg.), operational (oper.), and tactical (tact.) levels.

3.4 Flow of Information

In addition to the characteristics of the MCA, the MCS can also be analyzed in terms of information flows. Information from the physical farm environment and the external world is the driving force for the information processing activities within the MCS; the MCA also transfer information to the physical farm environment and the external world. The description of these information flows defines the required exchange of information with the external world and the physical farm environment and helps to clarify the interactions among the MCA within the MCS.

3.4.1 Information Exchange with External Agents

Dairy producers have to deal with a large amount of information exchange with various external agents. For many dairy farms, the milk recording agency or DHIA is the most important external agent. Based on the milk samples collected regularly on the farm, DHIA provide the dairy producer with test day results, such as milk production, milk fat and protein, and somatic cell count, of the herd and of the individual cows. Other important external agents include the breed associations, AI units, veterinarians, feed companies, and the milk processing industry. Information flows not only from the external agent to the producer but might also flow vice versa. For example, in order to generate farm specific recommendations, external agents need information about the local farm conditions.

The information provided by external agents may be the result of macro-scale management activities. At this macro scale, management activities involve multiple farms in a region or even the whole dairy sector. Examples include comparisons among herds, national genetic evaluations, and the allocation of milk quota to regions. Although these macro-scale activities also consist of decision making, implementation, and assessment, they are not dealt with specifically within this framework. The information from these macro-scale management activities includes reference values against which the farm performance can be compared and general recommendations which need to be fine-tuned to the situation of the local farm. For example, information from national genetic evaluations needs to be filtered according to the personal preferences of the dairy producer for improvement of specific traits.

3.4.2 Information Exchange with the Physical Farm Environment

The MCS also exchanges information with the physical farm environment through observations of the state, behavior, and performance of the cows and other physical entities and also the physical implementation of decisions made within the MCS upon the physical farm environment. The observations of cow variables, such as the milk production and quality, feed intake, and the general state of health, are essential components of many dairy farm MCA. Additionally, observations are required concerning other physical entities on the farm, including the temperature of the milk in the bulk tank, the quantity and quality of the available feeds, and the climate in the barn. The process of making observations, whether by human beings or with sensors, can be seen as a type of MCA at the regulatory level. Within these MCA, decision making is required to translate signals from the physical farm environment to data, while implementation may be needed to trigger other MCA.

Decisions made within the MCS need to be implemented in the physical farm environment, either directly affecting the cows through feeding, milking, and other treatments or indirectly by changing the climate surrounding them. This implementation tends to be part of operational and regulatory level MCA, and can be performed by human beings or automatic devices and robots. The transfer of information from the MCS to the physical farm environment consists of signals that activate and control the physical implementation of the decisions made.

3.4.3 Information Exchange Within the MCS

Within the dairy farm MCS, the MCA need to exchange information with each other in order to manage and control the dairy farm properly. Decisions made at one level often have to be implemented at lower levels, while they may also affect decision making in other MCA at the same or higher levels. For example, the MCA that performs the formulation of feed rations (operational level) needs to communicate new rations to the MCA in charge of feeding (between the operational and regulatory level). On a large farm, employees may use worksheets for this information; in the case of automated feeding, a feed robot needs to be updated with those same inputs. The assessment function often needs information from lower level MCA of the same or other spheres, to be able to determine the performance of the system. For example, the information required to assess the effects of a change in ration includes the actual feed intake of the cows.

Figure 3.4 gives an example of the MCA and information flows involved in ration formulation and feeding. The large amount of information exchanged, both on the farm and between the farm and external agencies, is an important factor in the development of integrated computerized dairy farm information systems. The various components of these automated systems have to be able to communicate not only with the dairy producer, but also with each other and with agents in the external world.

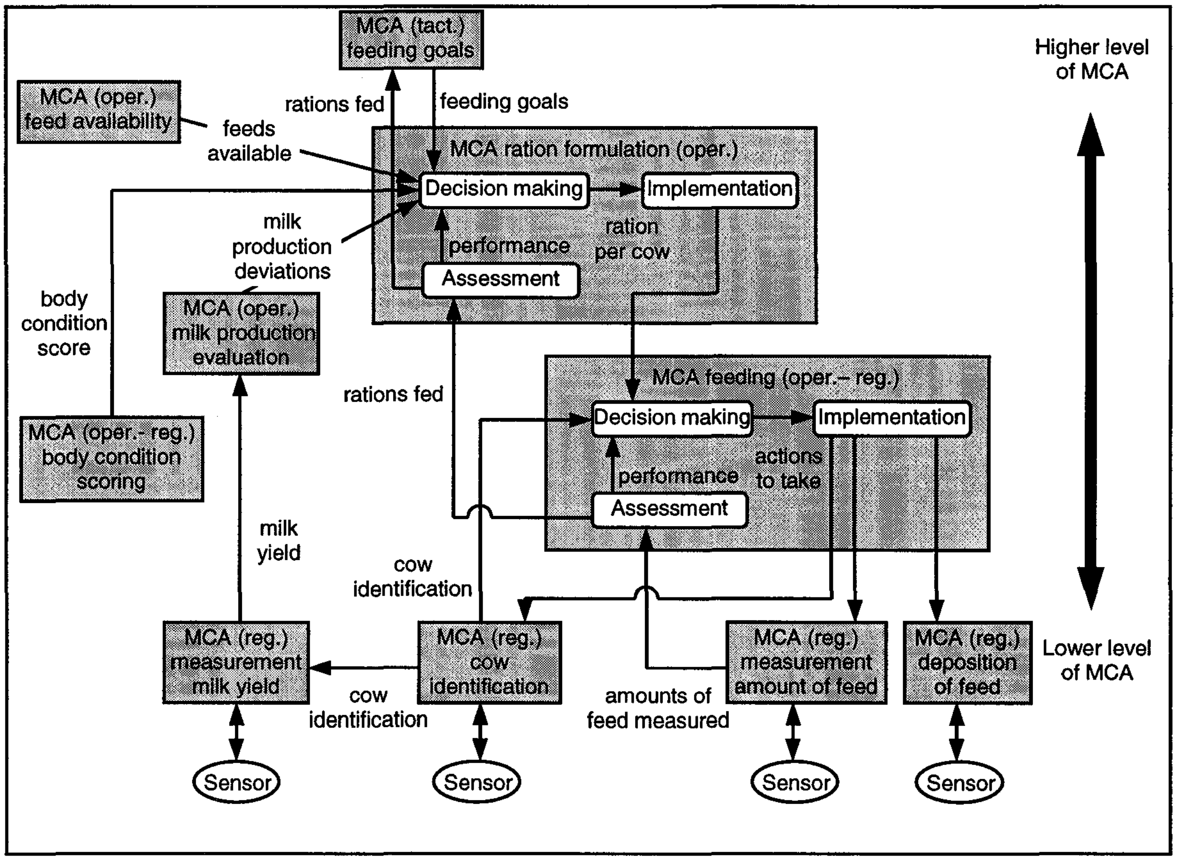

MCA (tact.) feeding goals
MCA (oper.) feed availability
rations fed
feeding goals
feeds available
MCA ration formulation (oper.)
Decision making
Implementation
milk production deviations
ration per cow
performance
Assessment
body condition score
MCA (oper.) milk production evaluation
rations fed
MCA feeding (oper.– reg.)
Decision making
Implementation
MCA (oper.- reg.) body condition scoring
actions to take
performance
Assessment
milk yield
cow identification
MCA (reg.) measurement milk yield
MCA (reg.) cow identification
MCA (reg.) measurement amount of feed
MCA (reg.) deposition of feed
cow identification
amounts of feed measured
Sensor
Sensor
Sensor
Sensor
Higher level of MCA
Lower level of MCA

Figure 3.4 Information flows among management and control activities (MCA) involved in ration formulation and feeding at the regulatory (reg.), operational (oper.), and tactical (tact.) levels.

3.5 Performance of MCA

The MCS described in the previous sections constitutes a general framework of dairy farming activities. However, how these MCA are performed and where they take place may vary considerably on specific dairy farms.

3.5.1 Automation of MCA

At the current level of technology, some of the dairy farm MCA can be automated, leading to a MCS in which MCA can be performed either by humans or by automatic devices, robots and computerized information systems. The possibility of automation varies with the level of MCA. At the regulatory and operational levels, decision making tends to be rather simple and can often be automated with a computer or regulation device. For example, making observations and implementing decisions in the physical farm environment can often be automated with sensor technology and process-control systems. At higher levels, decision making tends to be more complex and ill-structured. In these cases, DSS can be developed to help the human decision maker by automating parts of the decision-making process.

With sensors, some of the observations traditionally performed by dairy producers or their employees can now be automated, including the cow's milk production and activity level (Frost et al., 1997; Spahr, 1993; Tomaszewski, 1993). In addition, sensors allow for the observation of variables that could not be measured previously (e.g., milk temperature and electrical conductivity). With the on-farm implementation of sensors such as electronic milk meters, the frequency with which information of individual cows is being recorded increases from once per month (standard DHIA schedule) to multiple times per day which leads to an enormous increase in the amount of data that needs to be stored, treated, and interpreted. It is, therefore, essential for dairy farms that make use of this sensor technology to have computerized information systems (e.g. record-keeping systems and DSS) in place to support the handling, interpretation, and subsequent use of this information (Frost et al., 1997). The interpretation of data from sensors to detect estrus, for example, can be seen as an MCA in itself. With estrus detection, decision making is required to differentiate normal patterns in the data from those that indicate estrus. Implementation is then required to notify the producer, while assessment may be used to record the detected cases of estrus.

Automatic devices and robots are now available to (partially) automate traditionally labor intensive tasks such as feeding (Lévesque et al., 1994; Spahr and Puckett, 1986) and milking (Devir et al., 1993; Rossing and Hogewerf, 1997). These automated process-control units often consist of a collection of regulatory level MCA to perform sensing and the physical implementation of decisions and higher level MCA to control the overall process. Although these units perform their tasks autonomously, management and control by the dairy producer are required. The functioning in time of these units has to be analyzed to be able to optimize the entire automated process, and set points need to be updated on a regular basis. The automated process-control units also tend to produce large amounts of information that could be used to improve decision making in other MCA.

Human-based MCA that are too complex to be completely automated can be supported with DSS. These are generally defined as computer-based systems that support decision makers to deal with ill-structured decision situations by allowing the user to interact with data, tools, and models (Davis and Olson, 1985; Klein and Methlie, 1995). The term management-information system is often used as a synonym for DSS, although in some definitions a management-information system is restricted to providing the user access to information (Huirne, 1990). At the operational or tactical level, the decision-making process may be structured enough to allow a DSS to produce specific recommendations autonomously. A DSS for breeding decisions, for example, may support the dairy producer in choosing sires for a specific program of genetic improvement or in the ranking of heifers for preferential mating. It is, however, the human decision maker who has to make the final decision. The activities performed by a DSS can be seen as automated MCA, that involve 1) decision making (to generate a recommendation), 2) implementation (to show the user information), and 3) assessment (to record the user's final decision). A DSS may also support the implementation of the final decision made, by printing out worksheets for farm employees or by transferring new set points directly to automated process-control units, such as feed robots or climate computers.

3.5.2 Distribution and Location of MCA

The processing of farm-specific MCA can be centralized or distributed on the farm, and, in addition, some MCA may be processed externally. On a small farm without automation, all of the MCA may be performed in a centralized fashion by one person. However, more often MCA are performed by multiple processors, leading to a distributed type of processing. The decision-making activities may be shared by the dairy producer, external agents, and several processors on the farm, such as the central farm computer with record-keeping and decision-support software, a cow monitoring unit, and a feed robot (Figure 3.5). The different processors in such a distributed system need to be connected through an electronic network to facilitate the large amount of information exchange that exists. An important component of this information network is the central farm computer, which constitutes the main interface between the dairy producer and the other processors of this network.

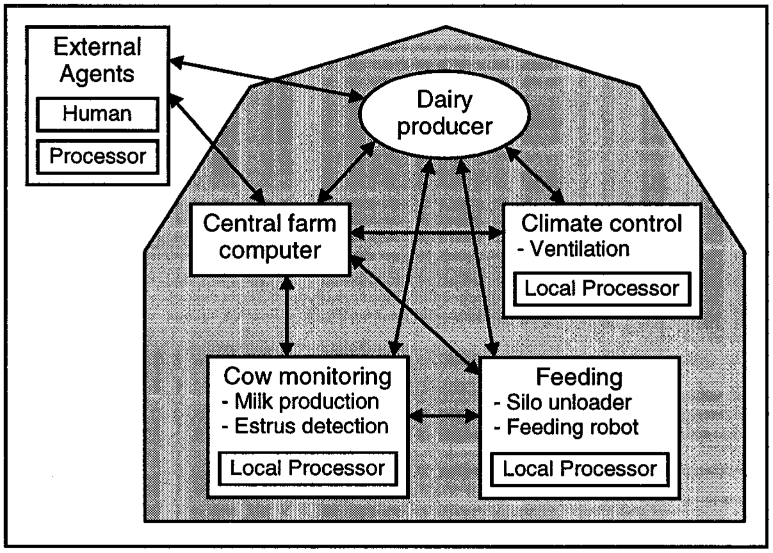


Figure 3.5 Distributed processing of dairy-farm management and control activities.

The most appropriate location to perform farm-specific MCA (on-farm versus external) depends on many factors, such as the frequency at which decision making is required, the expertise and interest of the dairy producer, and the availability of computer capacity to process data and compute performance indices. Operational and regulatory MCA are generally performed most efficiently on the farm due to their frequent use, as in the case of the detection of estrus or mastitis. The formulation of feed rations requires specialized expertise and is often performed by external agents, such as the DHIA or a feed company. The DHIA have traditionally recorded the amount of milk produced and the milk components of individual cows. Based on this data, the DHIA often perform farm-specific MCA such as the computation of performance indices and breeding values. Recent developments in sensor and communication technology may, however, change these patterns of where MCA are performed. On the one hand, the increased use of sensors on the farm (e.g., measurement of milk production and milk quality of each cow at each milking) requires computerized on-farm MCA for the acquisition and interpretation of data related to the cows' performance. On the other hand, the Internet can facilitate the exchange of information between the farm and external agents and may, therefore, support the external processing of MCA.

3.6 Computerized MCS

3.6.1 System Development

The MCS framework constitutes a starting point for the development of computerized management and control systems (CMCS). A CMCS is a subsystem of an MCS and consists of those MCA on a particular farm that are carried out autonomously by automatic devices, robots and computerized information systems. The actual development of dairy CMCS requires detailed analyses of the MCA involved and the information exchange among them. During this development, the MCS can function as a global model of the dairy farm to guide the development of integrated CMCS.

A major advantage of the concept of a network of MCA is that it matches well with an object-oriented approach to the development of computerized information systems (Booch, 1994). Each MCA could be considered as an object, containing both the procedures and the information to perform decision making, implementation, and assessment activities. This approach could ease the incremental development of integrated CMCS, because the object-oriented approach may increase the reusability of computer code after each iteration from initial prototype to final product (Power, 1996). Also, such an approach makes possible the development of farm-specific CMCS, by allowing choice and use of objects or CMCS components that correspond to the specific situation of each farm. An object-oriented approach may also facilitate the large amount of information exchange among the MCA. The objects, or MCA, would interact with each other by sending messages containing information (or requests for information) or by requesting the performance of specific actions.

3.6.2 Integrated and Distributed CMCS

The large degree of interdependence among decisions made at the various levels and spheres requires the development of integrated CMCS. Integration of decision making can only take place through the exchange of information among the various CMCS components, whether they reside on a central farm computer or are physically distributed. Many dairy producers already make use of computerized record-keeping systems to keep track of milk production, breeding, and health data. Newly developed CMCS components should, thus, be able to exchange information with these existing record-keeping systems. The processing of MCA within the CMCS may be physically distributed over multiple processors on the farm as shown in Figure 3.5. These processors, therefore, need to be interconnected through a communication network (e.g., LAN, or Local Area Network) to allow for the direct and efficient exchange of information (Gauthier and Kok, 1989; Kalter et al., 1992). Wireless data transfer may be an interesting avenue for such networks, especially for those units that are not stationary, such as a feed robot and milking units in a tie-stall system.

External agents often perform farm-specific MCA and, as such, are part of the distributed processing of the MCS (Figure 3.5). The information exchange between the farm and external agents has traditionally occurred on paper. However, in recent years, a number of projects have begun to transfer DHIA and other data electronically to the farm through direct modem connections (Lacroix et al., 1997; Tomaszewski, 1993) With electronic transfer of data, on-farm computerized information systems can make use of data directly, without time-consuming and error-prone manual data entry. The Internet makes the exchange of information between the farm and external agencies even easier. Additionally, the Internet may offer more flexibility in the location of decision making. In the future, agencies such as DHIA may provide decision-support services over the Internet, allowing the farm manager to use decision-support software that is located and maintained on a central computer at the external agency (Lacroix and Wade, 1996).

The exchange of information among CMCS components, however, is only possible when protocols for the representation and exchange of information and the calculation of performance indices are available and agreed upon. The development of national and international standards has in this respect many benefits. Such standards could allow dairy producers to purchase the different components of their CMCS from different software and hardware manufactures (Boulesteix et al., 1996; Tomaszewski, 1993), ease the exchange of information with external agents (Boulesteix et al., 1996), and facilitate the comparison of performance indices among dairy farms (Kroeze et al., 1996).

3.6.3 Knowledge-based Techniques

The various components of CMCS can be constructed with knowledge-based or artificial intelligence techniques in addition to the more traditional computing technologies, such as databases, mathematical models, and statistical analyses (Doluschitz, 1990; Hogeveen et al., 1991). These knowledge-based techniques allow for the use of the knowledge of human experts or other sources to automate or support MCA that deal with complex and poorly understood problems. In dairy farming, knowledge-based techniques seem to be especially appropriate for computerized information systems related to monitoring, diagnosis, and planning (Doluschitz, 1990), and in the past decade several such systems have been developed (Allore et al., 1995; Domecq et al., 1991; Grinspan et al., 1994; Schmisseur and Gamroth, 1993; Whittaker et al., 1989). Knowledge-based techniques may be useful at the operational level in domains that are poorly understood (e.g., to interpret data from sensors based on expert knowledge). These techniques may also be useful at the tactical and strategic levels, where the decision-making process tends to be more complex and knowledge from several domains is often required. Hogeveen et al. (1994) gave a comprehensive overview of various types of knowledge representation schemes and showed that the most appropriate representation of knowledge depends on the characteristics of the MCA for which a CMCS component is being developed.

3.6.4 Toward System Autonomy

Complete automation of MCA is at the present time mainly limited to the regulatory and operational levels. Dairy farm MCS with greater autonomy, involving automation of MCA at higher levels, may be possible in the future. Potential areas include the automatic adjustment of feed rations based on the measured cow performance and the automatic adjustment of the number of milkings per day for each individual cow on dairy farms with an automatic milking system. However, systems with greater levels of autonomy require more sophisticated control mechanisms to ensure the survival and stability of the automated system (Kok and Lacroix, 1993).

3.7 Conclusions

A framework has been developed to support the creation of computerized information systems for use in dairy farming. This framework describes the virtual part of dairy farming in terms of an MCS. This MCS consists of a network of MCA, which perform information processing activities and exchange information. The MCA can be oriented toward management or process control, and are classified according to the level at which they are performed and the sphere of dairy farming of which they are part. Information flows are treated in terms of the information exchange among MCA and the exchange between the MCS and both the physical farm environment and the external world. The MCA can be performed by human beings or automated systems and on the farm and externally, leading to MCS that are specific to each dairy farm. The large interdependence among MCA at the various levels and spheres, as well as the distributed decision making (which comes with the increased use of automated systems for monitoring and process control), necessitates CMCS that are integrated and that allow for easy exchange of information. Communication technologies such as Local Area Networks and the Internet are expected to play a major role in CMCS by facilitating the exchange of information on farm and between the farm and the various external agencies. This MCS framework is envisaged to support the development of CMCS by providing a description and categorization of the various kinds of information processing activities and information flows involved in dairy farming. In addition, this framework can act as a reference base for the analysis of existing dairy farm information systems.

Preface to Chapter 4

Knowledge-based systems may help dairy producers and their advisors to deal with the large amount of available information and complexity of decision making in dairy farming. The framework described in Chapter 3 was used to identify promising areas of knowledge-based system development. For example, at the operational level, milk-recording data could be used to find potential health problems, such as sub-clinical mastitis, or to detect management deficiencies related to nutrition. In consultation with dairy specialists, the analysis of group-average lactation curves was chosen as the application area for machine-learning assisted knowledge-based system development. A lactation curve is a graphical representation of the daily milk yield plotted against days after calving. Group-average lactation curve analysis involves the interpretation of lactation curves averaged for groups of cows with the objective to detect potential management deficiencies. For this purpose, cows are generally grouped according to their parity (i.e. lactation number).

In terms of the framework described in Chapter 3, the analysis of group-average lactation curves can be seen as part of a management and control activity within the spheres of nutrition and milk production, and at the tactical level. The decision-making function of this activity makes use of information from related management and control activities on the farm, such as ration formulation, body condition scoring, and health monitoring. External agents, such as the dairy herd improvement agency, veterinarian, and feed companies, also contribute information. Decision making may lead to adjustments to feeding procedures and changes to the base rations, which need to be implemented at the operational level. A decision-support system for the analysis of group-average lactation curves would help the decision-making and assessment functions of the management and control activity through the preprocessing of raw data into performance indices and by allowing the user to interact with appropriate performance representations such as graphs. Adding knowledge-based components to such a system, that contain the expertise of domain specialists would furthermore support decision making via the detection of abnormalities and potential management deficiencies.

This chapter describes a decision-support system for the analysis of group-average lactation curves. The overall problem area involved was decomposed into sub-problems and classification tasks to facilitate the development of knowledge-based modules through machine learning in subsequent research. In addition, case-acquisition functionality was added to the software to enable a domain specialist to efficiently analyze and classify a substantial number of example cases for machine learning.

This chapter was published in the Journal of Dairy Science (Pietersma, D., R. Lacroix, D. Lefebvre, E. Block, and K. M. Wade. 2001. A case-acquisition and decision-support system for the analysis of group-average lactation curves. J. Dairy. Sci. 84(3):730–739).

4 A Case-acquisition and Decision-support System for the Analysis of Group-average Lactation Curves

Abstract

A case-acquisition and decision-support system was developed to support the analysis of group-average lactation curves and to acquire example cases from domain specialists. This software was developed through several iterations of a three-step approach involving 1) problem analysis and formulation in consultation with two dairy nutrition specialists; 2) development of a case-acquisition and decision-support prototype by the system developer; and 3) use of the prototype by the domain specialists to analyze and classify milk-recording data from example herds. The overall problem was decomposed into three sub-problems: removal of outlier tests and lactation curves of individual cows; interpretation of group-average lactation curves; and diagnosis of detected abnormalities at the herd level through the identification of potential management deficiencies. For each sub-problem a software module was developed allowing the user to analyze both graphical and numerical performance representations and classify these representations using predefined linguistic descriptors. The example-based method for the development of the program proved to be very useful, facilitating the communication between system developer and domain specialists, and allowing the specialists to explore the appropriateness of the various prototypes developed. The resulting software represents a formalization of the approach to group-average lactation-curve analysis, elicited from the two domain specialists. In future research, the case-acquisition and decision-support system will be complemented with knowledge to automate identified classification tasks, which will be captured through the application of machine-learning techniques to example cases, acquired from domain specialists using the software.

4.1 Introduction

Dairy producers enrolled in a DHI program receive a large amount of milk-recording data following each test day. These milk-recording data can be a useful source for information to support dairy farm management and control activities, both at the operational (short term) and tactical (medium term) levels of decision making (Pietersma et al., 1998). At the operational level, results from the most recent test day can be used to monitor current performance. At the tactical level, milk-recording data collected (e.g., over the past year) can be used to analyze the performance of the cows averaged for the entire year or as a function of month of year or even stage of lactation. Several computerized information systems have been developed to support the analysis of milk-recording data. These include, at the operational level, a decision-support system for evaluating mastitis information (Allore et al., 1995), fuzzy-set based tools to monitor group-average milk yield and persistency values (Lacroix et al., 1998), and a prototype decision-support system for dairy cattle culling deployed over the Internet (Strasser et al., 1998). In addition, several expert systems have been developed to support tactical level dairy management related to reproduction (Domecq et al., 1991) and covering milk production, nutrition, reproduction, and health (Pellerin et al., 1994). One particular use of milk-recording data to support tactical level dairy management is the analysis of group-average lactation curves (Lefebvre et al., 1995; Skidmore et al., 1996). This type of analysis may, for example, reveal poor peak production for the group of cows in their first, second, or third and higher parity, which may be caused by deficiencies in areas such as nutrition or management of the dry period. However, proper interpretation of group-average lactation curves and additional milk-recording data tend to be time-consuming and complex. Use of a knowledge-based system (KBS) for the partial automation of this process would, thus, be advantageous; it would relieve dairy producers and their advisors from the tedious task of preprocessing the large amounts of raw data, required for such an analysis, and also provide them with expert interpretation (Whittaker et al., 1989).

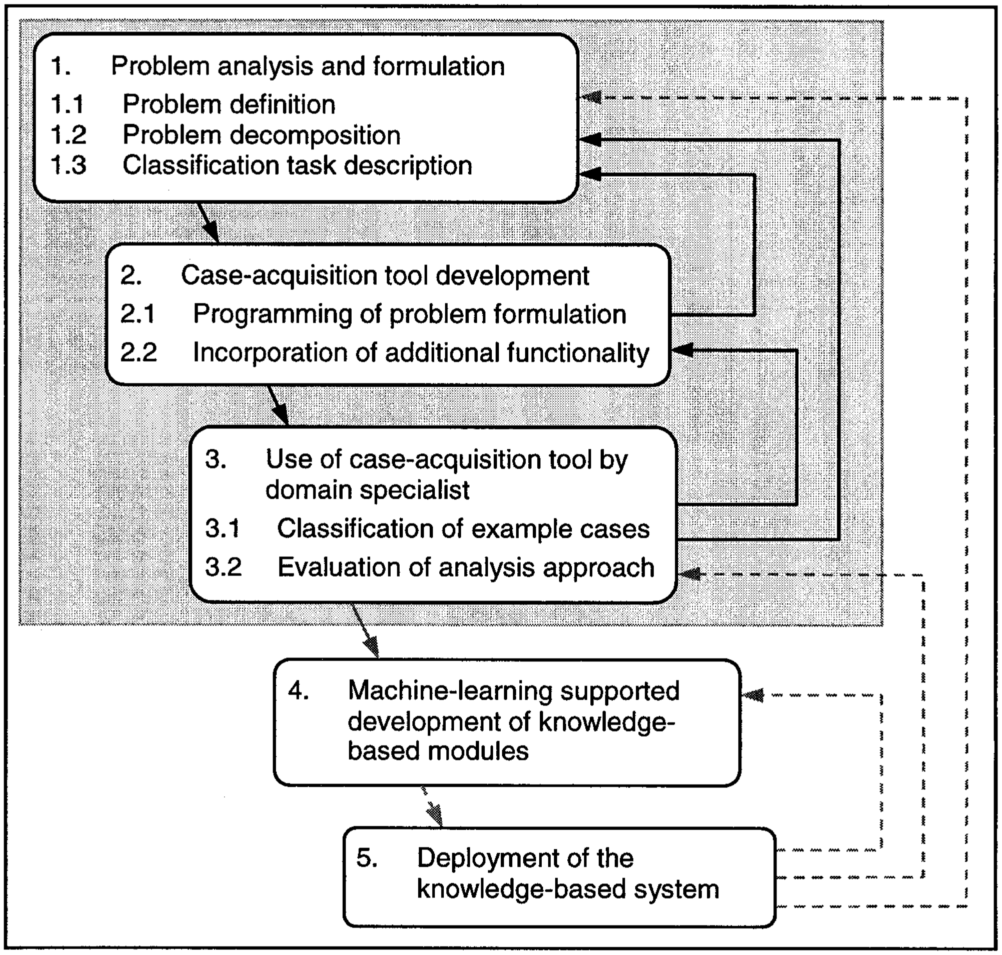
Traditionally, KBS have been developed based on interviews with domain experts, sometimes supplemented with other sources of knowledge such as documentation (Dhar and Stein, 1997; Durkin, 1994). However, the acquisition of knowledge through interviews has proven to be time-consuming and difficult, being referred to as the knowledge-acquisition bottleneck. Alternatively, the acquisition of knowledge from experts can be partially automated with machine-learning techniques (Dhar and Stein, 1997; Durkin, 1994; Langley and Simon, 1995). With this approach, a domain expert first classifies example cases of a particular problem, followed by the application of machine-learning techniques, such as decision-tree induction and instance-based learning, to discover and make use of the knowledge implicitly embedded in these example cases. However, new challenges arise with the application of machine-learning techniques to real-world problems, including task decomposition, acquisition of example cases of sufficient quality, selection and configuration of an appropriate machine-learning algorithm, and interpretation of the learned knowledge (Adriaans, 1997; Langley and Simon, 1995; Verdenius et al., 1997).

A research project was initiated to explore the use of machine learning for the development of KBS in dairy farming and focussed on the problem area of group-average lactation-curve analysis. The research presented in this paper dealt with the first part of this project and addressed the challenges related to task decomposition and acquisition of example cases. It was expected that decomposition of the problem area into classification tasks could be achieved through interviews with domain specialists, leading to a formalization of the approach to group-average lactation-curve analysis used by these specialists. In order to perform machine learning of the identified classification tasks, a substantial number of example cases would be required. The development of a software tool to automate the acquisition of example cases from domain specialists was expected to solve this case-acquisition bottleneck, while, as an implementation of the formalized approach to group-average lactation-curve analysis, this tool could also be used as the core software of the final KBS. The objectives of this research were 1) to formulate the overall approach to group-average lactation-curve analysis into classification tasks that allow for the application of machine-learning techniques and 2) to develop a case-acquisition tool to enable domain specialists to efficiently analyze and classify example cases of the analysis of group-average lactation curves.

4.2 Materials and Methods

4.2.1 Procedure

In order to support the analysis of group-average lactation curves, problem formulation and case-acquisition tool development were both carried out in consultation with two dairy nutrition specialists. Both specialists had practical experience with lactation-curve analysis and one of them had direct knowledge of the type of support currently provided by the DHIA advisors. The approach used involved three consecutive steps as indicated by the grayed area of the process model shown in Figure 4.1: problem analysis and formulation, case-acquisition tool development, and use of this tool by the domain specialists. These three steps are part of an overall process model, which was developed specifically for the context of using machine learning to support the acquisition



1. Problem analysis and formulation
1.1 Problem definition
1.2 Problem decomposition
1.3 Classification task description
2. Case-acquisition tool development
2.1 Programming of problem formulation
2.2 Incorporation of additional functionality
3. Use of case-acquisition tool by domain specialist
3.1 Classification of example cases
3.2 Evaluation of analysis approach
4. Machine-learning supported development of knowledge-based modules
5. Deployment of the knowledge-based system

Figure 4.1 A process model for machine learning to support knowledge acquisition for knowledge-based system development.

of knowledge from domain specialists for KBS development based on three methodologies (Brodley and Smyth, 1997; Langley and Simon, 1995; and Verdenius et al., 1997) for the application of machine learning to real-world problems in general. Additional steps in this process model include the development of knowledge-based modules through machine learning from example cases acquired with the case-acquisition tool and deployment of the KBS (Figure 4.1). This study was restricted to the first three steps and designed as an iterative process, which is represented by the various feedback loops in Figure 4.1.

The first step of the process model was carried out in consultation with the domain specialists and involved 1) definition of the overall problem, 2) decomposition of the overall problem into sub-problems (with reduced complexity to facilitate machine learning in subsequent research), and 3) reformulation of each identified sub-problem into one or multiple classification tasks for machine learning, including a description of the attributes and classes that characterize example cases for those tasks. The second step consisted of the development of a prototype case-acquisition tool by the system developer, based on the problem formulation resulting from the first step. In addition, the system developer incorporated alternative views of the data into the prototype based on data visualization techniques. In the third and final step, the two domain specialists used the developed prototype to analyze and classify a small data set consisting of milk-recording data from a selected number of dairy herds, and to evaluate the functioning of the program and the appropriateness of alternative views of the data. The deficiencies of the prototype were then discussed with the specialists, prompting a new iteration. During these discussions a computer with the case-acquisition prototype was used to support the analysis of the functioning of the program by stepping through specific example cases. This consultation led to adjustments in the problem definition and decomposition, the description of classification tasks, and the preferred views of the data representation, followed by the development of an improved prototype. For each sub-problem, several such iterations were required before the specialists were satisfied with the resulting analysis approach and case-acquisition program. The feedback loop from step two to step one represents situations in which the system developer required additional input from the specialists before releasing the next prototype for subsequent use. The loop from step three to step two occurred when a software bug was detected by the specialists and reported back to the system developer. During the project, a large amount of decision-support functionality was added to the case-acquisition tool enabling the specialists to analyze the milk-recording data efficiently using an approach to group-average lactation-curve analysis that emerged from the iterative development process. The resulting software was, therefore, referred to as a case-acquisition and decision-support system (CADSS).

4.2.2 Data

A data set, consisting of milk-recording data from 33 Holstein herds, was used throughout the development of the CADSS. These herds were randomly selected from herds enrolled in the provincial DHI program, while ensuring coverage of a wide range of rolling herd average milk production levels. For each herd the data were limited to one year of historical milk-recording data by choosing a so-called "most recent test date" and including only tests no more than 365 days prior to that date. In addition, only the data from lactations starting within the defined interval were included in the analyses. This resulted in a total of 1428 lactations, produced by 1419 different cows, and a total of 7684 tests. The data set for each cow and test day included such variables as milk yield (kg), fat %, protein %, SCC, and codes reflecting conditions that may have affected the performance of the cow on the test day. Milk urea nitrogen data were available for some of the tests. Ration data included DMI, NDF, non-structural carbohydrates, NEL, fat, CP, and undegraded intake protein. Total DMI was based on actual amounts of supplements fed (reported by the producer) and estimates of the forage DMI (determined by the Québec DHIA ration model and adjusted by the feed advisor). The quality of the feed ingredients was based on laboratory analyses of feed samples or, if no specific analyses were available, on standard values for the composition of feeds (National Research Council, 1989). The data set also included birth, calving and dry-off dates, body weight after calving, and the body condition score at four different stages of lactation. See Figure 4.2 for a summary of the data input variables. Standard lactation curves and standard peak levels for seven herd-average production levels and three parity groups (one, two, or three and higher) had previously been derived from 570,863 official Holstein test-day records at the provincial DHIA by Lefebvre et al. (1995) and were used as performance benchmarks for the group-average lactation curves. These standard curves were associated with herd-average cumulative 305-day milk production levels ranging from 6750 to 9750 kg in intervals of 500 kg. Additional standard curves and peak production levels were estimated through linear extrapolation to 4250 and 11,250-kg herd-average cumulative 305-day milk production to accommodate very low and very high producing herds.

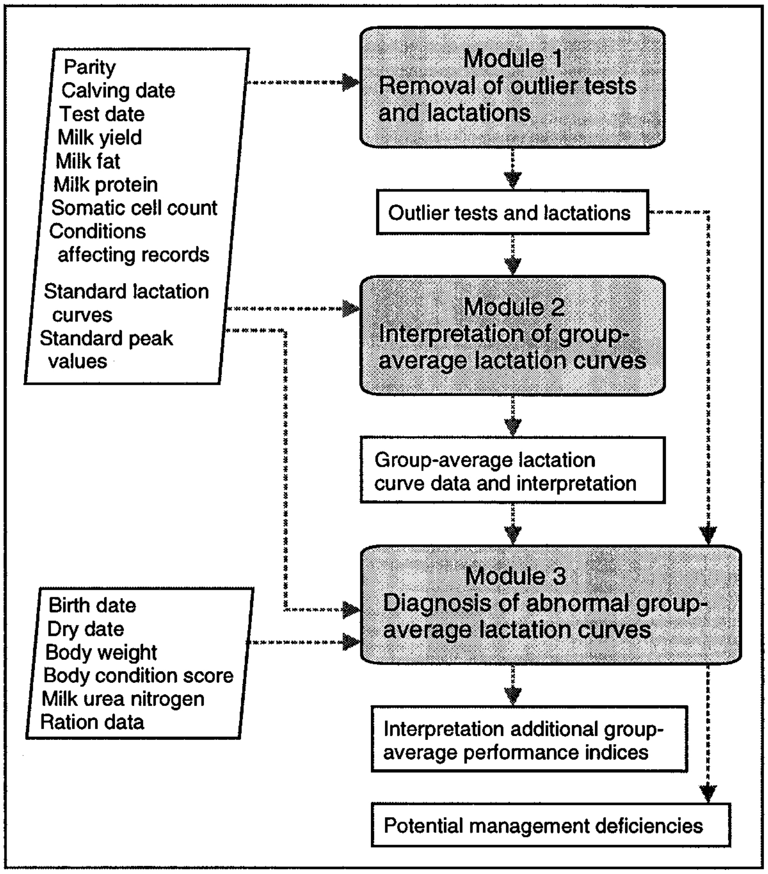

Parity
Calving date
Test date
Milk yield
Milk fat
Milk protein
Somatic cell count
Conditions affecting records
Standard lactation curves
Standard peak values
Module 1
Removal of outlier tests and lactations
Outlier tests and lactations
Module 2
Interpretation of group-average lactation curves
Group-average lactation curve data and interpretation
Birth date
Dry date
Body weight
Body condition score
Milk urea nitrogen
Ration data
Module 3
Diagnosis of abnormal group-average lactation curves
Interpretation additional group-average performance indices
Potential management deficiencies

Figure 4.2 Milk-recording data inputs, case-acquisition and decision-support system modules, and captured classifications.

The CADSS was developed using the Visual Basic (Microsoft Corporation, Redmond, WA) programming language. For each herd a relational database with three database tables was used to store the milk-recording data pertaining to the level of cow and test, cow and lactation, and herd and test. The classifications made by the user during a consultation session with the CADSS were also stored in a relational database. The Data Access Objects programming model (Microsoft, 1997) was used to enable the CADSS to access and manipulate the data in the database tables.

4.3 Results

4.3.1 Problem Definition and Decomposition

The two domain specialists described the analysis of group-average lactation curves as a first step in a tactical level management activity focussed on monitoring and improving the nutrition management on the farm. The specialists anticipated that this analysis process would lead to the detection of potential management problems and preliminary diagnoses.

The consultation sessions with the domain specialists resulted in a decomposition of the overall problem into three sub-problems and the development of three corresponding software modules (Figure 4.2). These sub-problems were 1) removal of outlier tests and lactations of individual cows, 2) interpretation of group-average lactation curves, and 3) diagnosis of abnormal group-average lactation curves. This decomposition corresponded to the analyses of milk-recording data at the levels of individual cow, group of cows, and the entire herd. The first sub-problem was related to the relatively small herd size of dairy herds enrolled in the Québec DHI program, which, in 1999, averaged 45 cows per herd (Programme d'analyse des troupeaux laitiers du Québec, 2000). With a small number of cows in a group, the interpretation of the group-average performance may be biased towards a single atypical cow. The specialists, therefore, considered the removal of outlier tests and lactations to be important, especially if these outliers were associated with explanatory information such as a high SCC, an extreme protein to fat ratio, or the existence of codes indicating specific events affecting the milk yield (e.g., clinical mastitis, displaced abomasum, or estrus). The second sub-problem involved the interpretation of group-average lactation curves and peaks for each of the three parity groups in order to detect abnormalities such as a poor peak production or an abnormal shape of the curve after the peak. The group-average lactation curve and peak were calculated from non-outlier milk yield data and could be compared with a standard lactation curve and standard peak level. The resulting group-average lactation curve interpretations for all three parity groups of a herd were analyzed in combination with additional group-averaged milk-recording data in the final sub-problem to diagnose detected abnormalities through the identification of potential management deficiencies.

The overall CADSS program and the three software modules corresponding to the identified sub-problems are explained in detail below.

4.3.2 CADSS Program

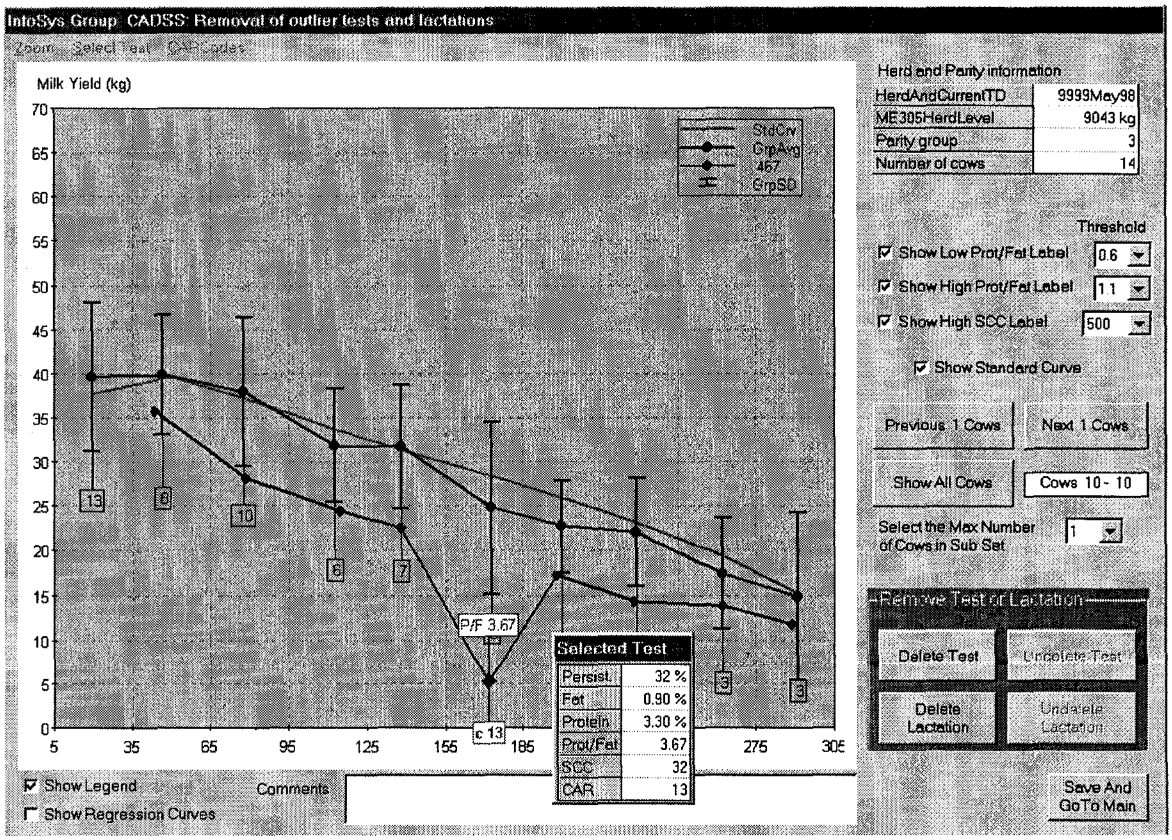
The CADSS program consists of a main module, to control the selection of herds, parity groups, and the sequence of analyses, and three additional analysis modules corresponding to the identified sub-problems. During a case-acquisition session, the CADSS program controls the sequence of analysis steps. After selection of a herd, the user is first directed to the module for removal of outlier tests and lactations for each of the three parity groups. The user may then continue with the module for the interpretation of group-average lactation curves. After this stage, the final module can be used to diagnose detected abnormalities. The user can go back to herds and parity groups, classified earlier, to review and, if necessary, change the classification decisions made.

The CADSS records the user interactions with the system and the classifications made by the user in a single relational database containing seven database tables. Four tables are used to record the classifications pertaining to four aggregation levels: test within lactation, lactation, parity group, and herd. The CADSS uses three additional tables to record the sequence of decision-making steps by the user for each of the three analysis modules of the program. The final classification decisions made by a specialist could be combined with potentially predictive attributes derived from the information shown at the time of decision making to generate example cases that could, in subsequent research, be used for machine learning. The recorded sequence of decision-making steps allowed for a replay of how the user interacted with the system and helped to define the gray or fuzzy zones of classification in cases where it was difficult for the specialist to choose between two classes.

The development of each module required several iterations involving the three consecutive steps of 1) discussion of the required functionality of the module, 2) prototype development, and 3) use of the prototype by the domain specialists (Figure 4.1). Shorter iteration cycles of only step one and two were used when adding alternative views of the data representation proposed by the system developer. Such additional functionality was first explained to the specialists during a consultation session using a computer with the prototype to show specific example cases. Based on the feedback from the specialists, the prototype was then improved and subsequently released for use by the specialists to complete a three-step iteration. The development of the first, second and third module involved, respectively, three, five, and six iterations. The modules were developed in parallel requiring a total of seven consultation meetings. The following three sections describe, for each module, the required milk-recording data, the classification tasks involved, the resulting functionality, and the acquisition of example cases.

4.3.3 Removal of Outlier Tests and Lactations (Module 1)

With the first module the user can analyze lactation curves of the individual cows within a parity group and perform two classification tasks: removal of outlier tests and/or removal of outlier lactations (Figure 4.3). The input for this module consists of eight different variables describing the lactation and the tests within lactation for each cow, and standard lactation curve values that can be used as a benchmark (Figure 4.2). The module results in the classification of each test and lactation as outlier or non-outlier.


InfoSys Group CADSS: Removal of outlier tests and lactations

Herd and Parity information

HerdAndCurrentTD	9999May98
ME305HerdLevel	9043 kg
Parity group	3
Number of cows	14

Threshold

☑ Show Low Prot/Fat Label 0.6

☑ Show High Prot/Fat Label 1.1

☑ Show High SCC Label 500

☑ Show Standard Curve

Cows 10 - 10

Select the Max Number of Cows in Sub Set 1

Selected Test

Persist.	32 %
Fat	0.90 %
Protein	3.30 %
Prot/Fat	3.67
SCC	32
CAR	13

☑ Show Legend

☐ Show Regression Curves

Comments

Figure 4.3 Screen capture of module 1: removal of outlier tests and lactations.

The user can choose to view the lactation curves of all cows belonging to the parity group or step through the lactation curves one at a time or in smaller selected groups. In addition to the individual cow curves, a group-average lactation curve is shown, averaged for non-outlier milk yield and DIM values within each of ten stages of lactation from 5 to 305 DIM. The group-average curve includes error bars representing the mean milk yield plus and minus the standard deviation. Labels are attached to individual milk yield tests in the case of codes indicating an event that may have affected the test results (such as clinical mastitis, displaced abomasum, or estrus). The user can choose to view additional labels to draw attention to a low or a high protein to fat ratio, or a high SCC, each with user-adjustable threshold values. A standard lactation curve can be displayed to represent the expected performance for the parity group, given the production level of the herd. Optionally, a regression line can be estimated and shown for each lactation curve using a multiple linear regression model proposed by Wilmink (1987) describing milk yield Y on DIM t as

$$Y(t) = b0 + b1\ t + b2\ \exp(-0.05\ t).$$

Selection of a particular test within a lactation curve by mouse click prompts the program to show additional information for that test, including persistency, fat and protein percent, protein to fat ratio, and SCC. The user can subsequently delete the selected test or an entire lactation by clicking on the "Delete Test" or "Delete Curve" buttons. Removal of all tests in a lactation causes the program to consider that lactation as deleted. The color of a deleted test or lactation is changed to gray to aid in the reselection of previously deleted tests or lactations. Each "delete" or "undelete" event prompts the program to recalculate the group-average lactation curve.

The CADSS records a flag for each deleted test or lactation to indicate it as an outlier. Table 4.1 shows examples of the classification of milk yield tests using the classification attribute “Test is outlier” and classes “True” and “False”. The example cases in Table 4.1 are described with a selection of nine of the many potentially predictive attributes that can be used for machine learning of this classification task.

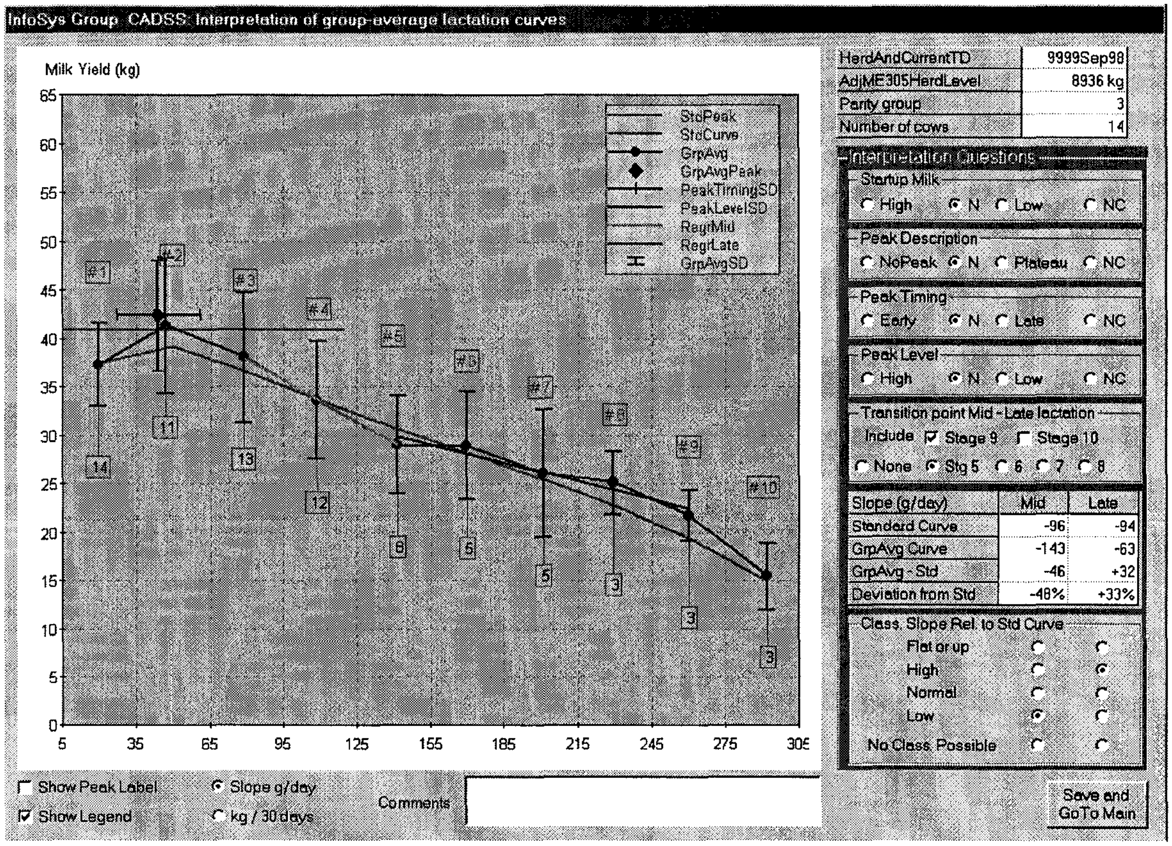
Table 4.1 Example cases of milk yield tests described by a selection of potentially predictive attributes for machine learning and classified by a domain specialist as outlier or non-outlier.

Potentially predictive attributes									Test is outlier
Persistency selected test (%)	Persistency next test (%)	Milk fat (%)	Milk protein (%)	Protein to fat ratio	SCC	Event code	Milk yield – GrpMilk[1] (kg)	(Milk yield – GrpMilk) / SDGrpMilk[2]	
44	146	6.5	4.2	0.65	5,890,000	Mastitis	- 16.3	- 1.9	True
89	123	2.6	2.9	1.11	101,000	0	- 8.5	- 1.3	True
79	107	3.4	3.0	0.89	1,176,000	0	+ 4.7	+ 0.9	True
97	90	4.1	3.3	0.80	40,000	0	- 3.2	- 1.2	False
89	95	3.4	2.9	0.86	19,000	Heat	+ 4.5	+ 0.9	False
79	96	5.0	4.0	0.80	1,233,000	0	- 3.7	- 1.1	False

[1] GrpMilk = group-average milk yield.
[2] SDGrpMilk = Standard deviation of group-average milk yield.

4.3.4 Interpretation of Group-Average Lactation Curves (Module 2)

With the second module the user has to characterize the group-average lactation curve of the selected parity group by choosing an option button for each of six linguistic descriptors or classification tasks. These tasks are listed in the right-hand section of the module (Figure 4.4) and consist of: start-up milk, peak description, peak timing, peak level, and slope of the curve during mid and late lactation. The input for this module consists of: the milk yield and date of each test, the parity and calving date of each lactation, standard lactation curves and standard peak level values to benchmark the group-average performance, and the outlier test and lactation flags from the first module. The interpretation of the group-average lactation curve results in a list of classes chosen by the user for each classification task (e.g., low start-up milk, normal peak description, normal peak timing, low peak level, high slope mid lactation, and normal slope late lactation).



InfoSys Group CADSS: Interpretation of group-average lactation curves

HerdAndCurrentTD	9999Sep98
AdjME305HerdLevel	8936 kg
Parity group	3
Number of cows	14

Interpretation Questions

Startup Milk
○ High ● N ○ Low ○ NC

Peak Description
○ NoPeak ● N ○ Plateau ○ NC

Peak Timing
○ Early ● N ○ Late ○ NC

Peak Level
○ High ● N ○ Low ○ NC

Transition point Mid - Late lactation
Include ☑ Stage 9 □ Stage 10
○ None ● Stg 5 ○ 6 ○ 7 ○ 8

Slope (g/day)	Mid	Late
Standard Curve	-96	-94
GrpAvg Curve	-143	-63
GrpAvg - Std	-46	+32
Deviation from Std	-48%	+33%

Class. Slope Rel. to Std Curve

Flat or up	○	○
High	○	●
Normal	○	○
Low	●	○
No Class. Possible	○	○

□ Show Peak Label
☑ Show Legend
● Slope g/day
○ kg / 30 days

Figure 4.4 Screen capture of module 2: interpretation of group-average lactation curves.

In this module, the group-average lactation curve is shown, averaged for non-outlier milk yield and DIM values of individual cows within each of ten stages of lactation from 5 to 305 DIM. Error bars represent the group-average milk yield plus and minus the standard deviation. The group-average peak level and timing are calculated based on the maximum milk yield of individual cows in the first 120 DIM. The group-average peak level and timing error bars represent the mean plus and minus the standard deviation. The program calculates a standard lactation curve for the selected parity group through linear interpolation between the two available standard curves closest to the mature equivalent production level of the herd in question. The standard peak level is derived similarly. The user can choose a class for each classification task by selecting an option button (for example, peak level can be "Low", "Normal", "High", or "No Classification Possible", while slope during mid and late lactation can be "Low", "Normal", "High", "Flat" or "No Classification Possible"). Selection of a transition point between mid and late lactation prompts the program to estimate and show two linear regression lines through the group-average curve data points from the third stage to the transition stage and from the transition stage to the last stage. The slopes of these two regression lines (in grams per day) are displayed in a table together with the slopes of regression lines through the standard lactation curve. This functionality was added to support the user with the classification of the slope of the group-average lactation curve during mid and late lactation.

Table 4.2 shows example cases of the group-average peak level, classified by a domain specialist as “Low”, “Normal”, or “High”. Potentially predictive attributes describing the group-average peak level in relation to the standard peak level may include the absolute and relative distance between the group-average and standard peak level, and the absolute distance, expressed in standard deviations. The selection of a transition point between mid and late lactation can be seen as a seventh classification task which may be useful to calculate predictive attributes for the classification of the slope of the lactation curve during mid and late lactation.

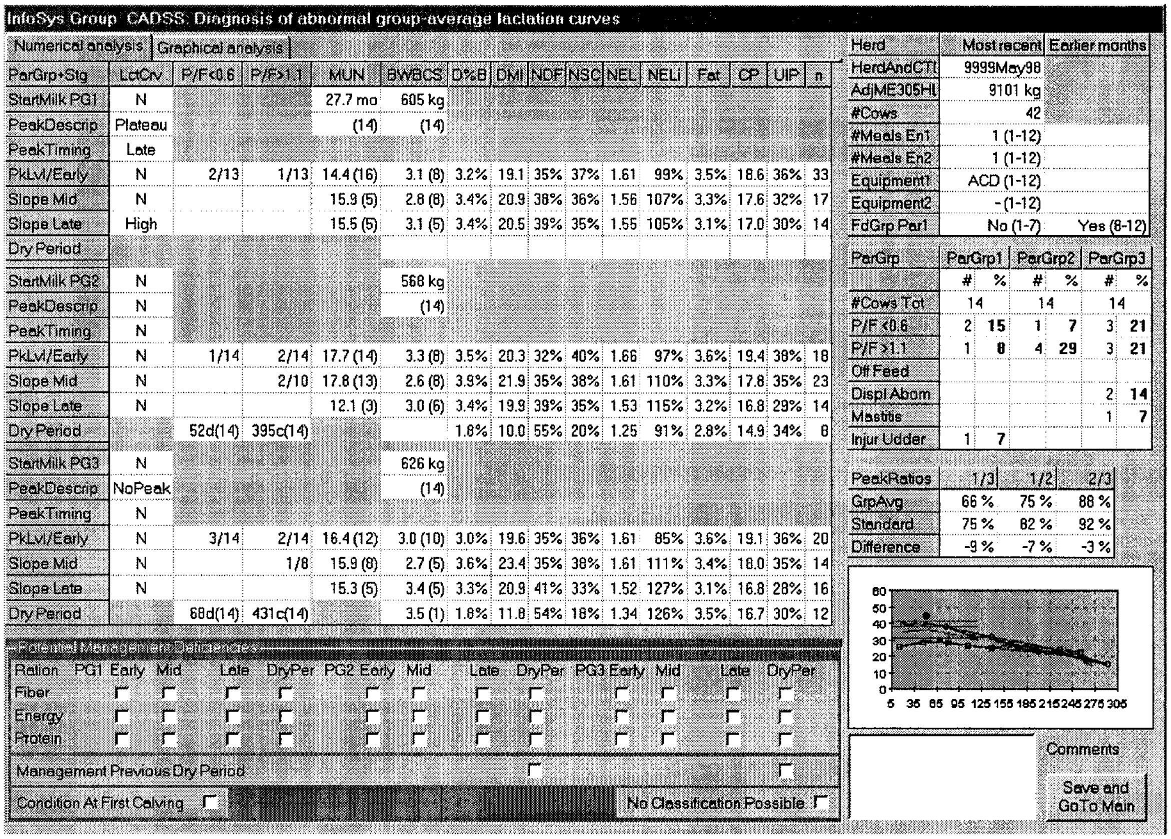
Table 4.2 Example cases of group-average peak levels described by a selection of potentially predictive attributes for machine learning and classified by a domain specialist as low, normal, or high.

Potentially predictive attributes			Peak level
GrpPL[1] – StdPL[2] (kg)	(GrpPL – StdPL) / StdPL (%)	(GrpPL – StdPL) / SDGrpPL[3]	
- 6.0	- 20	- 2.2	Low
- 4.1	- 9	- 0.8	Low
- 1.1	- 3	- 0.3	Normal
+ 0.9	+ 2	+ 0.2	Normal
+ 2.9	+ 10	+ 0.8	High
+ 1.6	+ 5	+ 0.4	High

[1] GrpPL = group-average peak level.
[2] StdPL = standard peak level.
[3] SDGrpPL = standard deviation of group-average peak level.

4.3.5 *Diagnosis of Abnormal Group-Average Lactation Curves (Module 3)*

The third and final module allows the user to compare the group-average lactation curve interpretations of the three parity groups with each other, and with additional group-average performance indices, and diagnose detected abnormalities through the identification of potential management deficiencies (Figures 4.5 and 4.6). The input for this module consists of: thirteen variables associated with each test within a lactation, nine variables associated with each lactation, standard lactation curves and standard peak-level values, the outlier results of the first module, and the group-average lactation curve data and interpretations from the second module. The module results in a maximum list of 40 potential management deficiencies related to 1) the general condition of the cows at first calving, 2) the overall management of the previous dry period for parity groups two and three, and 3) the fiber, energy, or protein aspects of the ration for early, mid, and late lactation, and the dry period of each of the three parity groups.



ParGrp+Stg	LctCrv	P/F<0.6	P/F>1.1	MUN	BWBCS	D%B	DMI	NDF	NSC	NEL	NELi	Fat	CP	UIP	n
StartMilk PG1	N			27.7 mo	605 kg										
PeakDescrip	Plateau			(14)	(14)										
PeakTiming	Late														
PkLvl/Early	N	2/13	1/13	14.4 (16)	3.1 (8)	3.2%	19.1	35%	37%	1.61	99%	3.5%	18.6	36%	33
Slope Mid	N			15.9 (5)	2.8 (8)	3.4%	20.9	38%	36%	1.56	107%	3.3%	17.6	32%	17
Slope Late	High			15.5 (5)	3.1 (5)	3.4%	20.5	39%	35%	1.55	105%	3.1%	17.0	30%	14
Dry Period															
StartMilk PG2	N				568 kg										
PeakDescrip	N				(14)										
PeakTiming	N														
PkLvl/Early	N	1/14	2/14	17.7 (14)	3.3 (8)	3.5%	20.3	32%	40%	1.66	97%	3.6%	19.4	38%	18
Slope Mid	N		2/10	17.8 (13)	2.6 (8)	3.9%	21.9	35%	38%	1.61	110%	3.3%	17.8	35%	23
Slope Late	N			12.1 (3)	3.0 (6)	3.4%	19.9	39%	35%	1.53	115%	3.2%	16.8	29%	14
Dry Period		52d(14)	395c(14)			1.8%	10.0	55%	20%	1.25	91%	2.8%	14.9	34%	8
StartMilk PG3	N				626 kg										
PeakDescrip	NoPeak				(14)										
PeakTiming	N														
PkLvl/Early	N	3/14	2/14	16.4 (12)	3.0 (10)	3.0%	19.6	35%	36%	1.61	85%	3.6%	19.1	36%	20
Slope Mid	N		1/8	15.9 (8)	2.7 (5)	3.6%	23.4	35%	38%	1.61	111%	3.4%	18.0	35%	14
Slope Late	N			15.3 (5)	3.4 (5)	3.3%	20.9	41%	33%	1.52	127%	3.1%	16.8	28%	16
Dry Period		68d(14)	431c(14)		3.5 (1)	1.8%	11.8	54%	18%	1.34	126%	3.5%	16.7	30%	12

Herd	Most recent	Earlier months
HerdAndCTI	9999May98	
AdjME305HI	9101 kg	
#Cows	42	
#Meals En1	1 (1-12)	
#Meals En2	1 (1-12)	
Equipment1	ACD (1-12)	
Equipment2	- (1-12)	
FdGrp Par1	No (1-7)	Yes (8-12)

ParGrp	ParGrp1		ParGrp2		ParGrp3	
	#	%	#	%	#	%
#Cows Tot	14		14		14	
P/F <0.6	2	**15**	1	**7**	3	**21**
P/F >1.1	1	**8**	4	**29**	3	**21**
Off Feed						
Displ Abom					2	**14**
Mastitis					1	**7**
Injur Udder	1	**7**				

PeakRatios	1/3	1/2	2/3
GrpAvg	66 %	75 %	88 %
Standard	75 %	82 %	92 %
Difference	-9 %	-7 %	-3 %

Potential Management Deficiencies

Ration	PG1 Early	Mid	Late	DryPer	PG2 Early	Mid	Late	DryPer	PG3 Early	Mid	Late	DryPer
Fiber	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Energy	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Protein	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Management Previous Dry Period								☐				☐

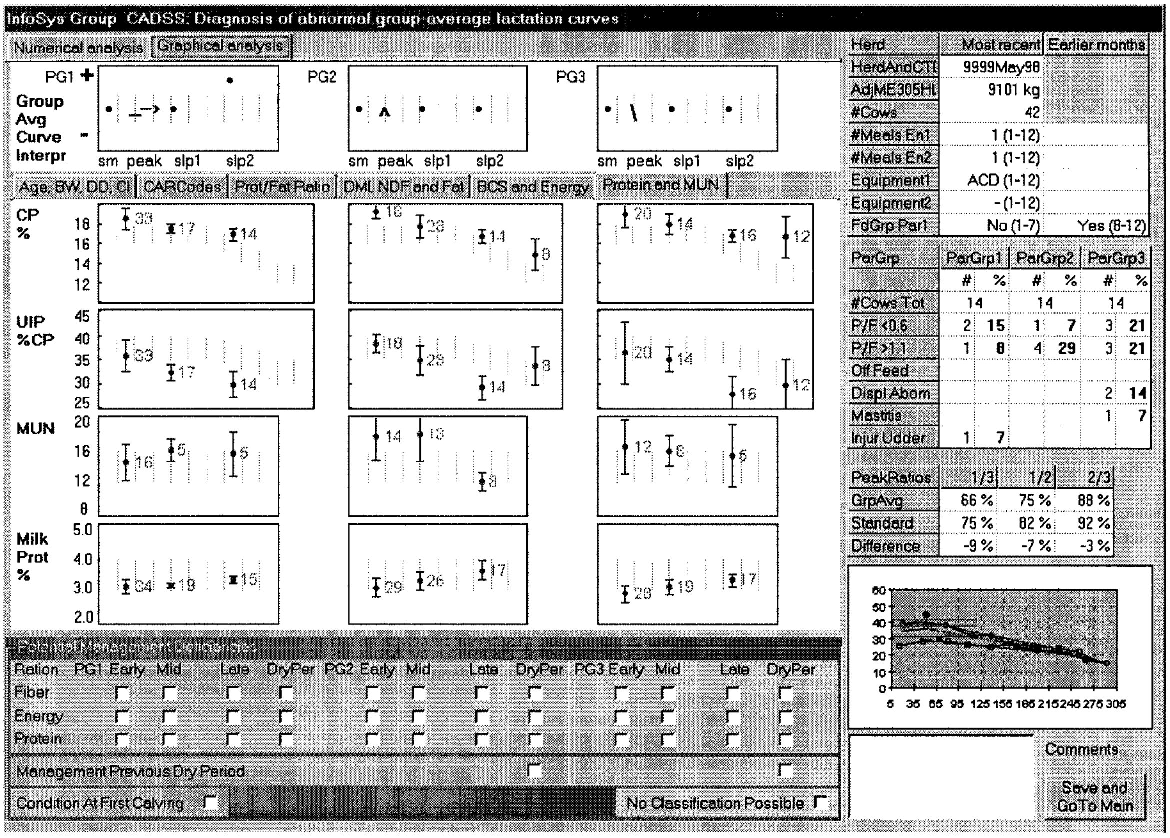
Condition At First Calving ☐

No Classification Possible ☐

Figure 4.5 Screen capture of module 3: diagnosis of abnormal group-average lactation curves using numerical analysis.

The main section of this module allows for the comparison of the group-average lactation curve interpretations with additional group-average performance indices for each of the three parity groups and for three stages of lactation and the dry period within each parity group. The user has two options to visualize this information: numerically as shown in Figure 4.5 or graphically as shown in Figure 4.6. The performance indices in this section include the proportion of cows with a low or high protein to fat ratio and group-average values for milk urea nitrogen, body condition score, and nine different descriptors of the ration. Also shown are the average age at first calving, body weight at calving for each parity group, and days dry and calving interval of the previous lactation for parity groups two and three. On the right of the main section (Figures 4.5 and 4.6), additional tables show information pertaining to the feeding system, percentage of cows with low or high protein to fat ratio and with specific event codes in each parity group, and peak ratios between parity groups one and two, one and three, and two and three. A graph shows the three group-average lactation curves and peaks simultaneously. As an intermediate analysis step before classifying the potential management deficiencies, the user can interpret and classify the group-average values of the numerical performance indices as "Normal", "High", and "Low" through successive mouse clicks on the cell in the table, showing the variable in question. These classification decisions prompt the program to change the color of the font in the cell successively from black ("Normal") to red ("High"), and blue ("Low"), and record the associated class using 176 fields in the herd table of the classification database. After analyzing the information shown in this module, the user can indicate potential management deficiencies that may explain detected problems with the group-average lactation curves and other abnormalities, by clicking on the check boxes in the bottom section of the module. Each one of these 40 potential management deficiency check boxes represents a classification task with the classes "True" and "False", and each decision is recorded in the classification database.

To support the analysis of the large amount of information shown in the main section, the numerical representation of the information was translated to a graphical format, using the visualization technique of multiple standardized graphs as described by Tufte (1997). With this approach each graph shows, for a particular parity group and performance index, the group-average mean with error bars representing plus and minus the standard deviation, plotted against days after calving (Figure 4.6). For each graph a normal range is indicated, allowing the user to quickly assess deviations of the group-average performance from normal expectations. Initial boundary settings were derived from the literature. However, the upper and lower level of the normal range in each graph may be fine-tuned through future machine learning of the decision boundaries between the “Normal” and “High” and the “Normal” and “Low” classes of each numerical performance index shown in the main section. Figure 4.6 shows a graphical representation of the group-average lactation curve interpretations for each parity group and a tab-sheet with graphs representing four performance indices related to milk and ration protein. The user can click on five additional tab-sheets to view graphs pertaining to the other performance indices shown in the main section.



Herd	Most recent	Earlier months
HerdAndCT(	9999May98	
AdjME305HL	9101 kg	
#Cows	42	
#Meals En1	1 (1-12)	
#Meals En2	1 (1-12)	
Equipment1	ACD (1-12)	
Equipment2	- (1-12)	
FdGrp Par1	No (1-7)	Yes (8-12)

ParGrp	ParGrp1		ParGrp2		ParGrp3	
	#	%	#	%	#	%
#Cows Tot	14		14		14	
P/F <0.6	2	15	1	7	3	21
P/F >1.1	1	8	4	29	3	21
Off Feed						
Displ Abom					2	14
Mastitis					1	7
Injur Udder	1	7				

PeakRatios	1/3	1/2	2/3
GrpAvg	66 %	75 %	88 %
Standard	75 %	82 %	92 %
Difference	-9 %	-7 %	-3 %

Potential Management Deficiencies

Ration	PG1 Early	Mid	Late	DryPer	PG2 Early	Mid	Late	DryPer	PG3 Early	Mid	Late	DryPer
Fiber	□	□	□	□	□	□	□	□	□	□	□	□
Energy	□	□	□	□	□	□	□	□	□	□	□	□
Protein	□	□	□	□	□	□	□	□	□	□	□	□
Management Previous Dry Period								□				□

Condition At First Calving □

No Classification Possible □

Figure 4.6 Screen capture of module 3: diagnosis of abnormal group-average lactation curves using graphical analysis.

A very large amount of variables were available to derive predictive attributes for each of the 40 classification tasks indicating whether or not a potential management deficiency existed. Table 4.3 shows example cases of the classification task to determine whether there is a potential problem with the management of the previous dry period of the second parity group cows. This table includes a selection of potentially predictive attributes for machine learning: startup milk, peak level, body condition score at calving for parity group two, the difference between the body condition score at calving for parity group two and the score at dry-off for parity group one, the percentage of cows in parity group two and early lactation with a low milk protein to fat ratio, and the ratio between peak levels of parity group one and two.

Table 4.3 Example cases of the classification of milk-recording data by a domain specialist to detect a potential problem with the management of the previous dry period in a group of second parity cows.

Potentially predictive attributes						Potential
Start-up milk P2[1]	Peak level P2	BCS[2] P2S1[3]	BCS P2S1 – BCS P1S4[4]	Cows with low protein to fat ratio P2S1 (%)	Relative peak level P1[5] over P2 (%)	problem
Low	Low	2.2	-0.9	16	86	True
Normal	Low	4.1	+0.8	28	96	True
Low	Normal	3.5	+0.5	12	80	True
Normal	Normal	3.9	+0.3	0	79	False
Normal	High	3.0	-0.6	12	76	False
Low	Normal	3.5	-0.2	8	83	False

[1] P2 = parity group two.
[2] BCS = body condition score.
[3] P2S1 = parity group two and stage one (early lactation).
[4] P2S4 = parity group two and stage four (dry period).
[5] P1 = parity group one.

4.4 Discussion

Successful application of machine-learning techniques to develop KBS requires proper formulation of the overall problem into classification tasks, conducive to machine learning (Langley and Simon, 1995, Verdenius et al., 1997) and the acquisition of an adequate number of example cases of sufficient quality (Adriaans, 1997). In this study, an iterative approach was developed to decompose the problem area of group-average lactation-curve analysis into sub-problems and classification tasks, and to create a software tool, the CADSS, to enable domain specialists to efficiently analyze and classify milk-recording data for a substantial number of example herds.

The process of analyzing group-average lactation curves turned out to be more complex than expected, involving multiple analysis steps, multiple views of performance representations, and a large degree of preprocessing of raw data. The case-acquisition tool resulting from this research can, therefore, be considered as a decision-support system, providing dairy producers and their advisors with a framework for the analyses of group-average lactation curves and automating the preprocessing of raw data into graphical and numerical performance representations. Once the program has been complemented with knowledge-based modules, generated through machine learning in a subsequent research project, field testing will be carried out to acquire input from the end-users regarding, e.g., the graphical user interface, the confidence they have in the knowledge-based modules, the amount of user-interaction required with each of the modules, and the costs and benefits of using the program to support dairy producers.

Although the development of the CADSS had been designed as an iterative process (Figure 4.1), many more iterations were required than initially anticipated until the specialists were satisfied with the resulting analysis approach and CADSS. After decomposing the overall problem into three sub-problems, it was difficult for the domain specialists to specify exactly how they wanted the large amount of raw data to be processed and represented in performance indices. The development, in itself, of the CADSS prototypes enabled the specialists to propose new ways of viewing and analyzing the data, not necessarily obvious or available to them before. In addition, the specialists were confronted with alternative views of the data proposed by the system developer, based on data visualization techniques. Many iterations were thus necessary to explore these new approaches to viewing and analyzing the data and to elicit the preferred method of group-average lactation-curve analysis, which was formalized in the resulting CADSS.

The focus on example cases was found to be useful throughout the development of the CADSS. The example-based approach facilitated communication between the system developer and the specialists during the consultation sessions in which the problem formulation and the functionality of the CADSS program were discussed. Additionally, the analysis of real-world example cases of milk-recording data with the CADSS prototypes helped the specialists to determine whether the chosen representations of the data were useful and if the classification tasks and classes adequately covered the sub-problems.

The CADSS developed here is expected to continue to evolve to higher levels of complexity in future research projects. The first step will be the acquisition from domain specialists of example cases of the removal of outlier tests and lactations. Machine-learning techniques will then be applied to these example cases to derive knowledge-based components that will be incorporated into the CADSS to automate the removal of outliers. This automation will make it easier for the domain specialists to analyze and classify a substantial number of example cases of the interpretation and diagnosis of group-average lactation curves using the second and third modules, while still allowing the user to override the system. The final KBS for use by dairy advisors could keep its case-acquisition functionality to allow for the acquisition of interesting new cases encountered in the field, which could be used for additional machine learning. In the long-term, machine-learning capabilities could be incorporated into the program, leading to a self-learning or adaptive KBS (Schmoldt, 1997).

Although the CADSS was developed to make use of the milk-recording data available at the provincial DHIA, the overall approach is expected to be applicable to any region and dairy support situation. The approach to group-average lactation-curve analysis presented in this paper is already possible with a basic set of milk-recording variables consisting of milk yield, fat, protein, and SCC. The case-acquisition functionality of the CADSS enables dairy nutrition specialists, familiar with the specific dairy farming situation, to analyze and classify milk-recording data of a selected number of dairy herds. This could be followed by the development of machine-learning generated knowledge-based components specific to that situation, which could be added to the core modules of the CADSS program.

The example-based approach to the development of decision-support systems, followed by case acquisition and machine learning to enhance these systems with knowledge-based components, may also be useful for other problem areas in dairy farming and agriculture in general. Information technology has made it possible to capture and store vast amounts of data from sources on the farm (e.g., sensors) as well as from external organizations. The challenge for agricultural producers is to interpret and utilize these data properly to improve decision-making (Doluschitz, 1990; Frost et al., 1997; Tomaszewski, 1993). Similar to the approach pursued in the research presented here, computerized support systems could be developed for the analysis and interpretation of data in these problem domains by making use of the expertise of domain specialists to explore new ways of analyzing the available data through example-based decision-support system development. Designing these systems to include case-acquisition functionality would allow for the acquisition of example cases, classified by domain specialists, and the application of machine-learning techniques to develop knowledge-based components which could be integrated into the developed decision-support system.

In conclusion, a CADSS was developed to support the analysis of group-average lactation curves and to enable domain specialists to analyze and classify example cases of this analysis process efficiently. In future research, machine-learning techniques will be used to discover and acquire the knowledge implicitly embedded in these example cases. The resulting knowledge-based modules will subsequently be incorporated into the CADSS and lead to a final KBS.

Preface to Chapter 5

The decision-support system for the analysis of group-average lactation curves described in Chapter 4 would greatly benefit from the inclusion of knowledge-based components to partially or completely automate the various decision-making tasks involved. Such components could, for example, automate the time-consuming removal of outlier tests and lactations of individual cows from group-average analysis based on the expertise of a domain specialist.

Machine learning has been successfully applied to real-world problems in many domains and may also be a useful approach to the development of knowledge-based components for the analysis of group-average lactation curves. However, several challenges remain, especially in contexts with limited availability of example cases, such as knowledge acquisition. With small data sets for learning, proper estimation of the classification performance with new data and comparison of the performance of different classifiers are difficult. Different approaches to this problem have been proposed in the literature, but additional research into these methodological aspects remains necessary. In addition, the classification performance achieved with machine learning may to a large extent be influenced by the type of prepocessing of the data and the proper tuning of the parameters of the machine-learning algorithm being used.

This chapter focuses on the use of decision-tree induction to acquire the domain knowledge involved in the filtering of lactations of individual cows for group-average lactation-curve analysis. This classification task is part of the "removal of outliers" module of the decision-support system described in Chapter 4, which represented the first step in the analysis of group-average lactation curves. In addition, this chapter describes a methodology for the analysis of the performance of classifiers generated through machine learning from small data sets and an approach to support the evaluation of the plausibility of induced decision trees. Experiments were carried out to determine the appropriate type of preprocessing of the data and configuration of the decision-tree induction algorithm. A series of decision trees was induced from the available data for implementation as knowledge-based components to automatically filter lactation curves in a decision-support system for the analysis of group-average lactation curves.

This chapter was submitted to the journal Computers and Electronics in Agriculture (Pietersma, D., R. Lacroix, D. Lefebvre, and K. M. Wade. Performance analysis of machine-learning induced decision trees for lactation-curve analysis).

5 Performance analysis of machine-learning induced decision trees for lactation-curve analysis

Abstract

Machine learning has been identified as a promising approach to knowledge-based system development. However, challenges, such as analysis of the performance achieved through learning from small data sets, remain. This study focussed, firstly, on the use of decision-tree induction for knowledge acquisition to filter lactations of individual cows for group-average lactation curve analysis, and, secondly, on the application of graphical and statistical techniques to analyze the results of machine learning. Data consisted of 1428 cases classified by a dairy-nutrition specialist as outlier (34 cases) or non-outlier. The performance of decision trees, induced from the entire data set, was estimated through ten-fold cross validation. Relative operating characteristic curves were used to visualize the achieved trade-offs between correctly classifying positive and negative cases. A performance index, representing the mean true positive rate of these curves for a limited range of false positive rate values, was developed to facilitate comparison among classification schemes. Analysis of variance was used to determine whether real differences existed for the expected performance on new data among the different combinations of data preprocessing and algorithm configurations evaluated in this study. In terms of data preprocessing, random assignment of herds to the folds of the cross validation did not perform significantly different from assigning cases to folds, while use of a special value to indicate attribute values that were irrelevant for the case in question significantly improved the performance over treating these values as missing. Tuning of the configuration of the decision-tree induction algorithm significantly improved the classification performance. Three final decision trees were induced from the entire data set. Their expected true positive rates were 52%, 68%, and 92%, at false positive rates of 1.5%, 3.5%, and 8.6%, respectively. However, due to the low prevalence of outlier lactations (cases), this performance was associated with many false positives. The specialist reviewed the final trees and adjusted two decision nodes. This study suggests that decision-tree induction has a role to play in the acquisition of knowledge involved in

the removal of outlier lactations. In addition, the application of ten-fold cross validation in combination with relative operating characteristic curves and analysis of variance was found to be useful in analyzing the results of machine learning from small data sets.

5.1 Introduction

Dairy producers who are enrolled in a milk-recording program, receive a large amount of data that can be used to improve dairy herd management. The analysis of group-average lactation curves, generated from such data, has been identified as a useful tool to support tactical-level nutrition management in dairy farming (Whittaker et al., 1989; Lefebvre et al., 1995; Skidmore et al., 1996). This type of analysis involves comparison of group-average curves with standard curves, and the analysis of additional explanatory data, that may lead to the detection of potential management deficiencies. Group-average lactation curve analysis tends to be lengthy and complex, and the use of a knowledge-based system (KBS) seems an obvious avenue of exploration. Such a system would automate the preprocessing of the large amount of raw data involved and provide dairy producers and their advisors with expert interpretation (Whittaker et al., 1989).

Traditionally, KBS have been developed based on interviews with domain experts, sometimes supplemented by other sources of knowledge such as documentation (Durkin, 1994; Dhar and Stein, 1997). However, the acquisition of knowledge through interviews has proven to be time-consuming and difficult. Experts often have difficulty expressing how they reason and make their decisions and, in addition, it is not easy to structure and encode the knowledge expressed through interviews into a representation that can be used as part of a KBS. Alternatively, acquisition of knowledge can be partially automated with machine learning (Langley and Simon, 1995; Dhar and Stein, 1997). With this approach, a domain expert first classifies example cases of a particular problem. Then a machine-learning technique, such as decision-tree induction, is used to learn how to classify new cases based on these examples. Machine learning may accelerate the knowledge-acquisition process (Dhar and Stein, 1997) and potentially lead to a more accurate representation of the expert's actions (Michalski and Chilausky, 1980; Ben-David and Mandel, 1995). However, only a few agricultural examples of machine-learning assisted knowledge acquisition were found in the literature. These included the application of rule induction to develop an expert system for soybean disease diagnosis (Michalski and Chilausky, 1980) and the use of decision-tree induction to support the creation of a KBS for tomato crop management in greenhouses (Mangina et al., 1999). Since machine-learning assisted knowledge acquisition has shown promising results in multiple domains, it should also be applicable to dairy farming and, specifically, to the complex area of lactation curve analysis.

Although machine learning may solve some of the difficulties associated with the traditional interview approach to knowledge acquisition, new challenges also arise. Firstly, machine learning can only take place if example cases of the problem at hand are available. In the context of knowledge acquisition the domain specialist may need to analyze and classify an adequate number of example cases for the specific KBS being developed (Pietersma et al., 2001a). Another difficulty is that the classification performance achieved with learning may be influenced by the type of preprocessing of the data set (Kubat et al., 1998; Witten and Frank, 2000) and the proper tuning of the algorithm parameters (Henery, 1994; Verdenius et al., 1997). This means that several machine-learning experiments may be required before achieving satisfactory results. An additional challenge is the analysis of the performance achieved with machine learning in experiments with small data sets. Different approaches to performance estimation and comparison of results have been explored (see, for example, Weiss and Kulikowski, 1991; Mitchell, 1997; Dietterich, 1998; Provost et al., 1998), but further study into the methodological aspects of performance analysis with small data sets remains necessary (Witten and Frank, 2000). The combination of these factors may explain the lack of attention for machine-learning assisted knowledge acquisition in dairy farming, and agriculture in general.

Research was initiated to explore the use of machine learning to support the development of a KBS for the analysis of group-average lactation curves. A case-acquisition and decision-support system (CADSS) was developed previously (Pietersma et al., 2001a) to enable domain specialists to analyze example cases of the analysis of group-average lactation curves and to capture their subsequent classifications. The project described in this current paper had two main objectives: firstly to develop a knowledge-based module to filter lactations of individual cows automatically for group-average lactation curve analysis through decision-tree induction, and secondly to explore the use of graphical and statistical techniques to analyze the results of machine-learning experiments with small data sets. Specific objectives were 1) to determine the appropriate type of preprocessing of the data set and configuration of the decision-tree induction algorithm, 2) to assess the ability of induced decision trees to discriminate between outlier and non-outlier lactation curves, and 3) to verify the plausibility of the induced decision trees.

5.2 Methods

The approach used in this study involved five sequential steps, indicated by the gray area and labeled with the numbers three through seven in Figure 5.1: classification of example cases by a domain specialist using a case-acquisition tool; analysis and

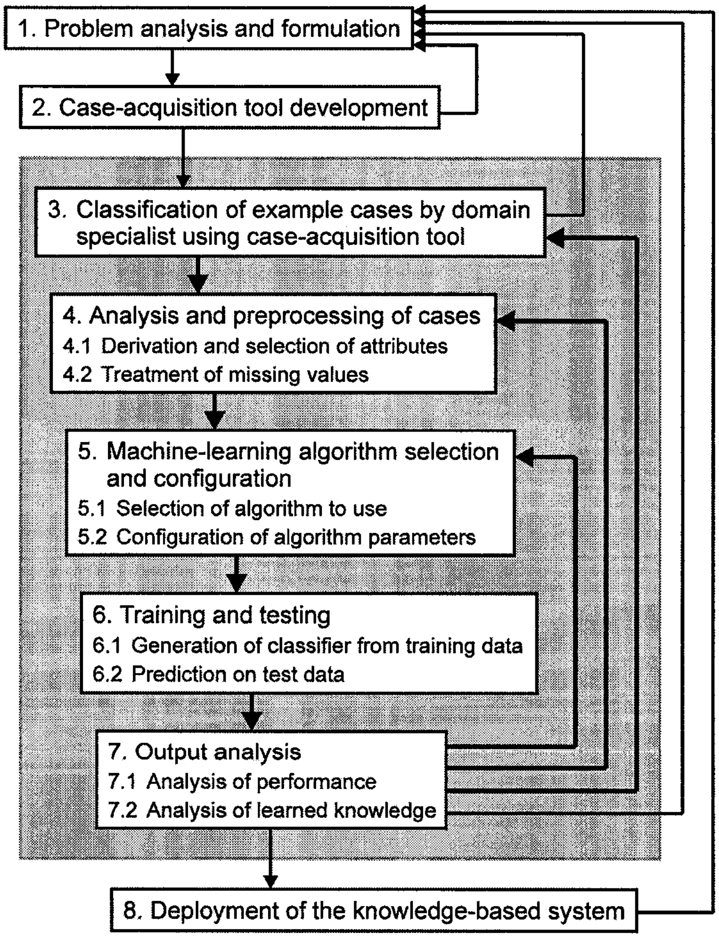
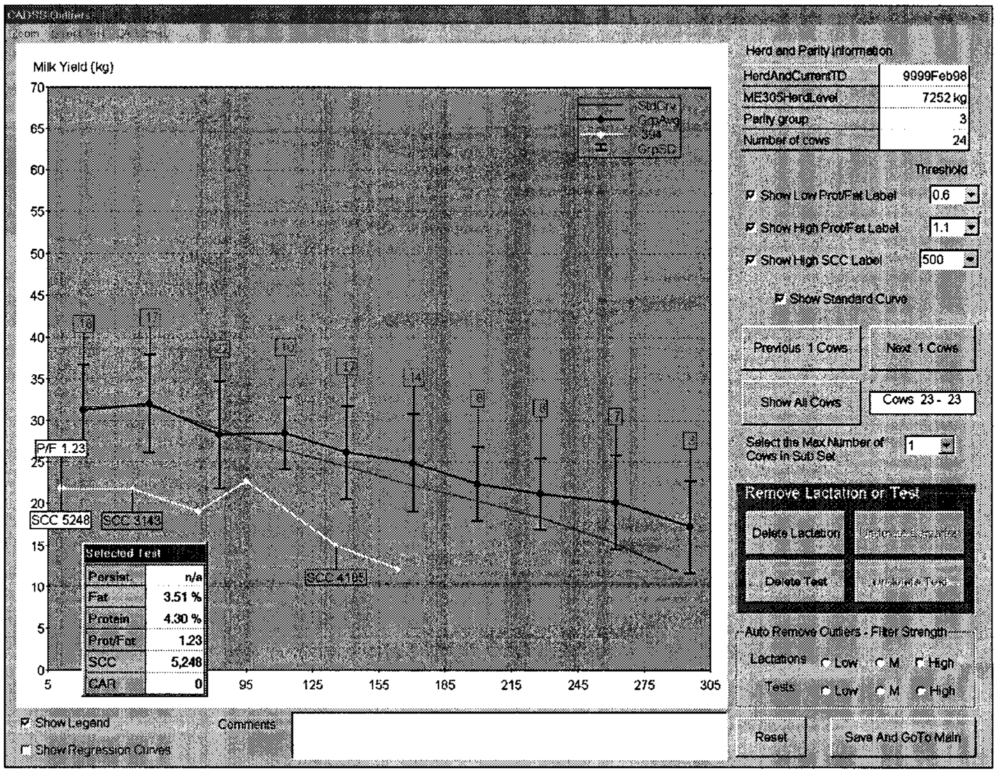

1. Problem analysis and formulation
2. Case-acquisition tool development
3. Classification of example cases by domain specialist using case-acquisition tool
4. Analysis and preprocessing of cases
4.1 Derivation and selection of attributes
4.2 Treatment of missing values
5. Machine-learning algorithm selection and configuration
5.1 Selection of algorithm to use
5.2 Configuration of algorithm parameters
6. Training and testing
6.1 Generation of classifier from training data
6.2 Prediction on test data
7. Output analysis
7.1 Analysis of performance
7.2 Analysis of learned knowledge
8. Deployment of the knowledge-based system

Figure 5.1 A process model for machine learning to support knowledge acquisition for knowledge-based system development (steps 1 and 2 are described in Pietersma et al., 2001a).

preprocessing of the acquired example cases; machine-learning algorithm selection and configuration; training and testing; and, finally, analysis of the results of machine learning. This approach to machine-learning assisted knowledge acquisition was adapted from different methodologies for the application of machine learning in general (Langley and Simon, 1995; Brodley and Smyth, 1997; Verdenius et al., 1997). In the following sections the five steps involved in this research are described in detail.

5.2.1 Acquisition and processing of example cases

Example cases of the removal of outlier lactations were generated by a dairy nutrition specialist using the CADSS for analysis of group-average lactation curves (Pietersma et al., 2001a). The removal of outliers was identified as the first step in the overall analysis process and considered important to avoid the interpretation of the group-average performance being biased by a few atypical lactations or tests. With the CADSS the specialist was able to compare the lactation curves of individual cows belonging to one of three parity groups with group-average and standard lactation curves (Figure 5.2).



Selected Test

Persist.	n/a
Fat	3.51 %
Protein	4.30 %
Prot/Fat	1.23
SCC	5,248
CAR	0

Herd and Parity information

HerdAndCurrentTD	9999Feb98
ME305HerdLevel	7252 kg
Parity group	3
Number of cows	24

Threshold
☑ Show Low Prot/Fat Label 0.6
☑ Show High Prot/Fat Label 1.1
☑ Show High SCC Label 500
☑ Show Standard Curve

Cows 23 - 23

Select the Max Number of Cows in Sub Set 1

Auto Remove Outliers - Filter Strength
Lactations ○ Low ○ M ○ High
Tests ○ Low ○ M ○ High

☑ Show Legend
☐ Show Regression Curves
Comments

Figure 5.2 Screen capture of the case-acquisition software module used to remove outlier tests and lactations.

Additional attributes available with the CADSS are listed in Table 5.1 and included, for each test of a lactation, milk protein to fat ratio, somatic cell count (SCC), codes indicating conditions affecting records (CAR) such as clinical mastitis or estrus, and a regression equation for the lactation curve proposed by Wilmink (1987). Details of the functioning of this CADSS can be found in Pietersma et al. (2001a). The domain specialist analyzed a data set consisting of the milk-recording data from 33 Holstein herds enrolled with the provincial dairy herd analysis service and representing a wide range of rolling herd-average milk production levels. The data set contained 1428 lactations of which 34 (2.4%) were classified as outlier.

Creating an effective representation of the data has been identified as an important step in the successful application of machine-learning techniques to real-world problems

Table 5.1 Listing of attributes describing example cases of the removal of outlier lactations available to the domain specialist (D) and used for machine learning (M).

	Attribute description
D	Days in milk (DIM), milk yield, percent fat, percent protein, somatic cell count, and conditions affecting records (CAR) code for each test of lactation curve
D	DIM, milk, standard deviation, and number of tests for each stage of group-average lactation curve
D	DIM and milk for each stage of standard lactation curve
D M	Regression equation for lactation curve $Y(t) = a + b\,t + c\exp(-0.05\,t)$ with parameters a, b, and c
D M	Parity and Parity group
D M	Number of tests in lactation
D M	Number of lactations in parity group
D M	Number of tests in parity group
D M	Number of lactations in herd
D M	Number of tests in herd
D M	Average mature equivalent 305-day milk production of the herd
M	Number and percentage of low and of high protein to fat ratio tests
M	Average protein to fat ratio
M	Number and percentage of high somatic cell count tests
M	Average somatic cell count
M	Average somatic cell linear score
M	Number and percentage of tests with any CAR code
M	Any test with CAR code abortion
M	Any test with CAR code foot rot
M	Any test with CAR code off feed during early lactation
M	Any test with CAR code abortion, milk fever, metritis, or displaced abomasum
M	Average absolute and relative deviation from group-average lactation curve
M	Average number of standard deviations (SD) from group-average lactation curve
M	Slope linear regression through entire curve and deviation from slope standard curve
M	Slope linear regression through tests after peak and deviation from slope standard curve
M	Min. slope linear regression through three consecutive tests and deviation from slope standard curve
M	Max. slope linear regression through three tests after peak and deviation from slope standard curve
M	Max. absolute and relative deviation of a test from line between a test before and a test after that test
M	Avg. SD of group-average lactation curve stages covered by lactation relative to stages not covered

(Langley and Simon, 1995). Thus, to support machine learning, additional attributes were constructed from the basic attributes available to the domain specialist. Such derived attributes included, for example, the mean SCC for the lactation, and the mean absolute and relative deviation from the group average lactation curve. An initial set of potentially predictive attributes was proposed by the system developer and additional attributes were suggested by the domain specialist, leading to a total of 41 attributes for machine learning (Table 5.1).

The resulting data set contained a substantial number of records with missing attribute values, some of which were missing due to errors in the data acquisition procedure (i.e., during milk recording) and considered as unknown. These unknown values were left blank as required by the machine-learning algorithm. The remaining missing attribute values were in fact irrelevant for the cases they described. Irrelevant attribute values occurred, e.g., with the attribute "slope after the peak" when a lactation did not include tests after the peak. Witten and Frank (2000) suggested that the type of treatment of irrelevant attribute values may have an impact on the achieved classification performance in machine learning. Thus, an experiment was carried out to investigate the effect of treating irrelevant attribute values as either unknown or with a special value beyond the range of possible values for the attribute.

5.2.2 Configuration of the algorithm

The decision-tree induction algorithm used in this study was CART - Classification And Regression Trees - for Windows version 3.6 developed by Salford Systems (Breiman et al., 1984; Steinberg and Colla, 1997). The algorithm learns in a top-down fashion, by splitting the training data recursively into two smaller subsets, choosing, at each split, the attribute that is most successful in discriminating among the classes of the classification problem. The CART algorithm induces a maximum tree which is pruned back to avoid overspecialization of the training data. The resulting decision tree consists of a series of decision nodes that, during classification, guide each new case to a leaf node indicating the predicted class. With CART the user can control many parameters affecting the type of decision tree induced and the classification performance achieved. More details regarding the CART algorithm are given in Appendix A, while a thorough description can be found in Breiman et al. (1984) and Steinberg and Colla (1997).

In this research, several iterations of algorithm configuration, training and testing, and analysis of the performance (steps five through seven in Figure 5.1) were used to tune the algorithm parameters to the type of data available. Based on preliminary experiments, three parameters were considered important for tuning: splitting and pruning criterion, minimum size of child nodes, and prior probability of outlier lactations. The effect of these parameter configurations on the classification performance was studied in combination with the different methods of data preprocessing: this is explained in detail in the experimental design section below.

5.2.3 Training and testing

To determine the appropriate type of data preprocessing and algorithm configuration, a comparison of the performance of the resulting decision trees was required. In addition, an estimate of the performance of the decision trees was needed to assess the ability to discriminate between outlier and non-outlier lactation curves. The standard approach to evaluating the performance of a classifier, derived through machine learning, involves training and testing on separate data sets, which is necessary, since machine-learning algorithms tend to overfit the training data (Weiss and Kulikowski, 1991). The apparent performance on the training data - also called resubstitution performance (Witten and Frank, 2000) - may thus be much better than the performance achieved when the classifier is used in the real world to classify entirely new data. However, in this research the size of the data set was rather small considering there were only 34 outlier lactations. In the case of small data sets the number of example cases is likely to be a factor limiting the classification performance (Cohen, 1995; Witten and Frank, 2000). Therefore, with limited data, all available labeled example cases should be used to train a final classifier for a real-world application (Witten and Frank, 2000).

To estimate the performance of classifiers generated from the entire data set of example cases, the stratified ten-fold cross validation approach to training and testing was used (Breiman et al., 1984; Weiss and Kulikowski, 1991; Witten and Frank, 2000). With this approach the entire data set is divided into ten mutually exclusive subsets or folds with approximately the same class distribution as the original data set. Each fold is used once to test the performance of the classifier, generated from the combined data of the remaining nine folds, leading to ten independent performance estimates. Assuming that the classification performance improves as more data are used for learning, the true performance of the classifier generated from the entire labeled data set is expected to be at least as good as the ten-fold cross validation estimate which is based on classifiers generated from 90% of the data.

The use of ten-fold cross validation requires the random assignment of example cases to the ten folds. However, when the example cases are grouped in batches (i.e., herds in this project), random assignment of cases to folds means that cases of the same batch will be distributed over multiple folds and end up in both the training and test sets. This may lead to a biased estimate of the performance since the final classifier is to be used on example cases belonging to entirely new batches (Kubat et al., 1998). An experiment was, therefore, performed to explore the effect of assigning example cases to folds at either the case or the herd level. Random assignment of the outlier and non-outlier cases to folds resulted in 142 or 143 cases per fold of which 3 or 4 were outliers. Entire herds were manually assigned to folds to achieve approximately the same class distribution in each fold, resulting in 126 to 223 cases per fold of which 3 to 6 were outliers (Figure 5.3). Assignment of example cases to folds at either the case or the herd level was considered as a type of data preprocessing and investigated together with the treatment of irrelevant attribute values and the different algorithm configurations.

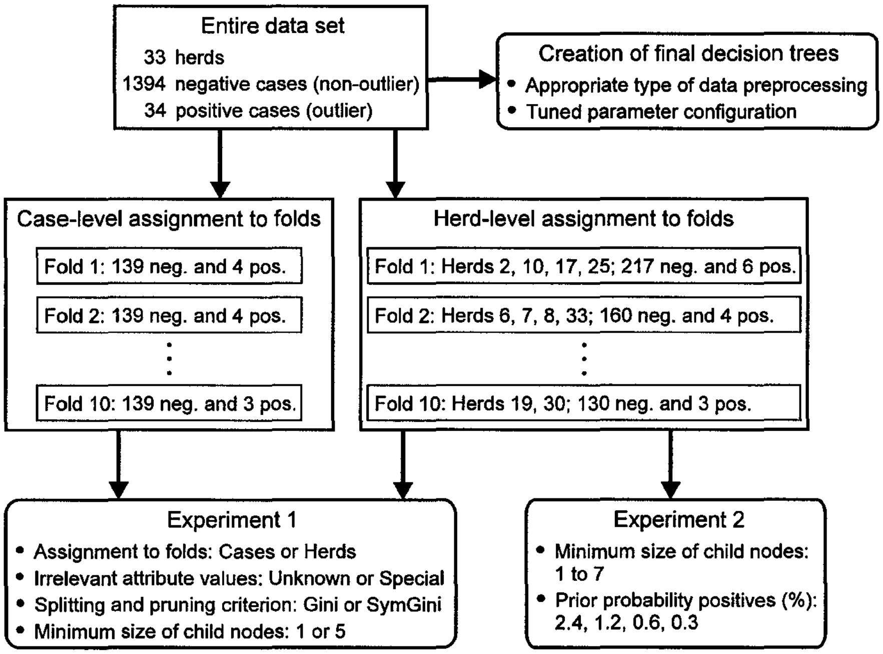

Entire data set
33 herds
1394 negative cases (non-outlier)
34 positive cases (outlier)
Creation of final decision trees
• Appropriate type of data preprocessing
• Tuned parameter configuration
Case-level assignment to folds
Fold 1: 139 neg. and 4 pos.
Fold 2: 139 neg. and 4 pos.
Fold 10: 139 neg. and 3 pos.
Herd-level assignment to folds
Fold 1: Herds 2, 10, 17, 25; 217 neg. and 6 pos.
Fold 2: Herds 6, 7, 8, 33; 160 neg. and 4 pos.
Fold 10: Herds 19, 30; 130 neg. and 3 pos.
Experiment 1
• Assignment to folds: Cases or Herds
• Irrelevant attribute values: Unknown or Special
• Splitting and pruning criterion: Gini or SymGini
• Minimum size of child nodes: 1 or 5
Experiment 2
• Minimum size of child nodes: 1 to 7
• Prior probability positives (%): 2.4, 1.2, 0.6, 0.3

Figure 5.3 Data treatment for experiments and final decision trees.

5.2.4 Performance analysis

In machine-learning literature, the performance of a classifier is often expressed as the overall error rate or as the overall accuracy with the implicit assumption that all types of misclassification are of equal cost and all types of correct classification are of equal benefit. However, in many classification problems, one type of misclassification may be considered less acceptable than another. Detailed analysis of the types of misclassifications leading to the achieved accuracy is thus often required. For example, in this study, a classifier that indicates each lactation as non-outlier would have an expected accuracy above 97.5%. However, this classifier would be completely useless for the classification task at hand since it would not remove any outliers. For the application in this study, with a very low prevalence of outliers, mistakenly classifying a non-outlier case as an outlier may, in fact, be less costly than misclassifying an outlier case as non-outlier.

The removal of outlier lactations is an example of two-class classification: a lactation was classified by the domain specialist as either outlier (positive) or non-outlier (negative). To distinguish among the different outcomes for two-class classification problems, a 2 x 2 contingency table or confusion matrix can be used (Swets, 1988; Weiss and Kulikowski, 1991; Witten and Frank, 2000). Figure 5.4 shows such a matrix in which the possible outcomes are denoted with A, B, C, and D. In this matrix, true positives (A) and true negatives (D) are correct classifications, a false positive (B) is an actual negative case incorrectly predicted as positive, and a false negative (C) is an actual positive case predicted as negative. These false positives and false negatives are equivalent to the concept of Type I and Type II errors used in statistics (Steel and Torrie, 1980).

	Actual class is positive	Actual class is negative	Total (predicted)
Predicted as positive	A	B	A + B
Predicted as negative	C	D	C + D
Total (actual)	A + C	B + D	A + B + C + D = N

TP rate = True positive rate = A / (A + C)
FP rate = False positive rate = B / (B + D)
PVP = Predictive value positive = A / (A + B)
PPR = Positive prediction rate = (A + B) / N

Prevalence of positive cases = (A + C) / N

Figure 5.4 Confusion matrix, performance indices, and prevalence for two-class classification.

From the 2 x 2 confusion matrix, several performance indices can be derived and four of these were used in this study: 1) the true positive rate (TP rate) or sensitivity, defined as A / (A + C) (Swets, 1988); 2) the false positive rate (FP rate) or 1 – specificity, defined as B / (B + D) (Swets, 1988); 3) the predictive value positive (PVP), defined as A / (A + C) (Weiss and Kulikowski, 1991); and 4) the quantity (A + B) / (A + B + C + D) (Swets, 1988), here referred to as the positive prediction rate (PPR). The prevalence of positive cases or prior probability of positives can also be determined from the confusion matrix as (A + C) / (A + B + C + D) (Swets, 1988). The TP rate and FP rate are both independent of the prevalence of positive cases and, thus, the characteristics of the classifier (Swets, 1988). Conversely, the PPR and PVP depend on the prevalence of positive cases and can be mathematically derived from the TP rate and FP rate for a given prevalence level using

$$PPR = \textit{Prevalence of positives} \times \textit{TP rate} + (1 - \textit{Prevalence of positives}) \times \textit{FP rate}$$

and

$$PVP = \textit{Prevalence of positives} \times \textit{TP rate} / PPR.$$

Machine-learning algorithms can generally be tuned to focus more on sensitivity and less on specificity, or vice versa. For example, with classification schemes that predict a continues value in the range from zero to one, such as artificial neural networks, a classifier with a particular trade-off between sensitivity and specificity can be generated by setting the cutoff value, or decision criterion, to a decimal number between zero and one above which cases are predicted as positive (Yang et al., 1999). With the CART algorithm, such a classifier can be generated through adjustment of the misclassification costs followed by the induction of a cost-specific decision tree (Breiman et al., 1984; Steinberg and Colla, 1997). Performance analysis in machine learning should thus investigate the performance profile of sensitivity and specificity combinations, achievable with a particular machine-learning algorithm, instead of focussing on one particular trade-off between sensitivity and specificity (Provost et al., 1998).

The entire range of achievable trade-offs between sensitivity and specificity can be visualized with "relative operating characteristic" (ROC) curves (Swets, 1988), consisting of the TP rate plotted against the FP rate (Figure 5.5). The lower left point (0,0) represents a classifier that assigns each case to the negative class, while the upper right point (100,100) represents a classifier that considers each case as positive. The upper left point (0,100) represents perfect classification, while the line y = x represents a random classification scheme with the probability of classifying a case as positive ranging from 0 to 1. The ROC curve of a particular classification system X is thus expected to be positioned above the random classification line (Figure 5.5). The closer the curve approximates the lines connecting (0,0) with (0,100) and (100,100), the better the performance. Each point on the ROC curve represents a classifier with a particular trade-off between sensitivity and specificity. For example, classifier X1 in Figure 5.5 has 80% TP rate at 10% FP rate. Given 2.4% prior probability of positive cases, this classifier would be expected to classify 12% of all cases as positive (PPR), while only 16% of those cases predicted as positive would be true positives (PVP). Classifier X2 has a higher TP rate (86%), but this is achieved at a much higher FP rate (20%), and would result in 22% PPR and 10% PVP.

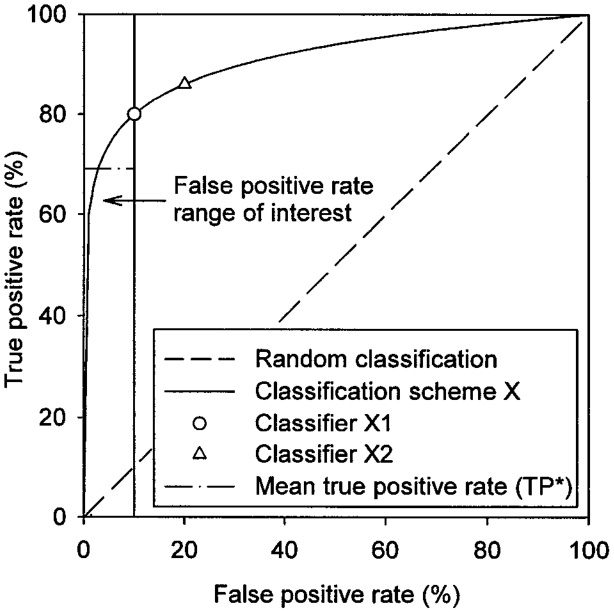


Figure 5.5 Relative operating characteristic (ROC) curve representing random classification and for classification scheme X, specific classifiers X1 and X2, and mean true positive rate (TP*) for scheme X covering the false positive rate range of interest.

Comparison of the performance of multiple classification schemes with statistical tools requires the information represented by the ROC curve to be collapsed into a single response variable. To this end, the area under the entire ROC curve was proposed as a suitable performance index by Swets (1988) and used in several machine learning studies (Bradley, 1997; Yang et al., 1999). However, the classification task to remove outlier lactations involved a highly unbalanced class distribution: only 2.4% of the cases had been labeled as positive (outlier) by the specialist. The range of FP rate values of the ROC curve was, therefore, limited to 10% at most. Classifiers with higher FP rates would lead to unrealistically high PPRs (i.e., predict too many lactations as outlier). Thus, instead of using the area under the entire ROC curve, a new performance index (TP*) was developed. This TP* was defined as the mean TP rate for the range of FP rates from 0% to 10% (Figure 5.5). Perfect classification would be represented by a TP* value of 100%, while random classification would lead to an expected TP* of 5%.

To generate an ROC curve for a particular combination of data preprocessing and algorithm configuration, a series of ten decision trees was generated using ten different values for the parameter of CART specifying the cost of false positives (ranging from, e.g., 5 to 120), while the cost of false negatives was kept at 1. The specific misclassification cost values were chosen to achieve ten classifiers covering the entire range of sensitivity and specificity trade-offs. For the different levels of the prior probability of outlier lactations parameter, different ranges of misclassification cost values were used to adjust for the effect of the prior class probability on the sensitivity and specificity trade-off. However, for a particular combination of data preprocessing and algorithm configuration, the same ten misclassification cost values were used for each of the ten folds of the cross validation. The ten misclassification cost levels combined with ten-fold cross validation resulted in 100 pairs of FP rate and TP rate values for each combination of data preprocessing and algorithm configuration. For each fold, an ROC curve was generated through interpolation between the FP rate and TP rate data points associated with the increasing misclassification costs. An average ROC curve was then calculated by taking the average of the ten interpolated TP rate values for each FP rate from 0 to 100% at 1% intervals (Provost et al., 1998). For each of the ten ROC curves, TP* was calculated, resulting in ten independent estimates of TP*. This allowed for the calculation of a standard error associated with the mean TP* estimate and the use of statistical tests. The ROC curves shown in this paper were all calculated using this averaging approach. An approximate 95% confidence interval of the TP rate of the average ROC curves was determined as the standard error of the ten interpolated estimates of the TP rate for a particular level of FP rate multiplied by a t-value of 2.26 (two-sided probability level of 95% and 9 degrees of freedom). A second method to averaging the 100 data points in ROC space was used to determine the classification performance of a final classifier generated from the entire data set and will be explained in the section describing the analysis of the learned knowledge.

5.2.5 Experimental design and analysis of variance

Two experiments were carried out to determine the appropriate type of data preprocessing and algorithm configuration (Figure 5.3). In the first experiment, two types of assignment of example cases to folds were studied in combination with two types of treatment of irrelevant attribute values and two types of splitting and pruning. To broaden the validity of the results of the experiment, the eight combinations of data preprocessing and algorithm configuration were repeated for two levels of the parameter for minimum size of child nodes. The parameter for the prior probability of outlier lactations was set to the observed probability in the data set (2.4%), which was considered as default. Thus, the first experiment involved a total of 16 distinct combinations of data preprocessing and algorithm configuration.

After deciding on the appropriate type of data preprocessing and configuration of the splitting and pruning parameter, a second experiment was designed to study the effects of seven configurations of the minimum child size parameter (1 - being the default - through 7) in combination with four configurations of the prior probability of outlier lactations parameter (2.4%, 1.2%, 0.6%, and 0.3%). Based on the results of the first experiment, data preprocessing was fixed at assigning herds to folds and special-valued irrelevant attribute values, and SymGini was chosen as the splitting and pruning criterion. Thus, the second experiment involved a total of 28 distinct combinations of algorithm configuration.

For both experiments, analysis of variance (Steel and Torrie, 1980; Cohen, 1995) was used to assess whether observed differences of TP* among the types of data preprocessing and algorithm configuration were likely to hold for new data and not achieved by chance. The analysis of variance (ANOVA) procedure allows one to analyze multiple factors of interest simultaneously and investigate the interactions between them. In addition, it is able to account for variability in the results due to other known factors such as the folds in the experiments described here. An important assumption associated with the use of ANOVA is that the observations for each combination of factors being studied are independent of each other (Steel and Torrie, 1980; Cohen, 1995). With ten-fold cross validation this requirement is met since the ten test sets are mutually exclusive: each example case is used only once for testing. However, with ten-fold cross-validation, the ten training sets are slightly different from each other since each pair of training data differs by 1 in 9 (or 11%) of the data. The ten observed performance values on the test sets are thus estimates of the performance of ten different classification schemes, generated from different training sets. Each of these training sets contains 90% of the available data, while the objective is to estimate the performance of classifiers, generated from the entire data set. In the agricultural domain this would be analogous to using the results of fertilizer trials on ten related varieties of corn, grown at ten randomly selected locations in a region (representing the test sets), to determine the optimum amounts of fertilizer to use on a new, related, variety of corn. In that situation, the assumption is made that the optimum level of fertilizer, determined for the ten tested varieties, will also hold for the new variety. Similarly, with ANOVA on ten-fold cross-validation data, the assumption is made that the detected differences in classification performance, determined with classifiers generated from the ten different training sets, will also hold for classifiers trained on the entire data set. This assumption was considered reasonable since each training set consists of 90% of the entire data set.

In this study, ANOVA was performed with the Mixed procedure of SAS for Windows version 8 (SAS Institute Inc., Cary, NC). The observations for TP* in the first experiment were described by the following model:

$$TP^{*}_{ijklm} = \mu + FA_i + Fold_{ij} + IA_k + SP_l + MC_m + \textit{interactions among fixed effects} + e_{ijklm}$$

where TP* = dependent variable, μ = overall mean, FA_i = fixed effect of fold assignment (i = cases or herds), $Fold_{ij}$ = random effect of fold within fold assignment level i (j = 1 to

10), IA_k = fixed effect of treatment of irrelevant attribute values (k = unknown or special), SP_l = fixed effect of splitting and pruning criterion (l = Gini or SymGini), MC_m = fixed effect of minimum size of child nodes (m = 1 or 5), and e = random residual term. The model for the second experiment was as follows:

$$TP^*_{ijk} = \mu + MC_i + PP_j + MC_i \cdot PP_j + Fold_k + e_{ijk}$$

where MC_i = fixed effect of minimum size of child nodes (i = 1 to 7), PP_j = fixed effect of parameter for prior probability of positive cases (j = 2.4%, 1.2%, 0.6%, or 0.3%), and $Fold_k$ = random effect of fold (k = 1 to 10). Within each level of assignment of cases to folds, each combination of the other factors of interest was trained and tested on the same training and testing sets. The factor treatment of irrelevant attribute values and the three factors representing the configuration of algorithm parameters were, therefore, considered as being repeated within the factor fold. A compound symmetry covariance structure was used to account for the covariance among the repeated observations within each fold (Littell et al., 1996).

5.2.6 Analysis of the learned knowledge

In addition to the numerical evaluation of the performance, a qualitative analysis of the learned knowledge was performed. A series of three final decision trees for implementation in a KBS was induced from the entire data set using the appropriate type of data preprocessing and algorithm configuration as determined in the two experiments (Figure 5.3). Each decision tree was induced with a different setting for the misclassification cost of positive cases parameter, allowing end-users of the KBS to choose from three specific trade-offs between sensitivity and specificity. These three decision trees were considered as representing a low, medium, and high filtering intensity for the removal of outlier lactations. A decision tree, induced with a relatively low setting for the cost of misclassifying positive cases, was expected to filter out relatively few cases (low PPR), while a high misclassification cost was associated with a high filtering intensity (high PPR).

To determine the classification performance of these decision trees, induced from the entire data set, it was necessary to get an estimate of the FP rate and TP rate, associated with each setting of the misclassification cost of positive cases parameter. Thus, a second approach to averaging the classification results of ten-fold cross validation was used for the chosen type of data preprocessing and algorithm configuration. This involved averaging the ten FP and TP rate pairs associated with each level of misclassification costs, resulting in ten average data points in ROC space, which were connected to yield a so-called pooled ROC curve (Bradley, 1997).

To verify how closely the classification performance of the three final decision trees induced from the entire data set resembled the sensitivity versus specificity trade-off observed with the cross-validated decision trees, the resubstitution FP and TP rates, determined through testing on the data used for training, were analyzed. In addition, The three final decision trees were evaluated by the domain specialist to verify the plausibility of the induced rules and to allow for manual adjustment of the decision at each node. To support this plausibility analysis, the single decision nodes of the final trees and, in some cases, pairs of decision nodes, were evaluated as small pieces of knowledge and tested for classification performance against the entire data set. Decision nodes performing poorly in this test may be the result of peculiarities in the training data, in which case they would not hold on new data. These decision nodes, which may be considered as counter-intuitive and unacceptable (Pazzani, 2000), were marked as suspicious and discussed in detail with the domain specialist.

5.3 Results

5.3.1 Experiment 1

Table 5.2 lists the average performance index for each level of the three factors of interest in the first experiment. Assigning herds to folds showed, on average, a 1.6% lower estimate of TP*, the mean TP rate for the range of FP rates from 0% to 10%, than assigning cases to folds.

Table 5.3 Mean of performance index TP* (%) for each of the two levels of the four factors studied in experiment 1.

Factor of interest	Description level 1	Mean TP* level 1	Description level 2	Mean TP* level 2	Difference TP* level 1 and 2
Assignment to folds	Case	59.4	Herd	57.7	-1.6 ^{ns}
Irrelevant attribute values	Unknown	57.0	Special	60.2	3.2 $^{P=0.054}$
Splitting and pruning criterion	Gini	55.1	SymGini	62.1	7.0 ***
Minimum size of child nodes	1	57.5	5	59.6	2.1 ^{ns}

Table 5.3 lists the mean and standard error of TP* for each of the 16 different combinations of data preprocessing and algorithm configuration involved in the first experiment. Of the eight comparisons between assigning herds and assigning cases to folds, three comparisons actually showed a higher estimate of TP* for assigning herds to folds. Analysis of variance indicated that the observed difference of TP* between assigning herds and cases to folds was not statistically significant (Table 5.2). Figure 5.6 shows the average ROC curves for the four combinations of data preprocessing for the Gini splitting and pruning criterion and the minimum size of child nodes set to 1. From 0% to 4% FP rate, assigning herds and assigning cases to folds resulted in similar TP rates. Between 5% and 10% FP rate, the difference between the two levels varied from

Table 5.2 Performance index TP* for each combination of the four factors studied in experiment 1.

		Minimum size of child nodes			
Splitting and pruning criterion	Treatment of irrelevant attribute values	1		5	
		Assignment to folds		Assignment to folds	
		Case level	Herd level	Case level	Herd level
		(%) (s.e.)	(%) (s.e.)	(%) (s.e.)	(%) (s.e.)
Gini	Unknown	55.4 (4.8)	54.6 (5.1)	53.8 (5.3)	52.8 (5.7)
Gini	Special	57.7 (4.4)	53.8 (5.6)	57.9 (4.0)	54.4 (5.7)
SymGini	Unknown	56.3 (6.9)	57.9 (4.5)	62.1 (5.2)	62.7 (5.7)
SymGini	Special	65.3 (5.6)	59.0 (5.3)	66.4 (5.4)	66.7 (5.8)

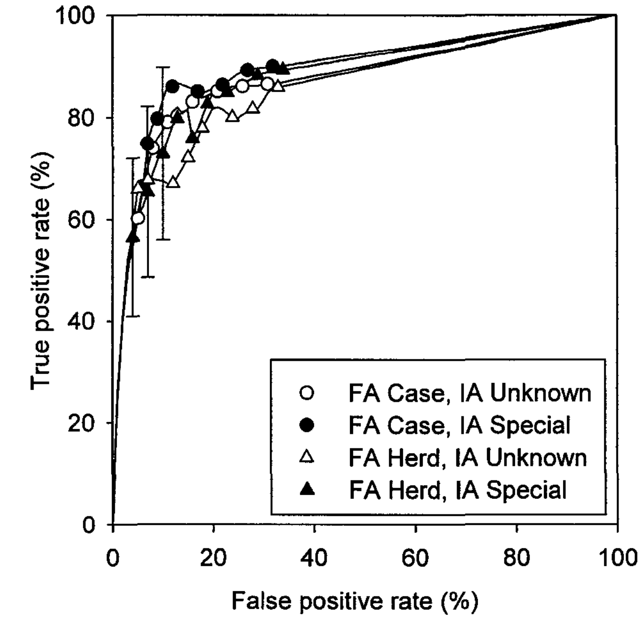


Figure 5.6 Average relative operating characteristic (ROC) curves for two levels of assignment of cases to folds (FA) combined with two levels of treatment of irrelevant attribute values (IA), using the Gini splitting and pruning criterion and minimum number of child nodes equal to one.

6% TP rate in favour of assigning herds to folds to 10% TP rate in favour of assigning cases to folds. As indicated with the error bars, representing approximate 95% confidence intervals for the TP rate of the combination of assignment of herds to folds and special valued irrelevant attribute values, the TP rate values showed a large variability from one fold to another. Analysis of the ROC curves for the other 12 combinations of data preprocessing and algorithm configuration revealed similar patterns regarding the difference between assigning herds or cases to folds. Although assigning herds to folds caused a much larger variability in the number of cases and class distribution among folds than assigning cases to folds, this did not result in a large increase in the variability of performance estimates among folds (Table 5.3). Thus, assigning herds to folds was used in the second experiment to ensure that the calculated performance indices were truly estimates of the expected performance on data from new dairy herds.

Use of a special value to deal with irrelevant attribute values showed, on average, a 3.2% higher estimate of TP* than considering these values as unknown (Table 5.2). However, detailed analysis of the eight comparisons between the two levels of the treatment of irrelevant attribute values showed that in one comparison, using a special value for irrelevant attribute values actually resulted in a lower estimate of TP* (Table 5.3). In addition, the standard errors for the TP* estimates in Table 5.3 are large and exceed the mean difference between the two levels of treating irrelevant attribute values. Thus, one may conclude that it is not possible to indicate that one type of treating irrelevant attribute values leads to a better performance than another. However, the variability in the results may to a large extent be attributable to the effect of fold. Figure 5.7 shows how the TP* performance varies with fold for the two levels of the treatment of irrelevant attribute values, combined with the two levels of splitting and pruning criterion and using assignment of herds to folds and minimum size of child nodes equal to 1. For some of the folds (e.g., 3 and 8) it was apparently more difficult to classify the test cases with the classifier generated from the training set than for other folds. Analysis of variance, taking into account the variability due to folds, indicated a P-value equal to 0.054 for the difference between the two levels of treatment of irrelevant attribute values. The treatment of irrelevant attribute values with special values was thus considered to improve the classification performance significantly.

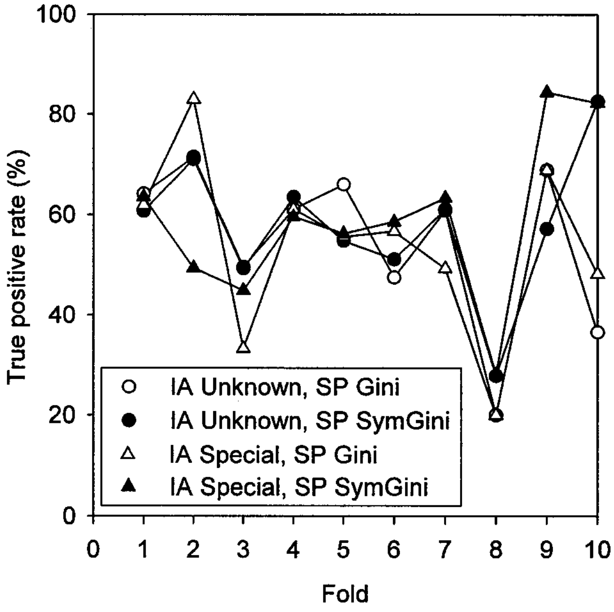


Figure 5.7 Performance index TP* plotted against fold for two levels of treatment of irrelevant attribute values (IA) combined with two levels of splitting and pruning criterion (SP), using herd level assignment to folds and minimum size child nodes equal to one.

The SymGini splitting and pruning criterion showed, on average, a 7.0% higher estimate of TP* than the Gini criterion (Table 5.2). The SymGini criterion also outperformed Gini in all of the comparisons shown in Table 5.3. Although the ANOVA indicated a statistically significant difference between the two types of splitting and pruning ($P<0.001$), a substantial interaction effect between the configuration of the parameters for splitting and pruning and minimum size of child nodes was detected ($P=0.094$). The results in Table 5.3 indicate that this interaction effect was due to a different response to changing the minimum size of child nodes from 1 to 5 for each level of the splitting and pruning criterion. Specifically, changing the minimum size of child nodes from 1 to 5 did not increase TP* in combination with the Gini criterion, but caused a substantial improvement with SymGini. Thus, SymGini splitting and pruning criterion was considered to substantially improve the classification performance compared to the default Gini criterion.

5.3.2 Experiment 2

Table 5.4 lists the mean of TP* for the 28 combinations of algorithm configuration involved in the second experiment. The mean performance for each level of the minimum size of child nodes parameter showed an optimum at level 6, which was 6.9% higher than the mean performance of the default level 1. Analysis of variance indicated statistically significant differences ($P<0.01$) among the levels of the minimum size of child nodes parameter. In pair-wise comparisons, each of the minimum size of child nodes configurations from 4 to 7 was found to be significantly different ($P<0.01$) from default level 1.

Table 5.4 Mean of performance index TP* (%) for each combination of the two factors studied in experiment 2.

Prior			Minimum size of child nodes					Mean of prior
probability	1	2	3	4	5	6	7	probability
2.4 %	59.0	60.9	59.7	63.6	66.7	65.9	65.3	63.0
1.2 %	63.0	63.5	62.7	67.4	67.5	68.4	64.6	65.3
0.6 %	60.2	63.5	64.0	67.1	68.0	68.4	67.7	65.6
0.3 %	59.0	61.4	62.1	65.6	65.6	66.2	65.9	63.7
Mean MC[†]	60.3	62.3 ^{ns}	62.1 ^{ns}	65.9 **	67.0 ***	67.2 ***	65.9 **	64.4

[†] Mean MC: mean for each level of minimum size child nodes and indication of statistical significance of deviation from mean TP* at minimum child nodes level one.

Figure 5.8 shows the average ROC curves for four different algorithm configurations using assignment of herds to folds and special values to treat irrelevant attribute values. These include the default algorithm configuration (Gini splitting and pruning criterion, minimum size of child nodes equal to 1, and 2.4% prior probability of outlier lactations) and three subsequent levels of parameter tuning. Error bars represent an approximate 95% confidence interval for the combination of SymGini splitting and pruning criterion, minimum size of child nodes equal to 6, and 0.6% prior probability of outlier lactations. Changing the minimum size of child nodes from 1 to 6 resulted, on average, in a substantial improvement in the TP rate between 4 and 10% FP rate (Figure 5.8). Thus, this factor was considered to significantly improve the classification performance.

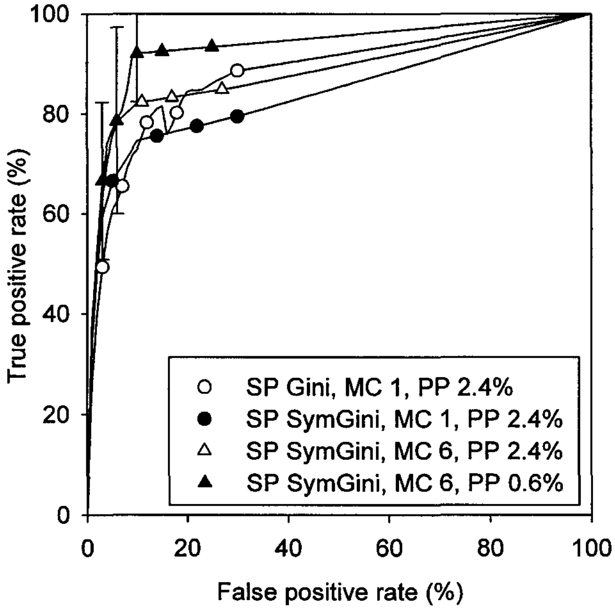


Figure 5.8 Average relative operating characteristic (ROC) curves for four algorithm configurations involving splitting and pruning criterion (SP), minimum size of child nodes (MC), and prior probability of outlier lactations (PP), using assignment of herds to folds and special values to treat irrelevant attribute values.

Decreasing the prior probability of outlier lactations parameter from the default 2.4% to 0.6% improved TP*, on average, by 2.6%. (Table 5.4). However, ANOVA indicated that the differences in performance among the four levels of the prior probability of outlier lactations parameter were not statistically significant. Detailed analysis of the average ROC curves revealed that decreasing the prior probability of outlier lactations parameter from the default 2.4 to 0.6% did not change, on average, the performance for FP rate values below 9%, but it was associated with a large increase in TP rate for 9 to 10% FP rate (Figure 5.8). The ANOVA was therefore rerun using an adjusted TP*, calculated as the mean TP rate of the ROC curves from 9 to 10% FP rate, and was limited to the levels of minimum size of child nodes that were significantly different from the default (4 through 7). For these constrained conditions, ANOVA indicated statistically significant differences among the four levels of prior probability of outlier lactations (P<0.01). In pair-wise comparisons, prior probabilities 1.2% and 0.6% were found to be significantly different (P<0.01) from default 2.4% prior probability. Thus, changing the prior probability of outlier lactations from 2.4% to 0.6% was also considered as a significant improvement on classification performance.

5.3.3 Final decision trees

Based on the results of the two experiments, the final tuned algorithm configuration consisted of the SymGini splitting and pruning criterion, a minimum size of child nodes equal to 6, and a 0.6% prior probability of outlier lactations. With this algorithm configuration, a series of three final decision trees was induced from the entire data set using three different levels of the misclassification costs of positive cases (20, 40, and 680). Table 5.5 shows for these decision trees the pooled performance estimates determined through cross validation. The specific misclassification cost settings were chosen to achieve PPR values of approximately one, two, and four times the observed prevalence of outlier lactations for, respectively, the low, medium, and high filtering intensity trees (Table 5.5). For the low filtering intensity, a decision tree with a PPR below the prevalence of outlier lactations was not used since this would have resulted in an unreasonably low TP rate. The pooled estimates of the FP rate for the low, medium,

Table 5.5 Cross-validation performance of decision trees associated with a low, medium, and high filtering intensity.

Filtering intensity	Relative cost of false negatives	False positive rate (%) (s.e.)	True positive rate (%) (s.e.)	Positive prediction rate (%)	Predictive value positive (%)
Low	20	1.5 (0.5)	51.7 (8.8)	2.7	45.1
Medium	40	3.5 (0.7)	68.3 (8.0)	5.0	32.5
High	680	8.6 (1.5)	91.7 (4.5)	10.5	20.9

and high filtering intensity trees induced from the entire data set were respectively 1.5%, 3.5%, and 8.6%, while the pooled estimates of the TP rate were 51.7%, 68.3%, and 91.7% (Table 5.5). Given a 2.4% prior probability of outlier lactations, the low filtering intensity decision tree was expected to remove 2.7% of the lactations (PPR) with 45.1% of the removed lactations being true positive cases (PVP), while at the high filtering intensity setting, 10.5% of the lactations would be removed with 20.9% of those being true positives.

Table 5.6 shows the size and resubstitution performance for the decision trees induced during the cross validation and for the trees generated from the entire data set. For the low and medium filtering intensity levels, the number of leaf nodes and the resubstitution performance of the entire data set trees closely resembled the average performance of the cross-validation trees. However, for the high filtering intensity, the entire data set tree, induced by the CART algorithm, showed a much lower resubstitution FP rate and TP rate, respectively 4.8% and 94.1%, than the average resubstitution FP rate and TP rate of the cross-validation trees, which were 7.8% and 99.4%. This unexpected result can be explained by the splitting and pruning approach that CART uses, involving an internal ten-fold cross validation to determine the optimum size of the decision tree. For the entire data set tree this resulted in a smaller optimum tree (7 nodes) than observed with the cross validation (8.5 nodes on average). Therefore, a slightly larger decision tree with ten leaf nodes was chosen from the series of decision trees induced by the CART algorithm for the misclassification cost of 680 (instead of using the optimum tree with seven leaf nodes). The resubstitution performance of this decision tree with ten leaf nodes showed a FP rate and TP rate of 6.4% and 100%, which was similar to the performance of the cross-validation trees (Table 5.6).

Table 5.6 Size and resubstitution performance of decision trees generated during cross-validation and of optimal and size-adjusted trees induced from the entire data set.

Filtering intensity	Cross-validation			Entire data set					
				Optimum size of tree			Adjusted size of tree		
	Leaf nodes (#) (s.e.)	FP rate (%) (s.e.)	TP rate (%) (s.e.)	Leaf nodes (#)	FP rate (%)	TP rate (%)	Leaf nodes (#)	FP rate (%)	TP rate (%)
Low	4.2 (0.2)	1.5 (0.1)	68.3 (1.7)	4	1.6	67.6	--	--	--
Medium	6.3 (0.4)	3.3 (0.3)	88.5 (1.1)	6	2.7	85.3	--	--	--
High	8.5 (0.2)	7.8 (0.6)	99.4 (0.4)	7	4.8	94.1	10	6.4	100.0

Figure 5.9 shows the final decision tree for the medium filtering intensity. The attribute "average SCC" was considered to be most important by the decision-tree induction algorithm. This attribute was chosen at the root node and again at the fourth decision node. The other attributes were "any test with CAR of code abortion", "average relative deviation from group-average lactation curve", and "regression parameter b". The three decision trees showed considerable overlap, with the first three nodes being exactly the same for all trees. Additional attributes appearing in the high filtering intensity decision tree included "regression parameter a", "percentage of tests with any CAR code", and "number of tests in lactation with a high protein to fat ratio".

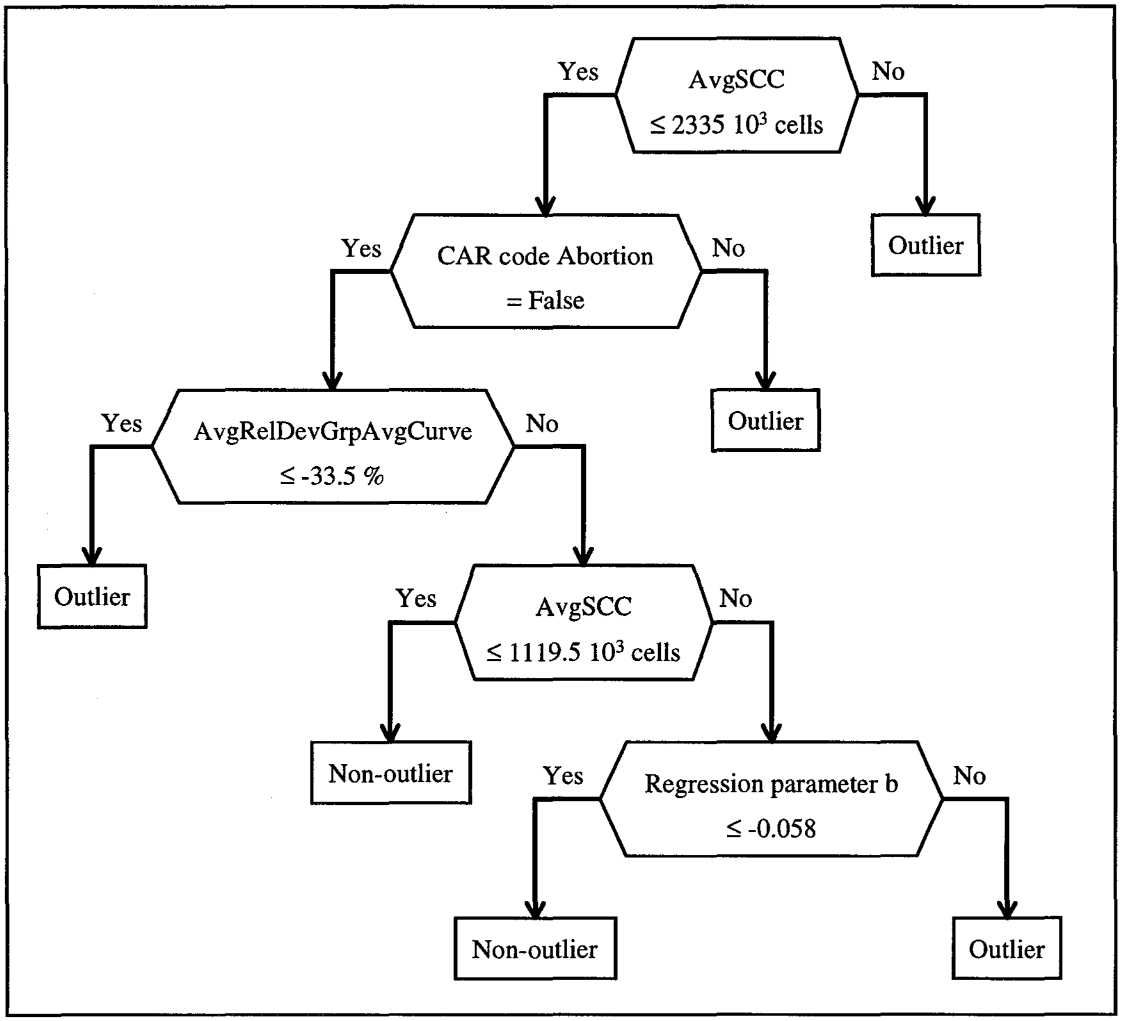


Figure 5.9 Decision tree induced from the entire data set for a medium filtering intensity.

The plausibility of the induced decision trees was discussed with the domain specialist, who considered the trees as being easy to understand. Table 5.7 lists all the different nodes of the three decision trees. The performance of the condition predicting outlier lactations at each node was tested within the context of decision tree (i.e., on the example cases of the data set that ended up at that node) and over the entire data set.

Table 5.7 Classification performance of individual decision nodes tested within the context of the decision tree and over the entire data set.

Decision node	Condition at decision node or nodes predicting outlier lactation	Tested within decision tree context		Tested over entire data set	
		Predicted positives (#)	PVP[†] (%)	Predicted positives (#)	PVP (%)
1	AvgSCC > 2335 10^3 cells/ml	8	87.5	8	87.5
2	CARcode_Abortion = True	7	85.7	7	85.7
3	AvgRelDevGrpAvgCurve ≤ -33.5 %	30	33.3	38	44.7
4	PctTestsAnyCARcode > 75.0 %	6	16.7	14	35.7
5	AvgSCC > 1119.5 10^3 cells/ml	39	15.4	52	30.8
6	AvgSCC > 940 10^3 cells/ml	18	5.6	71	23.9
7	PctTestsAnyCARcode > 31.5 %	42	2.4	68	19.1
8	NumTestsHighProteinToFatRatio > 0	6	16.7	294	2.7
7 and 8	PctTestsAnyCARcode > 31.5 % and NumTestsHighProteinToFatRatio > 0			10	30.0
9a	RegrParam_b > -0.058	22	27.3	816	3.4
9b	SlopeAfterPeak > -38.5 g/day	13	30.8	512	3.7
5 and 9b	AvgSCC > 1119.5 10^3 cells/ml and SlopeAfterPeak > -38.5 g/day			18	33.3
10a	RegrParam_a ≤ 18.0	9	22.2	11	27.3
10b	RegrParam_a ≤ 18.0 and RegrParam_b > 0	9	22.2	10	30.0

[†] PVP: predictive value positive.

Decision node 8 in Table 5.7 showed a very high number of cases predicted as positive (294) and a very low PVP (2.7%) when tested against the entire data set. However, in the decision tree, node 8 is preceded by node 7 and the combination of the two conditions to predict outlier lactation (nodes 7 and 8) showed reasonable performance when tested over the entire data set. Decision node 9a, "regression parameter b", showed a very low PVP (3.4%) when tested against the entire data set and was thought to cause poor performance on new data. This decision node was thus replaced with a competitor split provided by the CART algorithm. The alternative node 9b, "slope of the lactation curve after the peak", reduced the detection of actual positive cases from 6 to 4 in the context of the decision tree, but improved the performance when tested over the entire data set, especially in combination with the preceding node in the decision tree (nodes 5 and 9b). The decision node 10a "regression parameter a ≤ 18" showed reasonable performance over the entire data set. However, this attribute represented the intercept of the regression equation used to model lactation curves and was expected to lead to too many false positives when applied to low producing dairy herds. Thus, based on domain knowledge regarding the shape of lactation curves, modeled with the regression equation, a second condition, "regression parameter b > 0", was added. The resulting node 10b did not change the classification within the decision tree and improved the PVP over the entire data set slightly (Table 5.7). The two adjustments to the entire data set decision trees did not change the resubstitution performance of the low and high filtering intensity trees, but resulted in slightly lower resubstitution FP rate (from 2.7 to 2.2%) and TP rate (from 85.3 to 79.4%) of the medium level tree. However, these adjustments were expected to improve the classification performance on new data. The three final, modified, decision trees were considered as plausible by the domain specialist and were incorporated into the KBS for group-average lactation curve analysis.

5.4 Discussion

The example cases used in this study were acquired with a case-acquisition tool - the CADSS - created in consultation with two domain specialists specifically for the purpose of KBS development (Pietersma et al., 2001a). The data set for machine learning was, therefore, expected to be of high quality, with few unknown attribute values and a low level of mislabeling. The difficult and time-consuming process of data cleansing, generally required when machine learning is applied in the context of knowledge discovery from large existing data bases, could therefore be omitted. However, considerable time was spent on the creation of potentially predictive attributes, the analysis and treatment of unknown and irrelevant attribute values, and the extraction of separate training and testing data sets.

Relative operating characteristic curves and the performance index TP* were used to analyze and compare the performance of various classification schemes generated through machine learning. The ROC approach was found to be useful into visualizing the trade-off between sensitivity and specificity, achieved with these classification schemes. However, two different approaches to averaging the FP rate and TP rate data from the cross-validation experiments were required. Generating ten separate ROC curves, followed by averaging, allowed for the use of statistical tools to analyze the results. Averaging the FP rate and TP rate for each misclassification cost level was used to estimate the performance of the final decision trees induced from the entire data set.

Performance index TP*, the mean TP rate of the ROC curve for a limited range of FP rate values, was developed to enable comparison of the performance of multiple classification schemes with statistical tools. For classification tasks involving a highly unbalanced class distribution, this approach may be more appropriate than using the "area under the ROC curve" performance index. However, with TP*, domain expertise is required to target the performance analysis to the region in ROC space with FP rate values that are considered reasonable.

Analysis of variance was found to be a useful technique to support the analysis of the classification performance achieved with the experiments. Although this technique has only sporadically been used with machine-learning experiments (see e.g. Bradley, 1997), it offers several advantages to the common approach (see e.g. Mitchell et al., 1996; Kubat et al., 1998; Salehi et al., 2000) of limiting the performance analysis to comparisons of the mean results of multiple runs. Analysis of variance is able to account for the variability in performance due to the folds in k-fold cross validation experiments. It can also separate the main effects of the factors of interest from the interactions among these factors and can help to discern whether differences among classification schemes are likely to hold with new data or were just due to chance. However, ANOVA adds complexity to the analysis process and the results need to be expressed in a single performance index such as accuracy, area under the ROC curve, or the mean TP rate for the FP rate of interest. It should be noted that the combination of ten-fold cross validation and ANOVA may not be appropriate for studies where, instead of a specific classifier, as in this paper, a machine-learning system is implemented in the field to generate a classifier from new example data. In experiments to estimate the performance of such systems, both the test sets and the training sets used to generate the classifier need to be independent of each other. This is clearly violated with ten-fold cross validation, in which each pair of training sets has 89% of the data in common, leading to an elevated probability of incorrectly detecting differences among the tested machine-learning systems (Dietterich, 1998).

Contrary to the results of a study by Kubat et al. (1998), the expected positive bias in classification performance estimates associated with randomly assigning individual example cases instead of batches of cases to the folds in ten-fold cross validation did not prove to be significant. Thus, the prediction of example cases of a particular herd did not substantially improve when example cases of that herd were included in the data set used to induce the classifier. This suggests that in this study each of the different types of outlier and non-outlier cases appeared in multiple herds instead of being limited to a single herd.

The treatment of irrelevant attribute values with a special value to indicate a "not applicable" situation improved the performance over treating these values as unknown, as suggested by Witten and Frank (2000). The decision tree induction algorithm was thus able to make use of the additional information indicating whether a missing attribute value was unknown or irrelevant.

Tuning of the parameter configuration of the decision-tree induction algorithm greatly improved the classification performance compared to the default configuration. A large performance improvement was achieved by changing the default Gini splitting and pruning criterion to SymGini and the prior probability of positive cases to a quarter of the frequency observed for the entire data set. The algorithm thus seemed to benefit from focussing during tree growth on correctly classifying negative cases to reduce the FP rate and during pruning on correctly classifying positive cases to increase the TP rate. Further improvement was achieved by changing the minimum number of cases in a child node from the default 1 to 6, preventing the creation of very small nodes that were unlikely to be predictive on independent test data.

The final decision trees showed fairly good classification performance. For example, the decision tree associated with a medium filtering intensity had an expected sensitivity of 68% at 3.5% FP rate. Given the observed 2.4% prevalence of outliers, this classifier was expected to remove 5.0% of the lactations in a parity group with 33% of the removed lactations being true positive cases. However, the repeated ten-fold cross validation runs for parameter tuning may have led to an overspecialization of the classifiers for the entire available data set, resulting in a positively-biased estimate of the performance on new data (Henery, 1994; Witten and Frank, 2000). For large data sets, this so-called over-tuning can be avoided by keeping a sub-set of the data apart for final testing (Henery, 1994). For small data sets and applications where an unbiased performance estimate is critical (e.g., in the medical domain), the computationally expensive approach of ten-fold cross validation within ten-fold cross validation (Witten and Frank, 2000) might be appropriate.

Although the classification performance achieved in this research seems quite reasonable, further improvement may be possible. First of all, the quality of the existing data set may be improved by allowing the domain specialist to reevaluate the example cases misclassified by the decision trees. This may reduce the number of mislabeled cases and narrow the fuzzy zone around the decision boundary between outlier and non-outlier lactation curves. Secondly, the acquisition of additional example cases may also improve the classification performance. To explore this potential, so-called learning curves (Cohen, 1995) could be generated by removing a decreasing proportion of the data from the folds available for training.

In this study, the decision-tree induction approach to machine learning resulted in relatively small decision trees that were easy to understand by the domain specialist and allowed for manual adjustment of the decision nodes. This high level of understandability was considered an important advantage in the context of machine learning for knowledge acquisition, allowing the domain specialist involved to verify the plausibility of the results of learning and making it possible for end-users of the system to view a justification of the decisions made.

In a practical application of the removal of outlier lactations, the trade-off between sensitivity and specificity may depend on factors such as number of lactations available for the parity group being analyzed, prevalence of outliers, and end-user preference. The machine-learning approach to knowledge acquisition allows for the generation of a series of classifiers with increasing importance of correctly classifying positive cases from the single data set classified by the domain specialist. Implementation of a series of classifiers with increasing filtering intensity in the final KBS for group-average lactation curve analysis allows the end-users to move along the ROC curve and use the classifier with the desired sensitivity versus specificity trade-off. In addition, the end-user still has the ability to override the classifications made by a decision tree and exclude additional lactations or reconsider, and even undelete, some of those removed automatically.

5.5 Conclusions

This research suggests that automatically induced decision trees are quite good at mimicking the removal of outlier lactations as performed by a domain specialist. However, factors such as data preprocessing and algorithm configuration had a significant effect on the achieved performance. This research also explored the use of ten-fold cross validation to estimate the performance of classifiers generated from the labeled data available, visualization of the classification performance through ROC curves, assessment of differences of the mean true positive rate using analysis of variance, and evaluation of the plausibility of the induced decision trees via analysis of the performance of individual decision nodes. These methods were found to be useful for the development of a knowledge-based module to filter milk-recording data for group-average lactation curve analysis and may also be of use in research involving the application of machine learning in general.

5.6 Appendix A: Description of the CART algorithm

The CART (Classification And Regression Trees) algorithm performs binary recursive partitioning to automatically induce a decision or a regression tree from training data. Starting at the root of the decision tree, the algorithm considers each attribute and its corresponding attribute values as a potential rule for splitting the data into two subsets or child nodes, and selects the splitting rule leading to the largest reduction in heterogeneity or impurity of the observed classes. This process is repeated for each subsequent node in the decision tree. To quantify impurity, CART uses by default the Gini index *i(t)*:

$$i(t) = 1 - \sum_{j} p(j \mid t)^2$$

where *p(j|t)* is the probability of class *j* in node *t*. These probabilities are estimated from the observed relative frequencies of each class in the node, corrected for user-specified prior class probabilities and misclassification costs. The Gini index equals zero if the node consists of cases of only one class and has a maximum value if all classes are equally distributed. The impurity of a candidate split is calculated as the probability-weighted average of the impurity of the two child nodes. The algorithm chooses the splitting rule leading to the least impurity and uses that rule to assign the training cases associated with the parent node to the two child nodes. Training cases with missing values for the chosen splitting attribute are assigned to the child nodes using so-called surrogate splitting rules, which are selected to simulate the assignment to child nodes by the main splitting rule. The algorithm continues splitting nodes until all cases associated with the node have the same class or until a minimum number of cases in the parent or child node have been reached. The resulting maximum tree is then pruned back to avoid overspecialization to the training data and subsequent poor performance on independent test data. The optimum tree size is determined, by default, through stratified ten-fold cross validation. For each fold, the maximum tree is pruned back in a stepwise manner, leading to a series of trees with an associated size and average misclassification cost on the test set. Finally, a maximum tree is induced from the entire training data and pruned back to the size associated with the minimum expected misclassification cost as determined with the cross validation. With CART, classification of new cases involves dropping each case down the tree until a leaf node is reached, at which point the class with the highest relative frequency at that leaf node, corrected for misclassification costs and prior class probabilities, is assigned.

The CART algorithm includes a large number of user controllable parameters to customize the learning process. For example, the default Gini splitting and pruning criterion can be replaced with the so-called "SymGini" approach, which uses, unlike Gini, symmetric (equal) misclassification costs and user-specified priors during tree growth and employs, like Gini, user-specified misclassification costs and user-specified priors for pruning. The CART algorithm is described in detail by Breiman et al. (1984) and Steinberg and Colla (1997).

Preface to Chapter 6

The previous chapter dealt with challenges related to the appropriate preprocessing of example cases, configuration of the machine-learning algorithm, and analysis of the results of learning. Classifiers were generated to automatically remove outlier lactations of individual cows. Additional difficulties with the application of machine learning include the choice of an appropriate machine-learning technique and the development of potentially predictive attributes. Decision-tree induction has shown good classification performance in many domains and the generated decision trees tend to be much easier to understand than knowledge representations of other approaches to machine learning, which is an important advantage in the context of knowledge acquisition. However, the naïve-Bayes classifier, a probability-based approach to machine learning, may lead to better classification performance. In the previous chapter, considerable effort was required for the development of potentially predictive attributes to support the machine-learning process. However, the effect of availability of such attributes on the classification performance achieved with machine learning was not investigated.

In this chapter, machine learning is used to generate classifiers to automatically remove outlier tests within lactations of individual cows. This is complementary to the task of filtering entire lactations, which was dealt with in the previous chapter. Both tasks are part of the "removal of outliers" module of the case-acquisition and decision-support system described in Chapter 4. This chapter explores the effect of the machine-learning algorithm, decision-tree induction or naïve Bayes, combined with the effect of availability of potentially predictive attributes on the achieved classification performance. In addition, the knowledge representations of both the decision-tree induction and the naïve-Bayes approach are evaluated.

This chapter has been prepared for submission to the journal Transactions of the ASAE (Pietersma, D., R. Lacroix, D. Lefebvre, and K. M. Wade. Machine-learning assisted knowledge acquisition to filter lactation curve data).

6 Machine-learning assisted knowledge acquisition to filter lactation curve data

Abstract

Machine learning was employed to develop knowledge-based modules to filter test-day data of individual cows for group-average lactation-curve analysis. The importance of deriving predictive attributes and choice of machine-learning technique was explored. Data consisted of 1080 milk yield tests of which 108 had been classified as outliers by a dairy nutrition specialist and 972 cases had been classified as non-outliers. Two different approaches to machine learning were applied to these data: decision-tree induction and naïve-Bayes classification. Performance of the classifiers was estimated through ten-fold cross-validation while relative operating characteristic curves were used to visualize the achieved trade-off between sensitivity and specificity. Use of an initial set of derived attributes significantly improved the performance compared to limiting attributes to those available to the domain specialist. However, adding even more complex attributes did not necessarily improve the classification performance. Overall, the naïve-Bayes approach showed significantly better performance than decision-tree induction. For each machine-learning approach, three final classifiers, associated with a low, medium, and high filtering intensity, were generated from the entire data set. The expected true positive rate varied from 41% to 78% for false positive rates between 1.1% and 4.6%. However, due to the low prevalence of outlier tests, this performance was associated with a large number of false positives. The domain specialist considered the final classifiers of both approaches as plausible, but found the decision trees easier to understand than the naïve-Bayes classifiers. Machine learning was considered to be a promising approach to assist knowledge acquisition.

6.1 Introduction

Agricultural producers have access to increasing amounts of data, which may support on-farm decision making – a process that is becoming increasingly complex. However, exposure to too much data can easily lead to information overload and, thus, improper interpretations. Knowledge-based systems (KBS) have been identified as potentially useful tools to support producers and their advisors to interpret the available data properly and, possibly, provide them with expert recommendations (Doluschitz, 1990; Plant and Stone, 1991; Spahr et al., 1988). In the domain of dairy production alone, several such systems have been created (Allore et al., 1995; Grinspan et al., 1994; Pellerin et al., 1994). Traditionally, KBS have been developed based on interviews with domain experts, sometimes supplemented with other sources of knowledge such as documentation (Dhar and Stein, 1997; Durkin, 1994). However, the acquisition of knowledge through interviews has proven to be time-consuming and difficult. Alternatively, acquisition of knowledge can be partially automated with machine learning (Dhar and Stein, 1997; Langley and Simon, 1995). With this approach, a domain expert classifies example cases of the problem at hand. This is followed by the application of a machine-learning technique, such as decision-tree induction, to learn how to classify new cases from these examples. Machine learning may speed up the knowledge-acquisition process (Dhar and Stein, 1997) and lead to a potentially more accurate representation of the specialist's performance (Michalski and Chilausky, 1980; Ben-David and Mandel, 1995). However, a literature search revealed very few cases in agriculture where machine learning was used to assist knowledge acquisition. These included the application of rule induction to develop an expert system for soybean disease diagnosis (Michalski and Chilausky, 1980) and the use of decision-tree induction to support the creation of a KBS for tomato crop management in greenhouses (Mangina et al., 1999).

Although machine learning seems promising for knowledge acquisition, several challenges remain, including choice of an appropriate algorithm (Brodley and Smyth, 1997; Verdenius et al., 1997) and determining an effective representation for the data describing each example case (Langley and Simon, 1995). In the context of using machine learning to support knowledge acquisition, decision-tree induction might be considered as the default approach. Decision trees are very similar to the decision rules often used in KBS, and initial research into decision-tree induction was partly motivated by the difficulties associated with knowledge acquisition for KBS development (Michie et al., 1994). Decision trees tend to be easy to understand (Dhar and Stein, 1997; McQueen et al., 1995; Kononenko et al., 1998) and the approach has been applied successfully to many real-world problems (Langley and Simon, 1995). The so-called naïve-Bayes classifier - a probability-based approach to machine learning - represents an interesting alternative to decision tree induction, showing in some situations substantially better classification performance (Michie et al., 1994). Although the knowledge description generated with the naïve-Bayes approach may be more difficult to understand than decision trees, it still tends to be more transparent than the knowledge representations associated with other machine-learning approaches, such as artificial-neural networks and instance-based learning (Kononenko et al., 1998). Apart from the choice of an appropriate machine-learning algorithm, determining an effective representation for the data has been identified as a critical success factor in the application of machine-learning techniques to real-world problems (Langley and Simon, 1995). This involves the construction of attributes with potentially predictive value from the available data. Deriving predictive attributes may require considerable effort and might, therefore, benefit from consultation with domain specialists who are often able to provide suggestions (Langley and Simon, 1995).

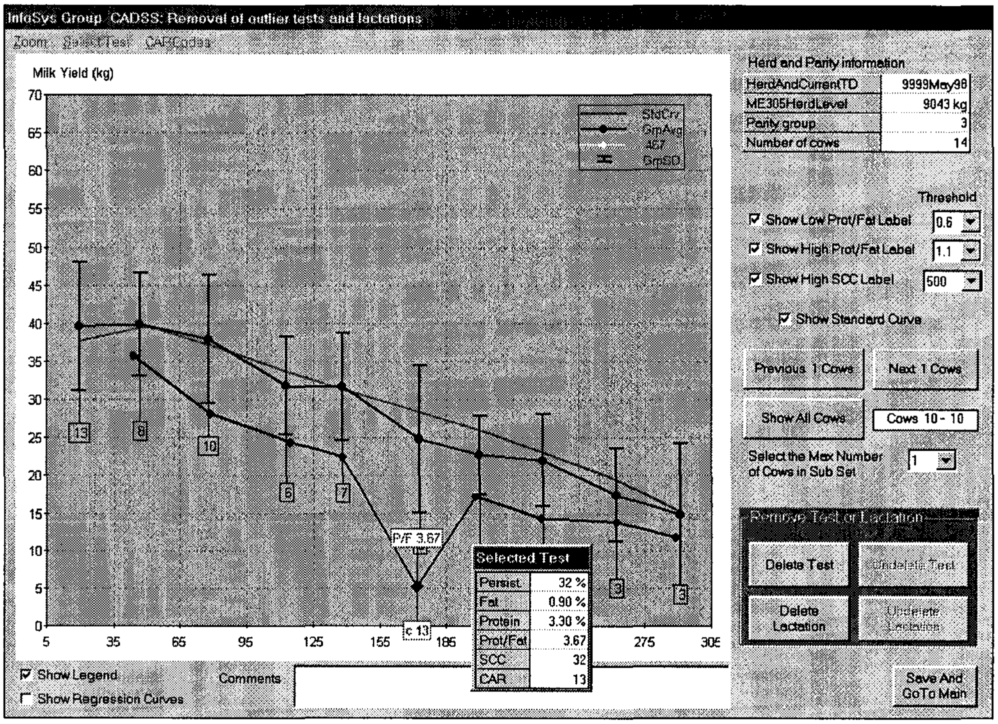
A research project was initiated to explore the use of machine learning to develop a KBS for the analysis of group-average lactation curves. This problem area involves comparison of group-average curves with standard curves as well as the analysis of additional explanatory data, with the objective of detecting potential management deficiencies. In a previous study (Pietersma et al., 2001a), the overall problem domain was decomposed into three sub-problems (removal of outlier data, interpretation of group-average lactation curves, and diagnosis of detected abnormalities). In addition, a case-acquisition and decision-support system (CADSS) was developed to enable domain specialists to work with example cases in the analysis of group-average lactation curves and to capture the resulting classifications (Pietersma et al., 2001a). The main goal of the research described in this paper was to develop knowledge-based modules to automate the removal of outlier tests within lactations of individual cows for group-average lactation-curve analysis using machine learning. Specific objectives were 1) to compare the classification performance achieved with decision-tree induction and naïve-Bayes classification, 2) to investigate the effect of availability of predictive attributes on the classification performance, and 3) to evaluate the plausibility of the knowledge represented by classifiers generated with either machine-learning technique.

6.2 Materials and Methods

6.2.1 Data

The removal of outliers was identified as the first step in the analysis of group-average lactation curves and considered important to avoid biasing the interpretation of the group-average performance by a few atypical lactations or tests (Pietersma et al., 2001a). With the CADSS, lactation curves (belonging to one of three parity groups) of individual cows could be compared with group-average and standard lactation curves (Figure 6.1). A user could select a particular test within the lactation curve of an individual cow to view additional information for that test, including the milk protein to fat ratio, somatic cell count, and codes indicating conditions affecting records such as clinical mastitis or estrus. Tests within lactations of individual cows and also entire lactations could be deleted to exclude them from group-average lactation-curve analysis (Figure 6.1). Details of the functioning of this CADSS can be found in Pietersma et al. (2001a).

A dairy-nutrition specialist used the CADSS to analyze the lactation curves of individual cows belonging to 33 Holstein herds enrolled with the Québec dairy herd analysis service and representing a wide range of rolling herd-average milk-production levels. With the CADSS the classifications of the specialist were captured resulting in a data set consisting of 7498 tests within lactations that each consisted of at least two tests. Milk yield tests belonging to lactations that consisted of only a single test were dealt with as entire lactations in previous research (Pietersma et al., 2001b) and, thus, excluded from this study.



Herd and Parity information

HerdAndCurrentTD	9999May98
ME305HerdLevel	9043 kg
Parity group	3
Number of cows	14

Threshold

☑ Show Low Prot/Fat Label 0.6

☑ Show High Prot/Fat Label 1.1

☑ Show High SCC Label 500

☑ Show Standard Curve

Cows 10 - 10

Select the Max Number of Cows in Sub Set 1

Selected Test

Persist	32 %
Fat	0.90 %
Protein	3.30 %
Prot/Fat	3.67
SCC	32
CAR	13

☑ Show Legend

☐ Show Regression Curves

Figure 6.1 Screen capture of the case-acquisition software module used to remove outlier tests and lactations.

The classification of test-day data of individual cows involved only two classes: a test could be labeled as an outlier (positive) or as a non-outlier (negative). Of the total number of tests, 108 (1.4%) were classified by the domain specialist as outliers. To reduce the computation time required for machine learning, only 972 randomly selected negative cases were used in addition to the positive cases, leading to a total of 1080 cases for training and testing (10% positive and 90% negative cases). The final classifiers for implementation in a KBS were also generated from these 1080 labeled cases.

6.2.2 Creation of attributes

Using the CADSS, the domain specialist had access to both graphical and numerical information to classify a test as either an outlier or as a non-outlier. Since a machine-learning algorithm cannot directly make use of graphical information, specific features or attributes describing such information need to be provided. However, the basic set of attributes representing the raw data used to draw lactation curves may provide only limited discrimination ability. Thus, the use of derived attributes that are constructed from the set of basic attributes and describe important aspects of the graphical information was expected to greatly help a machine-learning algorithm to discern between outlier and non-outlier cases.

In order to be able to study the importance of derived attributes for machine learning, three different levels of attribute availability were considered. The first level consisted of the 35 basic attributes available to the domain specialist using the CADSS. This basic set involved attributes representing the data used to create the lactation curves, such as the milk yield and days in milk on a test day for an individual cow, and the numeric attributes available with the CADSS, such as the somatic cell count on a test day. Table 6.1 lists the attributes used for machine learning, with multiple related attributes shown per row, and indicates the level of attribute availability at which they were used. The second level included, in addition to the level one, 16 attributes that were derived from the basic set, such as deviation of the test-day milk yield from the group-average lactation curve. The construction of these derived attributes focussed on fairly obvious aspects of the graphical information, such as deviations from the expected performance,

Table 6.1 Listing of attributes used for machine learning with three levels of attribute availability.

Level			Attribute description
1	2	3	Days in milk (DIM), milk, percent fat, percent protein, protein to fat ratio, somatic cell count, and conditions affecting records (CAR) code for the selected, prev., and next test of lactation curve
1	2	3	Sequence number of selected test within lactation, number of tests, and selected test is last test
1	2	3	Parity and Parity group
1	2	3	Persistency for the selected and the next test
1	2	3	Number of tests and standard deviation of group-average lactation curve for stage selected test
1	2	3	Average mature equivalent 305-day milk production of the herd
1	2	3	Number of test and lactations in parity group and in herd
	2	3	CAR code is unequal to zero for the selected, previous, and next test
	2	3	Absolute (abs.) deviation (dev.) slope between selected and previous test from slope std. curve
	2	3	Abs. dev. slope between selected and next test from slope standard curve
	2	3	Abs. and relative (rel.) dev. test from line between previous and next test
	2	3	Abs. and rel. dev. test from regression line through all tests in lactation including test
	2	3	Abs. and rel. dev. test from regression line through all tests in lactation excluding test
	2	3	Abs. and rel. dev. test from group average lactation curve
	2	3	Dev. of test from group average lactation curve expressed in number of standard deviations
	2	3	Abs. and rel. dev. test from standard curve
		3	CAR code abortion, milk fever, metritis, or displaced abomasum for selected, prev., and next test
		3	CAR code off-feed in early lactation for selected, previous, and next test
		3	Abs. and rel. dev. test from prediction previous test and shape standard curve
		3	Abs. and rel. dev. test from prediction previous non-outlier test and shape standard curve
		3	Abs. and rel. dev. test from prediction linear regression through previous three tests after peak
		3	If test is first test, abs. and rel. dev. test from prediction peak of lactation and shape std. curve
		3	If test is last test, CAR code and somatic cell count
		3	If test is last test, abs. and rel. dev. test prediction previous non-outlier test and shape std. curve

and required little input from the domain specialist. This second level was, thus, considered the default for attribute availability. Detailed analysis of the results of preliminary machine-learning experiments using the level-two attributes revealed poor classification performance for tests that were either the first or the last test of a lactation. Thus, a third level was created, which included, in addition to the level two, 18 derived attributes that were constructed in consultation with the domain specialist. These derived attributes aimed, for example, at improving the classification of the first and the last tests of a lactation and also included codes representing conditions affecting records considered most important by the domain specialist (Table 6.1).

The resulting three data sets contained a substantial number of records with missing attribute values. Some of these values were missing due to errors during milk recording and, thus, considered as unknown. However, missing attribute values that did not belong to this category were in fact irrelevant for the cases they described. For example, persistency (milk yield on test day n relative to the milk yield on test day $n - 1$) of the first test in a lactation cannot be determined. With the decision-tree induction approach, special values outside the range of possible values were used to indicate such irrelevant situations, which enabled the algorithm to consider these as a special group (Pietersma et al., 2001b; Witten and Frank, 2000). For the naïve-Bayes approach to machine learning, these irrelevant attribute values were, however, treated as unknown.

6.2.3 Machine-learning algorithms

In this study decision-tree induction was performed using CART for Windows version 3.6 developed by Salford Systems (Breiman et al., 1984; Steinberg and Colla, 1997). This algorithm learns in a top-down fashion, by splitting the training data recursively into two smaller subsets, choosing, at each split, the attribute and value that is most successful in discriminating among the classes of the classification problem. The CART algorithm continues splitting subsets until a maximum tree is reached. The tree is then pruned back to avoid overfitting the training data. The resulting decision tree consists of a series of decision nodes that, during classification, guide each new case to a leaf node indicating the predicted class. Preliminary experiments were performed to tune the settings of the parameters of the algorithm to the type of classification task and data involved in this research. The same parameter configuration was used for all three attribute levels: the SymGini splitting and pruning criterion, the minimum number of cases in child nodes set to four, and the parameters for prior probability of positive and negative cases set to the observed frequencies in the data set. The misclassification cost parameters were used to focus the decision-tree algorithm on correctly classifying positive cases over negative cases. For the remaining parameters, default algorithm settings were used. More details regarding the CART algorithm can be found in Breiman et al. (1984) and Steinberg and Colla (1997).

The probability-based algorithm used in this research was an extension of the naïve-Bayes classifier called "selective naïve Bayes" (Langley and Sage, 1994). Bayesian methods have been used for many years in the field of pattern recognition (Duda and Hart, 1973) and many applications for the classification of agricultural produce with machine vision have been reported (Howarth et al., 1992; Steinmetz et al., 1994). These applications tend to involve only numeric attributes, which allows for the use of a Bayesian classifier that assumes a multivariate normal probability density function for each class to account for the correlations among attributes (Duda and Hart, 1973). In addition, pattern recognition tends to be focussed on achieving high classification accuracy with little importance given to the understandability of the results of learning. In this study, machine learning was used to support knowledge acquisition for decision support, with as much emphasis on the understandability of the generated knowledge as on the classification performance. In addition, this study involved both numeric and categorical attributes. Thus, the naïve-Bayes classifier was chosen, which allows for evaluation of the evidence provided by each attribute value for each class and is able to deal with categorical attributes (Kononenko et al., 1998). In this study, numeric attributes were categorized into ten ranges or intervals, each with a width equal to a tenth of the difference between a preset minimum and maximum value for the attribute.

The naïve-Bayes classifier makes use of Bayes theorem and the simplifying assumption of independence of attributes within each class (Duda and Hart, 1973; Mitchell, 1997). Given this assumption, naïve-Bayes classification of a new instance involves the following equation:

$$c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{i=1}^{d} P(a_i \mid c_j)$$

with the class predicted using naïve Bayes (c_{NB}) determined as class c_j, belonging to a finite set of classes C, that has the maximum value for the prior probability of class c_j multiplied by the product of the probabilities for the observed values (a_i) of the d attributes of the new instance given class c_j (Mitchell, 1997). Thus, during classification of a new case the value of each attribute describing that case provides different amounts of evidence for each class, which is combined with the prior probability per class. For example, consider a classification problem with classes c_1 and c_2, corresponding to "True" and "False", and a new case with the values "High" and "Late" for attributes a_1 and a_2, respectively. Assume $P(c_1) = 0.6$ and $P(c_2) = 0.4$, $P(a_1=High \mid c_1) = 0.1$, $P(a_2=Late \mid c_1) = 0.2$, $P(a_1=High \mid c_2) = 0.2$, and $P(a_2=Late \mid c_2) = 0.5$. With this example, the class of the new case would be predicted as c_2 or "False" since (0.4) (0.2) (0.5) > (0.6) (0.1) (0.2). Learning with the naïve-Bayes classifier involves estimating the prior probability of each class and the probability of each attribute value given each class based on their frequencies in the training data (Mitchell, 1997; Witten and Frank, 2000). During classification, the contribution of attributes with a missing value can simply be excluded from the naïve-Bayes equation (Witten and Frank, 2000).

The naïve-Bayes classifier is robust to irrelevant attributes, since the conditional probability of an attribute value is expected to be the same for each class with such attributes, but sensitive to redundant or correlated attributes (Witten and Frank, 2000). For example, the addition of an attribute a_3 that is perfectly correlated with attribute a_2 doubles the weight of the evidence provided by attribute a_2 for each class without providing any new information to discriminate among the classes. Thus, in this study, the "selective naïve-Bayes classifier" (Langley and Sage, 1994) was chosen, which attempts to exclude highly correlated attributes through the selection of attributes with a hill-climbing search technique. The algorithm first determines the conditional probabilities given each class for each attribute from the frequencies observed with the available data. During attribute selection, the algorithm starts with an empty set of attributes and, at each iteration, the attribute leading to the largest increase in accuracy is permanently added to the subset of selected attributes. The accuracy is determined by testing the naïve-Bayes classifications using the selected attributes on the available data. This process continues until the addition of any of the remaining attributes would result in reduced accuracy (Langley and Sage, 1994). In addition to avoiding poor classification performance through the exclusion of highly correlated attributes, the attribute selection process also reduces the size of the knowledge representation of the naïve-Bayes classifier to those attributes considered important. A smaller number of attributes, each with a conditional probability distribution per class, leads to a more transparent knowledge representation than using all attributes available, which is advantageous in the context of machine-learning assisted knowledge acquisition.

In this study, the selective naïve-Bayes algorithm (SelNB) was implemented using Visual Basic (Microsoft Corporation, Redmond, WA). Database tables were used to store the parameter settings of the algorithm, description of the attributes, attribute values for the example cases, knowledge representation of the generated classifiers, and the classification results of training and testing experiments. The algorithm was extended to use misclassification cost as an attribute selection criterion instead of classification accuracy, and parameters were added to represent the cost of misclassifying class i as class j for all $i \neq j$. This allowed for putting more or less emphasis on correctly classifying positive cases versus correctly classifying negative cases, resulting in a cost-specific approach to attribute selection. To enable adjustment of the probability of classifying a new case as positive versus negative with a selected set of attributes, the prior probability values for each class were multiplied by a parameter to indicate the relative weight or importance of that class. This class-weight parameter provided a convenient way to adjust the decision threshold between positive and negative cases. To avoid conditional probabilities equal to zero for categorical or categorized numeric attributes, the count of cases in the training data belonging to a particular attribute value and class was initialized with the prior probability of that attribute value, regardless of the class (Witten and Frank, 2000). The total count of cases for each attribute value was initialized with one to avoid zero prior probabilities for any attribute value. The prior probabilities were set to the observed frequencies in the entire data set (0.014 for positive cases and 0.986 for negative cases). Based on preliminary experiments to tune the algorithm to the classification task and type of data available, the parameter representing the cost of mistakenly classifying positive cases as negative during attribute selection was set at four, while the cost of mistakenly classifying negative cases as positive was set at one.

6.2.4 Training and testing method

In order to estimate the performance of classifiers generated from the 1080 example cases, the stratified ten-fold cross-validation approach to training and testing was used (Breiman et al., 1984; Weiss and Kulikowski, 1991; Witten and Frank, 2000). With this approach the available data are divided into ten mutually exclusive subsets or folds with approximately the same class distribution as the original data set. Each fold is used once to test the performance of the classifier, generated from the combined data of the remaining nine folds, leading to ten "independent" performance estimates. Assuming that the classification performance improves as more data are used for learning, the true performance of the classifier generated from the entire labeled data set is expected to be at least as good as the ten-fold cross-validation estimate which is based on classifiers generated from 90% of the data. Entire herds, instead of individual example cases, were assigned at random to each of the ten folds to achieve an unbiased estimate of the performance on example cases belonging to entirely new herds (Kubat et al., 1998; Pietersma et al., 2001b). The assignment of herds to folds was constrained to achieve approximately the same class distribution in each fold as in the entire data set.

The same set of ten folds was used for training and testing throughout this study. Although three different data sets for attribute availability were used, these involved the same 1080 cases and the same assignment to folds.

6.2.5 Performance analysis

The removal of outlier tests represents a classification problem involving two classes. With such classification tasks, there are four possible outcomes during the testing of a classifier: an actual positive case can be predicted as positive or negative and an actual negative case can be predicted as negative or positive. To allow for detailed analysis of the performance of the generated classifiers, the following performance indices were used: 1) true positive rate (TP rate), defined as correctly predicted positives as a proportion of actual positives; 2) false positive rate (FP rate), defined as incorrectly predicted positives as a proportion of actual negatives; 3) predictive value positive (PVP), defined as correctly predicted positives as a proportion of all cases predicted as positives; and 4) positive prediction rate (PPR), defined as all predicted positives as a proportion of all cases (Swets, 1988; Weiss and Kulikowski, 1991; Witten and Frank, 2000). In some domains, the TP rate is referred to as the sensitivity and the FP rate as 1 – specificity. The prevalence of positive cases or prior probability of positives was estimated from the entire data set of example cases classified by the domain specialist as the actual positives as a proportion of all cases. The TP rate and FP rate are both independent of the prevalence of positive cases and, thus, the characteristics of the classifier (Swets, 1988). Conversely, the PPR and PVP depend on the prevalence of positive cases and can be mathematically derived from the TP rate and FP rate for a given prevalence level as follows:

$$PPR \ = \ \textit{Prevalence of positives} \ \times \ \textit{TP rate} \ + \ (1 - \textit{Prevalence}) \ \times \ \textit{FP rate}$$

$$PVP \ = \ \textit{Prevalence of positives} \ \times \ \textit{TP rate} \ / \ PPR.$$

Relative operating characteristic (ROC) curves (Swets, 1988) were used to visualize the trade-off between correctly classifying outlier cases and correctly classifying non-outlier cases. An ROC curve consists of the TP rate plotted against the FP rate (Figure 6.2). In ROC space, the lower left point (0,0) represents a classifier that assigns each case to the negative class, while the upper right point (100,100) represents a classifier that considers each case as positive. The upper left point (0,100) represents perfect classification, while the line y = x represents an ROC curve that can be achieved with random classification. Thus, the closer an ROC curve approximates the lines connecting (0,0) with (0,100) and (100,100), the better the performance. The ROC curve represents the entire range of trade-offs between sensitivity and specificity that can be achieved with a particular classification scheme. Each point on the ROC curve represents a specific classifier.

Comparison of the performance of multiple classification schemes with statistical tools requires the information represented by the ROC curve to be collapsed into a single performance index. To this end, the area under the entire ROC curve was proposed as a suitable performance index by Swets (1988) and used in several machine learning studies (Bradley, 1997; Yang et al., 1999). However, in this research the expected prevalence of positive cases in new data was very low (1.4%) and a classifier with a high FP rate would lead to too many false positives to be of practical use. Thus, instead of the area under the entire ROC curve, a performance index TP*, defined as the mean TP rate for the range of FP rates of interest, was used (Pietersma et al., 2001b). For this application, the FP rate of interest was limited to ≤ 5%. Figure 6.2 shows the ROC curve and associated TP* for an example classification scheme X.

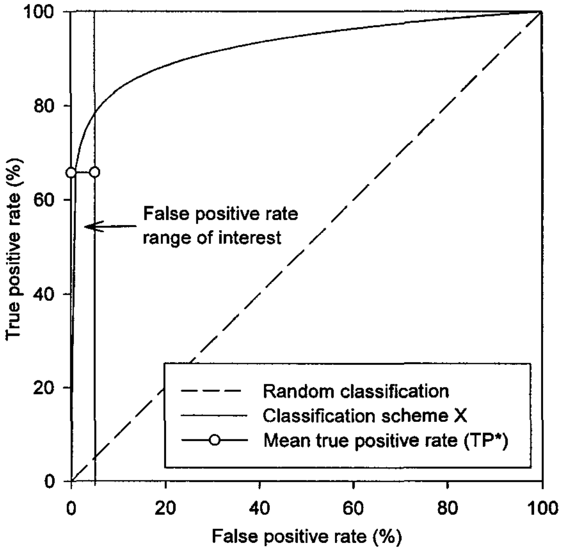


Figure 6.2 Relative operating characteristic (ROC) curves for random classification and an example classification scheme (X), and the mean true positive rate (TP*) for a specific range of false positive rate values.

To generate an ROC curve with CART, a series of ten decision trees was generated using ten different values for the parameter specifying the cost of false positives (ranging from, e.g., 1 to 100), while the cost of false negatives was kept at 1. With SelNB, ten different values were used for the parameter specifying the weight of the positive class, keeping the weight of the negative class at 1, to generate ten different points in ROC space. The same ten misclassification costs (CART) and class-weight values (SelNB) were used for each of the ten folds of cross-validation. This resulted in 100 pairs of FP rate and TP rate values for a particular combination of attribute availability and algorithm. For each fold, an ROC curve was generated through interpolation between the FP rate and TP rate data points associated with the increasing misclassification cost or class weight. An average ROC curve was then calculated by taking the average of the ten interpolated TP rate values for each FP rate from 0 to 100% at 1% intervals (Provost et al., 1998). For each of the ten ROC curves, TP* was calculated, resulting in ten independent estimates of TP*. This allowed for the calculation of a standard error associated with the mean TP* estimate and the use of analysis of variance. The ROC curves shown in this paper were all calculated using this averaging approach. Approximate 95% confidence intervals of the TP rate of the average ROC curves were determined as the standard error of the ten interpolated estimates of the TP rate for a particular level of FP rate multiplied by a t-value of 2.26 (two-sided probability level of 95% and 9 degrees of freedom).

6.2.6 Experimental design and analysis of variance

An experiment was designed to determine the effect of attribute availability and machine-learning algorithm on the classification performance achieved. The two machine-learning algorithms (CART and SelNB) were applied to each of the three attribute levels, leading to six different combinations of attribute availability and algorithm. For each combination, ten independent observations for TP* were obtained through the ten-fold cross-validation.

Analysis of variance was used to assess whether the observed differences of TP* among the attribute levels and machine-learning algorithms were likely to hold for new data (Pietersma, 2001b). Analysis of variance was performed with the Mixed procedure of SAS for Windows version 8 (SAS Institute Inc., Cary, NC). Each combination of attribute availability and algorithm was trained and tested on the same ten sets of training and testing cases. Thus, with the analysis of variance, the factors attribute availability and algorithm were considered as being repeated within the factor fold.

6.2.7 Evaluation of final classifiers

To allow end-users to choose classifiers at different points along the ROC curve, a series of three final decision trees and one final SelNB classifier with three different settings for the class-weight parameter was generated from the 1080 example cases described with all available attributes. These classifiers were considered as representing a low, medium, and high filtering intensity for the removal of outlier tests. A decision tree induced with a relatively low setting for the cost of misclassifying positive cases and a naïve-Bayes classifier with low value for the relative weight of the positive class, were expected to filter out relatively few cases (low PPR), while a high misclassification cost or high class-weight value was associated with a high filtering intensity (high PPR). To determine the classification performance of these specific classifiers, generated from the entire data set, it was necessary to obtain an estimate of the FP and TP rates, associated with each setting of the misclassification cost of positive cases parameter or the class-weight parameter for the positive class. This involved averaging the ten FP and TP rate pairs associated with each level of misclassification costs or class weight from the cross-validation, resulting in ten average data points in ROC space; these were then connected to yield a so-called pooled ROC curve (Bradley, 1997).

To evaluate the plausibility of these final decision trees, a quantitative and a qualitative assessment was carried out. Although for a classifier generated from the 1080 cases the true performance for new data can only be estimated, the apparent performance - also called resubstitution performance (Witten and Frank, 2000) - can be determined through testing of this classifier using the training data, i.e. the entire data set. Thus, the resubstitution FP and TP rates were used to quantitatively verify how closely the performance profile of the classifiers, induced from the entire data set, resembled the performance of the cross-validated classifiers. Manual adjustment of the level of pruning of the maximum tree induced with CART was used to achieve three final decision trees with the intended trade-off between sensitivity and specificity (Pietersma et al., 2001b). After the quantitative assessment, the three final decision trees and the final listing of conditional probabilities of attributes selected using the naïve-Bayes approach were analyzed and discussed with the domain specialist to qualitatively verify the plausibility of the generated knowledge representations.

6.3 Results

6.3.1 Attribute availability and machine-learning algorithm

Table 6.2 shows the mean and standard error of TP* for each of the six combinations of attribute availability and machine-learning algorithm. The standard error of the ten-fold cross-validation results per combination was fairly high, ranging from 3.7% to 5.6%, indicating a large variability in performance from fold to fold. For both algorithms, attribute level two showed a much higher TP* than level one. Increasing the number of available attributes from level two to three showed a small increase in performance for CART and a small decrease in performance for SelNB. However, analysis of variance indicated that this difference in response to attribute availability - the interaction between attribute availability and algorithm - was not statistically significant. The mean TP* values for attribute levels 1, 2, and 3 were, respectively, 43%, 60%, and 61%. The overall effect of attribute availability was statistically significant (P<0.01). Pair-wise comparison between attribute level two, considered as the default level of attribute availability, and level one showed a statistically significant difference (P<0.01), while attribute level three did not differ significantly from level two.

Table 6.2 Mean and standard error of performance index TP* (%) for each combination of machine-learning algorithm and level of attribute availability.

Algorithm	Attribute availability			Mean TP*
	1	2	3	algorithm
CART	39.7 (5.4)	55.3 (3.8)	58.6 (5.3)	51.2
SelNB	47.0 (4.6)	64.2 (5.6)	62.9 (3.7)	58.0 *
Mean TP* attribute	43.4 **	59.7	60.8 ^{ns}	54.6

Figures 6.3a and 6.3b show the average ROC curves for each of the six combinations of attribute availability and machine-learning algorithm. The entire curves shown in Figure 6.3a indicate that the relative performance among the six variants depended to a large extent on the FP rate. Figure 6.3b focuses on the range of FP rate values of interest and allows for detailed analysis of how the differences in TP rate among the six variants vary with the FP rate. In Figure 6.3b, error bars representing 95% confidence intervals for the TP rate are shown for the ROC curve representing the default

algorithm (CART) and default attribute level (2). The ROC curve for CART and attribute level 1 is situated entirely below this confidence interval, which suggests that CART and attribute level 1 showed significantly lower TP rates than level 2 throughout the FP rate range of interest. However, for CART and attribute level 3 the ROC curve did not extend beyond the 95% confidence interval. For SelNB, the ROC curve of attribute level 1 showed substantially lower TP rates throughout the FP rate of interest than the ROC curve for level 2. For SelNB and between 0 and 2% FP rate, the curve for attribute level 3 showed a substantially lower TP rate than the curve for level 2, while between 2 and 5% FP rate, attribute levels 2 and 3 showed fairly similar TP rates. Thus, throughout the FP rate of interest, restricting the attributes for machine learning to only those available to the domain specialist (level one) resulted in a significant reduction of the TP rate, while attribute level 3 did not significantly improve the performance over level 2.

At each attribute level, SelNB showed a higher TP* than CART, with a decreasing difference as more attributes were available for learning (Table 6.2). However, since the interaction between attribute availability and algorithm did not prove to be statistically significant, the results expressed in TP* did not provide enough evidence to declare that CART was more sensitive to the availability of predictive attributes than SelNB. The mean TP* values for CART and SelNB were, respectively, 51% and 58% (Table 6.2), and

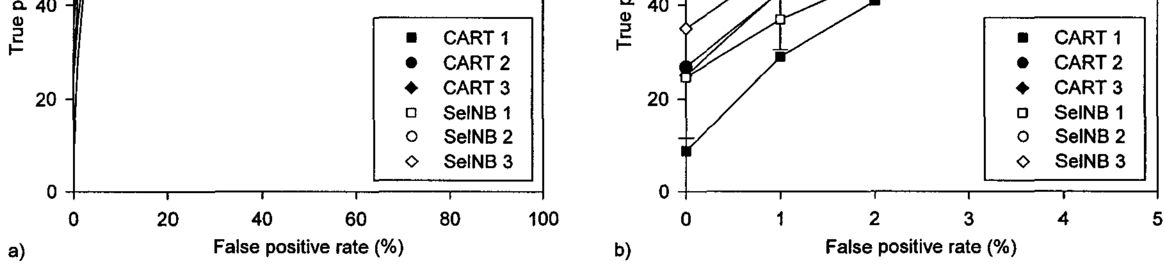


Figure 6.3 Average relative operating characteristic (ROC) curves for two machine-learning algorithms (CART and SelNB) and three levels of attribute availability (1, 2, and 3). Error bars represent an approximate 95% confidence interval. Figure 3a shows entire ROC curves while 3b focuses on the false positive rate interval of interest.

their difference was statistically significant (P<0.05). Detailed analysis of the ROC curves in Figure 6.3b suggest that the significant difference in TP* values between CART and SelNB was mainly due to the differences between the two algorithms for the range of FP rate values between 0 and 1%.

6.3.2 Classifiers generated from the entire data set

Using the entire data set and the largest number of attributes available, a series of three classifiers was generated for each algorithm. Although the classification performance of the highest attribute level did not significantly differ from the performance achieved with the second (default) attribute level, this third level was chosen since it included attributes that had been suggested by the domain specialist. Classifiers that included these attributes were expected to be considered by the specialist as more plausible than classifiers without such attributes. The three classifiers for each algorithm were chosen to be associated with a PPR approximately equal to one, two, and four times the observed prior probability of outlier tests (1.4%). Given the achieved ROC performance, a classifier with PPR below the observed prevalence of outlier tests was expected to have a TP rate too low for practical use.

Table 6.3 shows the misclassification costs or class weight settings and associated cross-validation performance of the three classifiers for CART and SelNB. Three different decision trees were induced with the misclassification cost of positive cases set at 3, 5, and 40. The low filtering intensity tree was expected to remove only 41% of the cases indicated as outliers by the domain specialist (TP rate). Assuming a 1.4% prior

Table 6.3 Characteristics and cross-validation performance of classifiers associated with a low, medium, and high filtering intensity.

Algorithm	Filtering intensity	Cost or class weight†	False positive rate	True positive rate	Positive prediction rate	Predictive value positive
			(%) (s.e.)	(%) (s.e.)	(%)	(%)
CART	Low	3	1.1 (0.3)	40.9 (6.8)	1.6	35.4
CART	Medium	5	1.7 (0.4)	59.6 (5.8)	2.5	33.1
CART	High	40	4.6 (0.8)	77.6 (4.6)	5.7	19.2
SelNB	Low	0.01	0.6 (0.4)	47.4 (7.3)	1.3	51.4
SelNB	Medium	0.3	2.1 (0.9)	62.3 (6.3)	2.9	29.8
SelNB	High	4	4.6 (1.5)	76.4 (4.9)	5.6	19.0

† Cost or class weight: cost of false negatives relative to cost of false positives for CART, weight of the positive class over the negative class for SelNB.

probability of outliers, this classifier was expected to predict 1.6% of the tests as outliers (PPR), but only 35% of these tests were expected to be truly outliers (PVP). The high filtering intensity tree had a much higher TP rate (78%) than the low filtering intensity tree. However, this was achieved by predicting 5.7% of all tests as outlier with only 19% PVP.

The size and resubstitution performance of the trees induced from the entire data set was compared with the average size of the classifiers generated with the cross-validation and with the average resubstitution performance determined by testing each classifier of the cross-validation on the training data used to generate that classifier (Table 6.4). For each filtering intensity, the entire data set tree induced by the CART algorithm had a size and resubstitution performance profile quite different from the cross-validation. For example, for the low filtering intensity, the entire data set tree had only two leaf nodes, much smaller than the average 5 nodes in the cross-validation, and the resubstitution TP rate was 25%, much lower than the average 45% TP rate in the cross-validation. This discrepancy can be explained by the splitting and pruning approach that CART uses involving an internal ten-fold cross-validation to determine the optimum size of the decision tree. Based on the entire data set, the algorithm chose a final tree with characteristics quite different from what was observed on average for the ten trees of the external cross-validation, which were induced from 90% of the data. To achieve a series of three final decision trees that more accurately reflected the intended trade-offs between sensitivity and specificity, the level of pruning of the maximum tree, considered optimum

Table 6.4 Size and resubstitution performance of cross-validation classifiers and of optimal and size-adjusted classifiers induced from the entire data set.

Algo-rithm	Fil-ter[†]	Cross-validation			All data optimum size			All data adjusted size		
		Nodes or attributes	False pos. rate	True pos. rate	Nodes or attributes	False pos. rate	True pos. rate	Nodes	False pos. rate	True pos. rate
		(#) (s.e.)	(%) (s.e.)	(%) (s.e.)	(#)	(%)	(%)	(#)	(%)	(%)
CART	L	4.9 (1.0)	0.1 (0.0)	45.4 (5.3)	2	0.0	25.0	5	0.0	45.4
CART	M	9.8 (0.8)	0.2 (0.0)	69.7 (3.3)	14	0.4	80.6	10	0.2	67.6
CART	H	11.9 (0.6)	3.3 (0.3)	89.8 (0.5)	28	3.2	99.1	13	3.3	89.8
SelNB	L	28.1 (2.1)	0.2 (0.0)	56.6 (4.9)	30	0.3	58.3	--	--	--
SelNB	M	28.1 (2.1)	1.0 (0.2)	73.0 (3.7)	30	1.0	75.0	--	--	--
SelNB	H	28.1 (2.1)	2.0 (0.1)	88.8 (1.8)	30	2.3	89.8	--	--	--

[†] Filter: low (L), medium (M), and high (H) filtering intensity; False pos. rate: false positive rate; True pos. rate: true positive rate.

by CART, was manually adjusted. For the low filtering intensity, less pruning than considered optimum by CART was chosen, leading to a larger tree. For the medium and high filtering intensities, more pruning was chosen, leading to smaller trees (Table 6.4).

Figure 6.4 shows the decision tree induced from the entire data set for the medium filtering intensity. The attribute representing the relative deviation of the milk yield of the test from the regression line between the previous test and the next test of the lactation was considered as most important. This attribute was chosen at the root node with

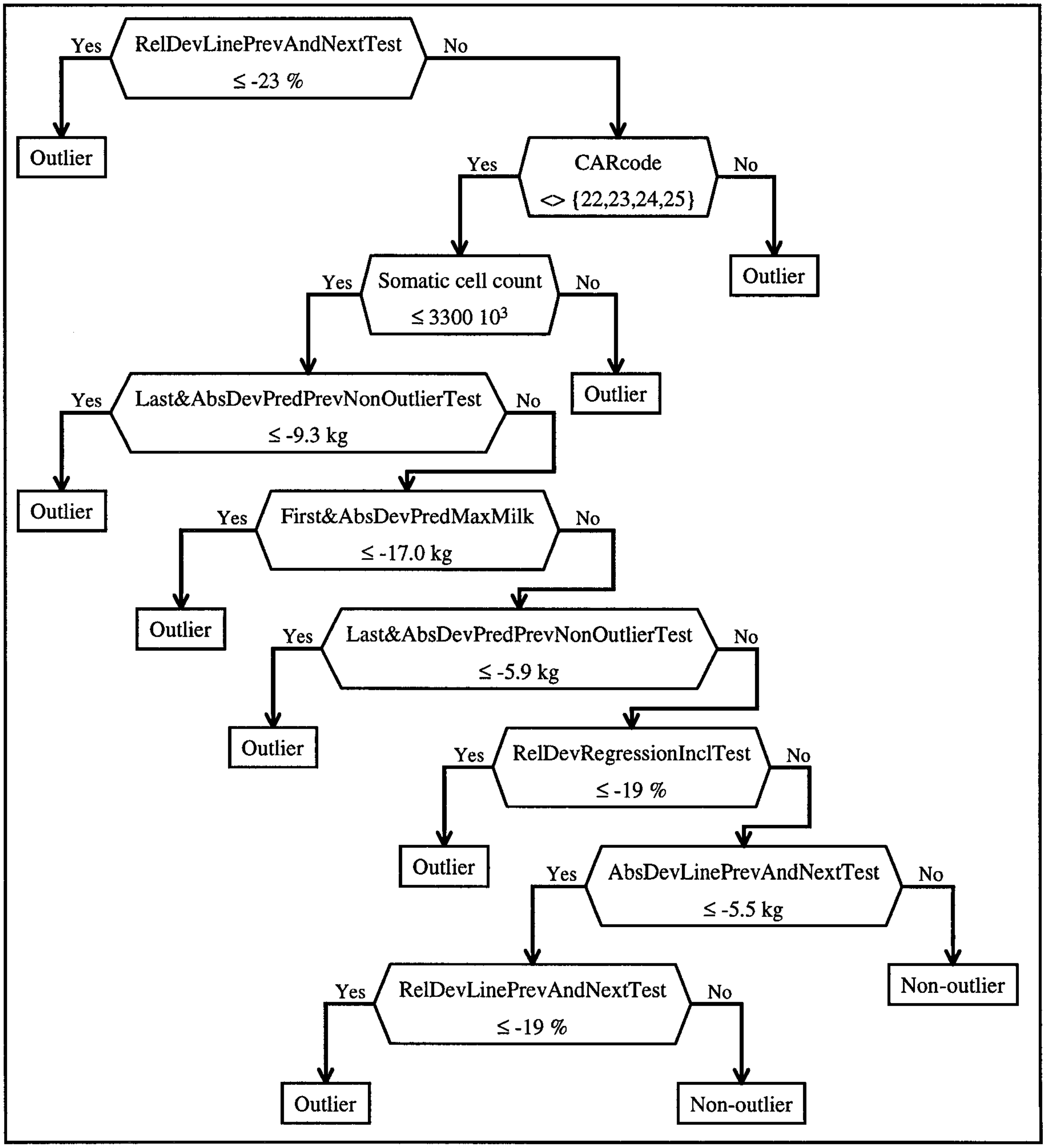

RelDevLinePrevAndNextTest ≤ -23 %
Yes
No
Outlier
CARcode <> {22,23,24,25}
Yes
No
Outlier
Somatic cell count ≤ 3300 10³
Yes
No
Outlier
Last&AbsDevPredPrevNonOutlierTest ≤ -9.3 kg
Yes
No
Outlier
First&AbsDevPredMaxMilk ≤ -17.0 kg
Yes
No
Outlier
Last&AbsDevPredPrevNonOutlierTest ≤ -5.9 kg
Yes
No
Outlier
RelDevRegressionInclTest ≤ -19 %
Yes
No
Outlier
AbsDevLinePrevAndNextTest ≤ -5.5 kg
Yes
No
Non-outlier
RelDevLinePrevAndNextTest ≤ -19 %
Yes
No
Outlier
Non-outlier

Figure 6.4 Decision tree induced from the entire data set for a medium filtering intensity.

threshold value –23% and again at the bottom of the decision tree with threshold value – 19%. Of the seven distinct attributes of the medium filtering intensity tree, two belonged to attribute level one, three were added at attribute level two, and two, related to the first and last tests, were added at the third attribute level. The low filtering intensity tree consisted of the first four decision nodes of the medium level tree. The first eight decision nodes of the medium filtering intensity tree were also part of the high filtering intensity tree. The additional four decision nodes of the high filtering intensity tree consisted of two attributes that also appeared in the low and medium filtering intensity trees, but with different threshold values, and two new attributes. Of the nine distinct attributes of the high filtering intensity tree, two belonged to attribute level one, while four and three attributes were added at attribute levels two and three, respectively.

With the SelNB algorithm, the use of three different settings for the weight of the positive class relative to the weight of the negative class (0.01, 0.3, and 4) resulted in three classifiers with a low, medium, and high filtering intensity (Table 6.3). At low filtering intensity, the SelNB classifier showed a higher TP rate than the decision tree, 47% instead of 41%. However, the standard errors of these TP rate estimates were very high (approximately 7%). The SelNB classifier showed a lower FP rate than the decision tree, 0.6% versus 1.1%. The lower FP rate combined with the higher TP rate resulted in a substantially higher PVP than the equivalent decision tree, 51% versus 35%. At the medium and high filtering intensities, the SelNB classifiers showed essentially the same performance as the CART decision trees.

The SelNB algorithm selected a set of 30 attributes from the 69 available attributes when applied to the entire data set (Table 6.4). The number of selected attributes and the resubstitution FP and TP rates of the classifiers generated from the entire data set closely resembled that of the classifiers of the cross-validation. The classifiers generated from the entire data set with SelNB were, therefore, not adjusted in size.

Table 6.5 shows a listing of the conditional probability values for the 10 most important attributes of the 30 attributes selected by SelNB, when applied to the entire data set. The first selected attribute represented the relative deviation of the milk yield of a test from the prediction based on the previous non-outlier test and the shape of the standard lactation curve. In Table 6.5, the first row for this attribute shows the upper limits for each of the ten ranges that were created during categorization. Thus, an attribute value of ≤ –66 belongs to range 1, an attribute value between –66 and –52 belongs to range 2, etc. For each of the two classes, the distribution of conditional probabilities over the ten ranges is shown. A non-outlier case is most likely to belong to range 6 and 7: 44 and 30% of the non-outlier example cases, respectively, appeared in those ranges. Conversely, an outlier case is most likely to belong to range 4 or 5 with 38 and 27% of the outlier example cases, respectively, appearing in those two ranges. During the classification of a new example case, the value of an attribute provided a conditional probability as evidence for each class. For example, a new case with the value –30 for the first attribute in Table 6.5 was categorized into range 4. This attribute contributed a rounded conditional probability of 0.01 for the non-outlier class and a conditional probability of 0.38 for the outlier class. The evidence provided by this attribute value favored the outlier class over the non-outlier class with a factor 30. The fourth row in Table 6.5 shows the ratio between the largest and the smallest conditional probability for each range of the first attribute. This ratio was given a positive sign if the outlier class had the largest conditional probability value; a negative sign was applied if the value for the non-outlier class was largest. These ratios were added for each selected attribute to support the evaluation of the plausibility of the generated knowledge description.

Some of the attributes selected with SelNB were clearly correlated with another selected attribute. For example, the first and eighth attribute listed in Table 6.5 are the relative and absolute variant of the same deviation from the prediction using the previous non-outlier test. During the attribute selection process, with performance testing on the training data, adding such a highly correlated attribute apparently did not reduce the classification performance.

Although the SelNB algorithm used many more attributes in its knowledge representation, 30 instead of the 9 distinct attributes selected by CART, some overlap existed. Four of the nine different attributes that appeared in the decision trees were also selected by SelNB and were part of the first ten attributes selected (Table 6.5).

Table 6.5 Conditional probabilities and their ratio for the ten most important attributes selected from the entire data set with the naïve-Bayes algorithm.

R[†]	Attribute	Item	Range for categorized numeric attributes or categorical attribute value									
			1	2	3	4	5	6	7	8	9	10
1	RelDevPred	Range limit[‡]	-66	-52	-38	-24	-10	4	18	32	46	60
	PrevNon	Non-outlier (%)	0	0	0	1	14	44	30	7	2	1
	Outlier	Outlier (%)	5	3	10	38	27	6	1	5	2	2
	(%)	Ratio*	6510	6104	81	30	2	-7	-21	-1	-1	3
2	AnyCARcode	Attribute value	False	True								
		Non-outlier (%)	98	2								
		Outlier (%)	58	42								
		Ratio	-2	18								
3	First&AbsDev	Range limit	-21	-17	-13	-9	-5	-1	3	7	11	15
	PredMaxMilk	Non-outlier (%)	0	0	2	6	10	26	33	15	7	2
	(kg)	Outlier (%)	20	15	10	15	26	6	1	6	0	0
		Ratio	1370	1285	6	3	3	-4	-23	-3	-22	-18
4	First&RelDev	Range limit	-58	-46	-34	-22	-10	2	14	26	38	50
	PredMaxMilk	Non-outlier (%)	0	1	2	6	18	30	26	10	5	2
	(%)	Outlier (%)	15	15	30	25	6	1	6	0	0	0
		Ratio	1285	25	13	4	-3	-23	-4	-22	-21	-18
5	Last&AbsDev	Range limit	-17	-14	-11	-8	-5	-2	1	4	7	10
	PredPrevNon	Non-outlier (%)	0	0	0	1	3	15	41	27	12	1
	Outlier	Outlier (%)	5	5	9	32	32	10	2	1	5	0
	(kg)	Ratio	797	797	1061	53	11	-2	-26	-25	-2	-17
6	AbsDev	Range limit	-9.6	-7.2	-4.8	-2.4	0	2.4	4.8	7.2	9.6	12
	Regression	Non-outlier (%)	0	0	2	9	40	38	10	1	0	0
	InclTest	Outlier (%)	8	13	14	27	22	5	4	4	2	1
	(%)	Ratio	6419	6724	9	3	-2	-8	-3	3	4893	3672
7	SomaticCell	Range limit	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
	Count	Non-outlier (%)	91	5	2	1	1	0	0	0	0	0
	(10^3)	Outlier (%)	62	11	8	6	3	2	0	1	2	7
		Ratio	-1	2	3	5	6	9	9	4934	6575	8627
8	AbsDevPred	Range limit	-16.5	-13	-9.5	-6	-2.5	1	4.5	8	11.5	15
	PrevNon	Non-outlier (%)	0	0	1	3	16	41	29	9	2	0
	Outlier	Outlier (%)	1	13	19	27	24	5	0	4	5	2
	(kg)	Ratio	4072	7458	38	11	2	-8	-98	-3	2	6
9	Milk	Range limit	7000	7500	8000	8500	9000	9500	10000	10500	11000	11500
	Production	Non-outlier (%)	2	12	12	8	16	19	3	8	12	9
	LevelHerd	Outlier (%)	3	13	7	12	16	16	1	12	16	6
	(kg)	Ratio	2	1	-2	2	-1	-1	-4	2	1	-2
10	PrevTestCAR	Attribute value	False	True								
	{19, 23, 24, 25}	Non-outlier (%)	100	0								
		Outlier (%)	98	2								
		Ratio	-1	18								

[†] R: Rank of attribute in selection process.

[‡] Range limit: attribute value indicating the upper limit of the range for categorized numeric attributes.

* Ratio: largest conditional probability divided by lowest with positive and negative sign to indicate outlier and non-outlier, respectively, as class with largest conditional probability.

6.3.3 Evaluation of final classifiers by domain specialist

The three induced decision trees were evaluated by the domain specialist and, once a detailed explanation of the meaning of the derived attributes was provided, considered easy to understand. All the decision nodes were thought to be plausible and expected to properly classify new data.

The attributes selected by SelNB and their conditional probability values were also evaluated by the domain specialist. For each attribute, the two conditional probability distributions were analyzed to establish whether a general pattern existed that was consistent with the domain expertise. For 13 of the 30 selected attributes, such a pattern did not exist. These attributes were thought to be the result of overfitting the data and, thus, removed from the list. For example, the pattern of conditional probabilities of attribute nine in Table 6.5, representing the average mature equivalent milk production of the herd, was not expected to properly classify new data. The removal of these 13 attributes reduced the resubstitution performance on the entire data set. For example, for the high filtering intensity SelNB classifier, the resubstitution FP rate increased from 2.3 to 3.1% and the TP rate decreased from 90 to 86%. However, relying on the expertise of the domain specialist, these adjustments were expected to improve the performance for new data. Of the remaining 17 attributes, 3 attributes belonged to attribute level one, while 2 and 12 had been added at levels two and three, respectively.

Overall, the description of learned knowledge via decision trees was considered as more transparent than the listing of conditional probability values for each attribute selected with naïve Bayes. The induced decision trees were relatively small with each decision node consisting of a single attribute threshold value and a clear assignment to a class. Conversely, SelNB resulted in a fairly long list of attributes and for each attribute many conditional probability values were involved. Also, for the decision trees it was considered relatively easy to grasp how the different decision nodes interacted with each other leading to the final classification of a new case, whereas with the naïve-Bayes approach it was considered very difficult to obtain an overview of the combined effect of the conditional probability values of the selected attributes on the final classification as either outlier or non-outlier.

6.4 Discussion

In this study, the use of derived attributes constructed from the basic data available to the domain specialist significantly improved the classification performance. Thus, deriving potentially predictive attributes was a very important part of the process of machine-learning assisted knowledge acquisition. However, the addition of even more complex attributes following detailed analysis of preliminary results and consultation with the domain specialist did not lead to further improvement of the performance. Thus, after the addition of a first set of derived attributes to the original data, the availability of potentially predictive attributes seemed to be no longer a limiting factor to the improvement of the classification performance. When all attributes were made available for machine learning, several of the attributes added at the highest attribute level, were selected by the decision-tree induction and naïve-Bayes algorithms to be part of the final classifiers. Inclusion of such attributes improved the understandability and acceptability of the learned knowledge by the domain specialist.

Both machine learning approaches allowed for inspection of the plausibility of individual pieces of learned knowledge. With the decision trees, each decision node could be analyzed to verify the plausibility of the attribute, threshold value, and class assignment. With the naïve-Bayes approach, the conditional probability distributions of each selected attribute could also be inspected. With both approaches, individual decision nodes or attributes that were not expected to properly classify new data could be removed or manually modified. However, the interactions among these pieces of knowledge were much easier to grasp with the decision trees than with the conditional probability lists. The decision trees offered a crisp description of the decision boundary between outlier and non-outlier cases, whereas the conditional probability lists, with each attribute providing different amounts of evidence for each class, represented more of a fuzzy approach to classification. Thus, the domain specialist considered the representation of learned knowledge with decision trees to be more transparent than the conditional probability lists generated with naïve Bayes.

The achieved classification performance was considered as fairly poor with 78% TP rate achieved at a PPR equal to four times the prior probability of outlier tests observed in the data. Thus, in a practical application of the filtering of milk recording data for group-average lactation-curve analysis, a reasonable level of sensitivity would require a classifier that removed four times as many tests as the domain specialist. This may, to some degree, have been caused by the existence of mislabeled example cases in the data set. The domain specialist indicated that for some of the example cases the classification was rather subjective and that inconsistencies may have occurred. Thus, the quality of the data set may be improved by allowing the domain specialist to reevaluate misclassified example cases. Such re-evaluation is expected to narrow the fuzzy zone around the decision boundary between outlier and non-outlier cases and reduce the number of false positives and false negatives in cross-validation experiments. The inspection of potentially mislabeled cases by the domain specialist could be limited to those cases misclassified by both machine-learning approaches, as suggested by Brodley and Friedl (1996). Although a fairly large number of minority class cases was available (108), it seems that many different reasons exist for classifying a test as outlier. Some of these patterns may have a very low prevalence and additional example cases would support the learning of those situations. Thus, in addition to the re-evaluation of misclassified cases, the acquisition of additional example cases from additional herds may also improve the performance.

Although the naïve-Bayes algorithm used in this research already achieved better classification performance than CART, further improvement may be possible through the use of an enhanced attribute selection process. In this study, the SelNB algorithm selected approximately three times as many attributes as the number of decision nodes used by CART during the cross-validation with the highest attribute level. Of the 30 attributes selected by SelNB from the entire data set, more than 40% were considered by the domain specialist as unlikely to properly classify new data. This suggests that the attribute selection process was not optimal. Instead of using a hill-climbing search method and relying on the resubstitution performance as a selection criterion, as shown by Langley and Sage (1994) and implemented in this study, other approaches to attribute selection using more complex search methods and an internal cross-validation procedure (Kohavi and John, 1997) may result in improved classification performance. However, those alternatives would have increased the computation time dramatically.

In this study, choosing one particular machine-learning approach over the other would represent a trade-off between classification performance and understandability of the learned knowledge. The naïve-Bayes approach showed significantly better classification performance than decision-tree induction, but the domain specialist considered the decision trees as more understandable than the conditional probability lists generated with the naïve-Bayes classifier. Further investigation is required to determine the preferred machine-learning approach to automate the removal of outlier tests within lactations. The use of an improved data set following re-evaluation of misclassified example cases or a larger data set is expected to improve the classification performance, but may benefit one algorithm more than the other.

In a practical application of the filtering of milk-recording data for lactation-curve analysis, the trade-off between sensitivity and specificity may depend on factors such as number of tests available for the parity group being analyzed, prevalence of outliers, and end-user preferences. Implementation of a series of classifiers with increasing filtering intensity in the final KBS for group-average lactation-curve analysis allows the end-users to move along the ROC curve and use the classifier with the desired trade-off in sensitivity versus specificity. In addition, the end-user always has the ability to override the decisions made by the classifier.

6.5 Conclusion

This research suggests that machine learning is a promising approach to develop knowledge-based modules to mimic the filtering of lactation curve data by a domain specialist. Use of additional attributes derived from the basic data significantly improved the classification performance. Overall, the naïve-Bayes approach showed significantly better performance than decision-tree induction. However, the domain specialist considered the decision trees as more transparent than the conditional probability lists generated with naïve Bayes. Although the final classifiers were considered as plausible, improvement of the classification performance through, for example, re-evaluation of misclassified cases by the domain specialist, might be required for a practical application.

Preface to Chapter 7

A framework for the development of computerized information systems in dairy farming was established in Chapter 3 and formed the basis for the creation of a decision-support system for the analysis of group-average lactation curves in Chapter 4. This decision-support system included case-acquisition functionality to enable a domain specialist to analyze and classify a substantial amount of example cases resulting from the analysis of group-average lactation curves for machine learning. Chapters 5 and 6 dealt with several methodological aspects related to the use of machine learning for knowledge acquisition, including creation of potentially predictive attributes to support learning, choice of a machine-learning technique, and performance analysis of classifiers generated from small data sets. In addition, Chapters 5 and 6 resulted in classifiers for implementation in the first module of the decision-support system described in Chapter 4, to automatically exclude outlier lactations and tests within lactations of individual cows from group-average lactation-curve analysis.

This chapter focuses on the creation of classifiers for each of the classification tasks identified as part of the interpretation of group-average lactation curves, the second module of the system described in Chapter 4. For these classification tasks, the understandability of the results of learning was considered to be very important. Thus, based on the results obtained in Chapter 6, the decision-tree induction approach to machine learning was used instead of the naïve-Bayes classifier. In addition, a method was developed to enable use of the performance analysis approach established in Chapter 5 for classification tasks involving more than two classes.

This chapter has been prepared for submission to the journal Canadian Biosystems Engineering (Pietersma, D., R. Lacroix, D. Lefebvre, and K. M. Wade. Decision-tree induction to interpret group-average lactation curves).

7 Decision-tree induction to interpret group-average lactation curves

Abstract

Decision-tree induction was used to automatically learn to interpret group-average lactation curves in dairy farming. Lactation curves are a graphical representation of the daily milk yield after calving and can be analyzed together with additional information to support the detection of management deficiencies. A dairy-nutrition specialist analyzed 98 group-average lactation curves, representing 33 dairy herds, and classified these curves regarding predefined aspects of interpretation. For machine learning, seven main classification tasks and three secondary tasks, supporting one of the main tasks, were identified. For each task, potentially predictive attributes were created based on the graphical and numerical information available to the specialist. Five-fold cross-validation was used to estimate the classification performance, and relative operating characteristic curves were used to visualize the achieved trade-off between sensitivity and specificity. For five of the seven main classification tasks, a series of three final decision trees, with increasing sensitivity and associated with a low, medium, and high tendency of classifying new cases as abnormal, were induced from the entire data set. For two of the main tasks, alternative trees showed very similar performance. The medium tendency trees were chosen to lead to a probability of predicting new cases as abnormal similar to the observed prevalence of abnormal cases, given a population of cases with that prevalence. The decision trees induced for the main classification tasks showed good performance. For the medium tendency decision trees, the sensitivity was at least 80% and the number of truly abnormal cases as a percentage of all cases predicted as abnormal was at least 75%. For the secondary tasks, the performance was poor and domain expertise was required to select a plausible tree from alternative trees generated by the induction algorithm. The decision trees, ranging from two to seven leaf nodes, were evaluated by the domain specialist, and, after a few adjustments, considered as plausible. This study suggested that automatically induced decision trees are able to closely match the interpretation of group-average lactation curves as performed by a domain specialist.

Machine-learning assisted knowledge acquisition is expected to be especially appropriate for problem domains where specialists have difficulty expressing decision rules, such as the analysis of graphical information.

7.1 Introduction

Dairy producers enrolled in a dairy-herd improvement program have access to a large amount of data concerning the milk production of their cows. This milk-recording data may support many management and control activities in various spheres of dairy farming and at different levels of decision making (Pietersma et al., 1998). The analysis of group-average lactation curves derived from such data has been identified as a useful tool to support nutrition management in dairy farming (Lefebvre et al., 1995; Skidmore et al., 1996; Whittaker et al., 1989). This type of analysis involves interpretation of the shape of the composite lactation curve of a group of cows, comparison of these curves with standard curves, and analysis of additional explanatory data, with the objective to detect potential management deficiencies. Use of a knowledge-based system (KBS) to support the analysis of group-average lactation curves might be advantageous to dairy producers. Such a system was developed at the Texas A&M University (Fourdraine et al., 1992a; Whittaker et al., 1989) to automate the preprocessing of the large amount of raw data involved and provide dairy producers and their advisors with expert interpretation. A KBS for the analysis of group-average lactation curves might also be of benefit to dairy producers in Canada, but should take into account the relatively small size of dairy herds, the particular types of milk-recording data available, and standard lactation curves associated with the specific dairy-farming conditions.

The traditional approach to the acquisition of knowledge for KBS through interviews with domain specialists has proven to be difficult and time-consuming (Durkin, 1994; Dhar and Stein, 1997). Domain specialists often have difficulty expressing exactly how they make their decisions and it is not easy to organize and translate the knowledge expressed by the specialists into a representation that can be used in KBS. The elicitation of decision rules might be especially challenging with problem areas that involve the interpretation of graphical information, as in the case of group-average lactation-curve analysis. For a domain specialist it may be easy to take into account the large amount of information represented by a graph and classify the entire graph or a part of it as either normal or abnormal. However, the high information density of graphs makes it very difficult for the specialist to determine appropriate numerical features or attributes and to formulate rules for use in a KBS to automatically interpret the information described by the graph. An alternative approach to knowledge acquisition, that might be more appropriate for domains with graphical information, involves the application of machine learning to example cases classified by the domain specialist (Langley and Simon, 1995; Dhar and Stein, 1997). Machine-learning techniques are able to automatically generate a description of the knowledge embedded in the example cases to which they are applied. Decision-tree induction is an approach to machine learning that is particularly well suited to support knowledge acquisition. Decision trees tend to be easy to understand (Dhar and Stein, 1997; Kononenko et al., 1998; McQueen et al., 1995), which allows for evaluation of the learned knowledge by specialists and enables end-users of the KBS to view a justification of the decisions made. However, several challenges with the application of machine learning have been identified, including the decomposition of a complex problem into sub-problems, acquisition of an adequate number of labeled example cases of sufficient quality, deriving potentially predictive attributes, and analysis of the results of learning (Langley and Simon, 1995; Verdenius et al., 1997).

Research was initiated to explore the usefulness of machine-learning assisted knowledge acquisition for group-average lactation-curve analysis. A large amount of consultation with two dairy-nutrition specialists was required to elicit the domain vocabulary, decompose the overall problem area into three sub-problems (removal of outlier data, interpretation of group-average lactation curves, and diagnosis of detected abnormalities), and to develop a case-acquisition and decision-support system (CADSS) (Pietersma et al., 2001a). This CADSS included a graphical user-interface for each sub-problem, allowing users to interact with the information presented. In addition, the specialists identified several classification tasks for each sub-problem, each with a predefined number of classes. These tasks were included in the CADSS and functionality was added to capture the classifications made by a domain specialist. Although most of these tasks involved two classes, such as “True” and “False”, the interpretation sub-problem included many tasks with three or four classes. For example, the group-average peak production could be classified as “Low”, “Normal”, or “High”. Detailed analysis of the results of learning with these multi-class tasks remains a challenge since commonly used performance indices only apply to classification tasks involving two classes.

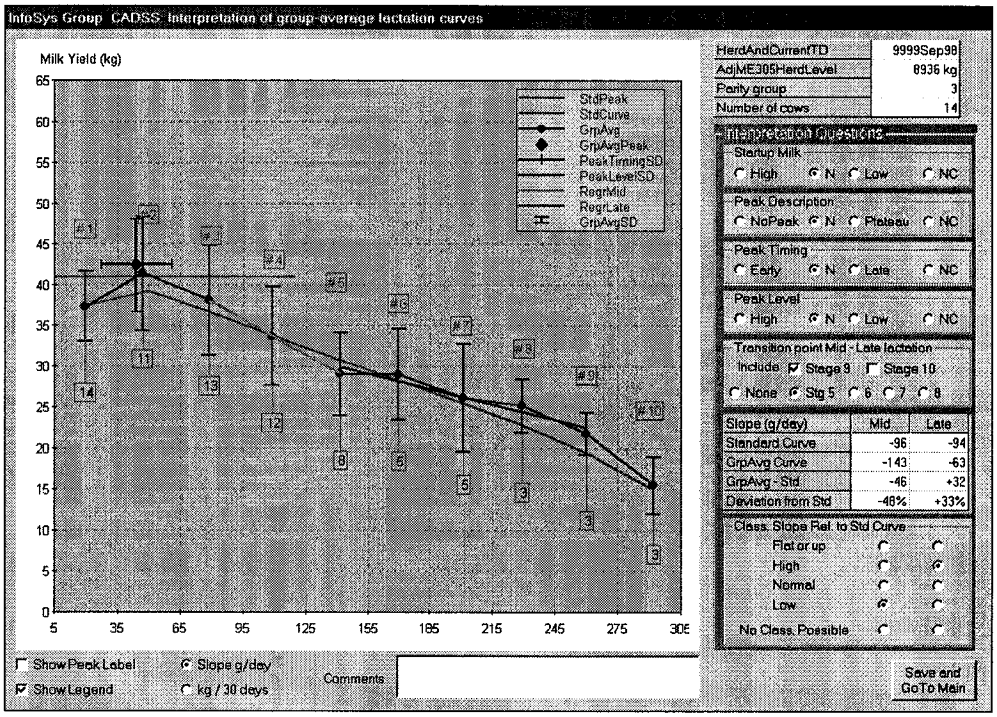
The goal of the project presented here was to develop a knowledge-based module for implementation in the CADSS to partially automate the interpretation of group-average lactation curves. The objectives were: 1) to induce decision trees for each classification task involved with the interpretation of group-average lactation curves; 2) to develop an approach to facilitate performance analysis for classification tasks involving more than two classes; 3) to determine the ability of the induced decision trees to mimic the classifications performed by the domain specialist; and 4) to verify with the specialist the plausibility of the induced decision trees.

7.2 Materials and Methods

7.2.1 Data

A dairy-nutrition specialist used the CADSS to analyze and classify the milk-recording data of 33 Holstein herds, enrolled with the provincial dairy herd analysis service (PATLQ). These herds represented a wide range of milk production levels. Within each herd and for each of three parity groups (parity 1, 2, and 3+), the lactation curve data of individual cows was first filtered by the domain specialist to remove outliers. The removal of outliers had been identified by domain specialists as the first step in the overall analysis process and considered important for relatively small sized dairy herds, to avoid the interpretation of the group-average performance being biased by a few atypical lactations or tests (Pietersma et al., 2001a). The CADSS allowed the specialist to compare the lactation curves of individual cows with group-average and standard lactation curves and view, for each individual test, additional information including the milk protein to fat ratio and the somatic cell count. In the second step of the overall analysis process, addressed in this research, non-outlier milk yield data was used to create, for each parity group, a group-average lactation curve and a group-average peak production. The specialist analyzed the group-average lactation descriptions for the 33 herds using the CADSS, which led to a total of 99 interpretations.

Figure 7.1 shows a screen capture of the CADSS module for the interpretation of group-average lactation curves. The group-average lactation curve represents the averaged non-outlier milk yield and days in milk values of individual cows within each of ten stages of lactation from 5 to 305 days in milk. The group-average peak level and timing, indicated with a diamond marker and both horizontal and vertical error bars in Figure 7.1, represent the mean of the maximum milk yield for the first 120 days in milk of individual cows, and the mean of the associated days in milk values, respectively. The error bars shown represent the group-average value plus and minus the standard deviation. In addition, the standard lactation curve and peak level are shown for the parity group in question and the milk production level of the herd.



InfoSys Group CADSS: Interpretation of group-average lactation curves

HerdAndCurrentTD	9999Sep98
AdjME305HerdLevel	8936 kg
Parity group	3
Number of cows	14

Interpretation Questions

Startup Milk
○ High ● N ○ Low ○ NC

Peak Description
○ NoPeak ● N ○ Plateau ○ NC

Peak Timing
○ Early ● N ○ Late ○ NC

Peak Level
○ High ● N ○ Low ○ NC

Transition point Mid - Late lactation
Include ☑ Stage 9 □ Stage 10
○ None ● Stg 5 ○ 6 ○ 7 ○ 8

Slope (g/day)	Mid	Late
Standard Curve	-96	-94
GrpAvg Curve	-143	-63
GrpAvg - Std	-46	+32
Deviation from Std	-48%	+33%

Class. Slope Rel. to Std Curve

Flat or up	○	○
High	○	●
Normal	○	○
Low	●	○
No Class. Possible	○	○

□ Show Peak Label
☑ Show Legend
● Slope g/day
○ kg / 30 days
Comments

Figure 7.1 Screen capture of the case-acquisition software module to interpret group-average lactation curves.

With the interpretation module, six main classification tasks had been identified by the specialists (Pietersma et al., 2001a). The first task involved the interpretation of the performance after calving using the so-called "Start-up milk" defined as the group-average milk yield for the first stage of lactation. The shape of the peak could then be interpreted using the classification task "Peak description". The third and fourth task involved the interpretation of the group-average peak timing and peak level, respectively. Finally, the shape of the group-average lactation curve after the peak could be interpreted as a whole or in two sections. The entire curve or the first section after the peak could be classified with the task "Slope mid lactation" and the second section after the peak with the task "Slope late lactation". To support the interpretation of the slope after the peak, three additional classification tasks were created. The specialist could include or exclude the group-average milk yield at stage 9 and at stage 10. In addition, a transition point between mid and late lactation could be identified (Figure 7.1).

For each task, a set of classes had been identified by the specialists (Pietersma et al., 2001a). For example, the peak level could be classified as either "High", "Normal" ("N" in Figure 7.1), "Low", or "Not Classifiable" ("NC" in Figure 7.1). Table 7.1 shows the distribution of cases per class for each classification task, as interpreted by the domain specialist. For example, for the task "Start-up milk", 60 lactation curves were considered "Normal", while 31 curves were interpreted as "Low" and 7 as "High". For each task, one case was interpreted as "Not Classifiable" due to limited data. In addition, for the tasks "Include stage 9" and "Include stage 10", several cases were not applicable due to the absence of a group-average milk yield for the stage in question. For many curves, the specialist did not specify a transition point between mid and late lactation (class "None" in Figure 7.1). In that case, the slope of the entire lactation after the peak was classified. As a result, the task "Slope late lactation" only consisted of the 62 cases for which the slope after the peak had been split into two sections.

Table 7.1 Classification tasks used by the domain specialist and number of cases per class.

Classification task	Class 0		Class 1		Class 2		Class 3		NC[†]
	Label	Cases (#)	Label	Cases (#)	Label	Cases (#)	Label	Cases (#)	Cases (#)
Start-up milk	Normal	60	Low	31	High	7			1
Peak description	Normal	62	No Peak	16	Plateau	20			1
Peak timing	Normal	59	Early	21	Late	18			1
Peak level	Normal	66	Low	28	High	4			1
Include stage 9	True	85	False	8					6
Include stage 10	False	65	True	11					23
Transition point	None	36	Stage 5	16	Stage 6	41	Stage 7	5	1
Slope mid lactation	Normal	45	Low	23	High	24	Flat	6	1
Slope late lactation	Normal	13	Low	17	High	27	Flat	5	37

[†] NC = Not classifiable.

7.2.2 Classification tasks for machine learning

In this research, seven of the nine classification tasks used by the domain specialist consisted of more than two classes (Table 7.1). However, methods for detailed analysis of the classification performance, as explained below, tend to be restricted to classification tasks with two classes. Thus, the following approach was developed to enable the use of two-class performance indices for multi-class tasks. Firstly, for tasks that consisted of a "Normal" class and two classes representing deviations from "Normal" in opposite directions, such as "High" and "Low" or "Early" and "Late", the three distinct classes were used for machine learning. However, for performance analysis, the two classes unequal to "Normal" were grouped into a single class called "Abnormal". Correctly classified "Abnormal" and "Normal" cases were considered as true positives and true negatives, respectively. With these tasks, misclassifications of, for example, a "Low" case as "High" or vice versa were not expected to occur, unless the data had been clearly mislabeled. This approach was used for the tasks "Start-up milk", "Peak timing", and "Peak level". For example, with the task "Start-up milk", the 60 "Normal" cases were considered as such during both decision tree induction and performance analysis. However, during tree induction, the 31 "Low" and 7 "High" cases were kept as distinct classes, while during performance analysis, these two classes were grouped together as "Abnormal" with a total of 38 cases (Table 7.2).

Table 7.2 Classification tasks used for machine learning and number of cases per class for machine learning and for performance analysis.

Classification task	Normal (machine learning and performance analysis)		Abnormal (machine learning)				Abnormal (performance analysis)
			Class 1		Class 2		
	Label	Cases (#)	Label	Cases (#)	Label	Cases (#)	Cases (#)
Start-up milk	Normal	60	Low	31	High	7	38
No peak	False	82	True	16			16
Plateau peak	False	62	True	20			20
Peak timing	Normal	59	Early	21	Late	18	39
Peak level	Normal	66	Low	28	High	4	32
Exclude stage 9	False	85	True	8			8
Single slope	False	62	True	36			36
Transition stage	Stage 6	41	Stage 5	16	Stage 7	5	21
Lactation slope	Normal	58	Low	40	High + Flat	62	102
Flat slope	False	51	True	11			11

For the four other multi-class tasks, the distinction between the classes unequal to "Normal" was not always obvious and misclassifications among these classes were thought to be quite possible. For these tasks, decomposition into a two-step process with two sub-tasks was used. For example, for the task "Peak description", the first sub-task determined whether a "No peak" description applied with classes "True" and "False". For the cases classified as "False" in the first step, a second sub-task determined whether the description should be "Plateau peak", again with classes "True" or "False" (Table 7.2). Cases classified as "False" in both steps represented a classification as "Normal" for the peak description task. The classification task "Transition point" was also decomposed into two steps: first to determine if a single slope after the peak should be considered, and, for the negative cases, to determine the specific transition stage, with stage 6 regarded as the default transition point between mid and late lactation. The two classification tasks "Slope mid lactation" and "Slope late lactation" were merged together and an additional attribute was used to identify whether the slope pertained to mid lactation, late lactation, or to a single slope after the peak. This resulted in 58 "Normal", 40 "Low", 51 "High", and 11 "Flat" cases. The merged task was then decomposed into two steps: first to determine the slope with classes "High" and "Flat" combined (Table 7.2), and, for cases considered either "High" or "Flat", a second task to determine whether the slope should be considered as "Flat".

For the task "Include stage 9", most cases were classified as "True" which was considered as the default situation (Table 7.1). To make the labeling of this task consistent with the other two-class tasks in this study, the labels "True" and "False" were reversed and the task was renamed as "Exclude stage 9" (Table 7.2). The domain specialist considered the classification task "Include stage 10" as having little influence on the classification of the slope after the peak. This task was, therefore, set by default to "False" and excluded from machine learning.

7.2.3 Creation of attributes

The domain specialist had access to complex graphical information to interpret the various aspects of the group-average lactation curves. The CADSS provided numerical data only for the slope during mid and late lactation (Figure 7.1). Thus, specific features or attributes had to be derived for each classification task to allow the machine-learning algorithm to learn to classify the information represented by the graphs. However, the attributes representing the raw data used to create the group-average and standard curves presented in the CADSS, such as the group-average milk yield, standard deviation of the milk yield, and days in milk for each of the ten stages of lactation, were expected to provide only limited discrimination ability. Thus, to make machine learning feasible with the relatively small number of example cases available for learning, considerable time was spent to craft attributes that were expected to be useful to discern between the classes.

With the CADSS, the domain specialist could compare the group-average performance standard lactation curves and peak levels that were used as benchmarks. Thus, to support machine learning, attributes were derived to represent this type of comparison. For example, for the task "Start-up milk", such attributes consisted of the deviation of the group-average start-up milk from the standard start-up milk, expressed in absolute terms, in relative terms, and as the number of standard deviations. Table 7.3 shows a listing of the attributes derived for machine learning, with codes such as "SM" for "Start-up milk", to indicate the classification task for which the attributes were used.

In addition to the attributes representing the deviation from a benchmark, several attributes related to the shape of the group-average curve were created. For example, for the start-up milk task, the domain specialist might take into account the group-average start-up milk yield in relation to the maximum milk yield of the group-average curve. Thus, a start-up milk yield expected with the observed maximum milk yield was estimated by adjusting the maximum group-average milk yield for stages two and three with the difference between the maximum and start-up milk yield for the standard curve. Three additional attributes were created, representing the deviation of the observed start-up milk yield from this expected value in absolute, relative, and number of standard deviation terms (Table 7.3). The attributes representing the deviation from benchmark performance were proposed by the system developer, but consultation with the domain specialist was required to create attributes related to the shape of the group-average lactation curve.

Table 7.3 Listing of potentially predictive attributes for each of the classification tasks used for machine learning.

Task[†]	Description of attribute or attributes
SM	Absolute (abs.), relative (rel.), and number of standard deviations (numSD) deviation of the group-average lactation curve (GrpAvgCrv) milk at stage 1 from the standard curve (StdCrv) milk at stage 1
SM	Abs., rel., and numSD deviation of GrpAvg milk at stage 1 from prediction based on maximum (max.) GrpAvg milk at stage 2 and 3 and the shape of StdCrv
NP, PP	Stage 1 has max. GrpAvgCrv milk
NP, PP	Abs., rel., and numSD deviation of GrpAvg milk at stage 1 from max. GrpAvg milk
NP, PP	Slope linear regression (LinRegr) GrpAvgCrv from stage 1 to 3, 1 to 4, 2 to 3, 2 to 4
NP, PP	Slope LinRegr through GrpAvgCrv from stage 1 to max. milk stage or from stage 1 to 2
NP, PP	Deviation slope LinRegr through GrpAvgCrv from slope StdCrve for stage 2 to 3 and 2 to 4
NP,PP,PT,PL	Days in milk, SD days in milk, milk, SD milk, and number of tests of GrpAvg peak
PT	Parity group
PL	Abs., rel., and numSD deviation of GrpAvg peak milk from Std peak
PL	Abs., rel., and numSD deviation of max. GrpAvgCrv milk from max. StdCrv milk
E9	SD and number of tests GrpAvgCrv at stage 9
E9	Abs., rel., and numSD deviation of GrpAvgCrv milk at stage 9 from prediction based on LinRegr through previous 2 and previous 3 stages
SS	Average (avg.) SD of stages of GrpAvgCrv after peak
SS	Max. numSD deviation of GrpAvgCrv milk from LinRegr GrpAvgCrv after peak
SS	Rel. deviation of root mean squares error (RMSE) for no transition point from minimum RMSE for any transition point
SS	Max. difference between rel. deviation of slope LinRegr GrpAvgCrv from slope of StdCrv for mid and late lactation, for transition points at stage 5, 6, and 7
SS, TS	Rank of RMSE of LinRegr GrpAvgCrv after peak or avg. RMSE for two regression lines for mid and late lactation for transition points at stage 5, 6, and 7
TS	Rank of the avg., max., or difference for the rel. deviation of slope LinRegr GrpAvgCrv from slope of StdCrv for mid and late lactation, for transition points at stage 5, 6, and 7
LS, FS	Type of section of GrpAvgCrv after peak (mid + late lactation, mid lactation, late lactation)
LS, FS	Avg. SD of GrpAvgCrv for section
LS, FS	Abs., rel., and numSD deviation slope GrpAvgCrv from slope StdCrv for section
FS	Slope GrpAvgCrv for section

[†] Task: classification task for which attributes were used: SM = start-up milk; NP = no peak; PP = plateau peak; PT = peak timing; PL = peak level; E9 = exclude stage 9; SS = single slope after peak; TS = transition stage between mid and late lactation; LS = normal or abnormal slope for mid + late, mid, or late lactation; FS = flat slope for mid + late, mid, or late lactation.

For some cases, certain attributes had attribute values that could not be determined. For example, the standard deviation of the group-average milk yield at a particular stage in lactation for which only one test was available could not be calculated. For such situations, a special value, such as 9999, was used to indicate the irrelevance (Pietersma et al., 2001b; Witten and Frank, 2000).

7.2.4 Decision-tree induction algorithm

In this study decision trees were induced with CART for Windows version 3.6 developed by Salford Systems (Breiman et al., 1984; Steinberg and Colla, 1997). The algorithm learns in a top-down fashion, by splitting the training data into two subsets recursively, choosing the attribute and value that is most successful in discriminating among the classes of the classification problem at each split. The CART algorithm continues splitting subsets until a maximum tree is reached, which is pruned back to the optimal size, determined through an internal training and testing procedure, to avoid overfitting the training data. The resulting decision tree consists of a series of decision nodes that, during classification, guide each new case to a leaf node indicating the predicted class.

Preliminary experiments were performed to tune the settings of the parameters of the algorithm to the type of classification tasks involved in this research. The same parameter configuration was used for all classification tasks and consisted of the so-called "Gini" splitting and pruning criterion, the minimum number of cases at a child node set at 3, and the minimum number of cases at a parent node set at 6. In addition, the parameter for the prior probability of the class representing a normal situation was set to the observed frequency in the data set for that class, while equal prior probability values were used for each of the remaining classes. The misclassification cost parameters were used to focus the decision-tree algorithm on correctly classifying one particular class over other classes. For the remaining parameters of the algorithm, the default settings were used. A thorough description of the CART algorithm can be found in Breiman et al. (1984) and Steinberg and Colla (1997).

7.2.5 Training and testing method

For relatively small data sets, the ten-fold cross-validation approach to training and testing has often been recommended (Breiman et al., 1984; Weiss and Kulikowski, 1991). With this approach, the entire data set is divided into ten subsets or folds, and each fold is used once for testing the classifier trained from the combined data of the remaining folds. The cross-validation performance can then be used as an estimate of the performance of the final classifier that is generated from the entire data set to classify new cases in the real world. However, in preliminary experiments, ten-fold cross-validation, with approximately 10 cases in each test fold, resulted in a large variability in the performance estimates from fold to fold. Thus, five-fold cross-validation was used, with two times as many cases per test set than were available with ten-fold cross-validation, and less variability in the performance estimates. Although five-fold cross-validation uses 80% of the entire data set for training (instead of 90% with ten-fold cross-validation), the performance was considered a fairly good estimate of the performance of classifiers generated from the entire data set. For each classification task, entire herds were randomly assigned to folds to avoid example cases of the same herd being part of both the training and the test sets, potentially leading to a biased estimate of the performance on data from entirely new dairy herds (Kubat et al., 1998; Pietersma et al., 2001b). To achieve approximately the same class distribution in each fold as in the entire data set, herds were first ranked according to the prevalence of the classes, followed by assigning the first five herds to folds one through five, respectively, and so on.

7.2.6 Performance analysis

With machine learning, accuracy, defined as all correctly classified cases as a proportion of all classified cases, is often used as a criterion to assess the performance of the generated classifiers (Weiss and Kulikowski, 1991; Witten and Frank, 2000). However, in real world applications, some types of misclassification may be worse than others. For example, it may be more costly to classify a person with a serious disease as healthy than to classify a healthy person as having that disease. Machine-learning algorithms can often deal with such situations by focussing more on correctly classifying one particular class at the expense of misclassifying the other class or classes. Performance indices have been developed to deal with this trade-off, but tend to be limited to classification tasks for which a case is either positive or negative. In this study, most classification tasks consisted of more than two classes, but an approach was developed to enable the use of performance analysis tools designed for two classes with multi-class problems, as explained above.

With classification tasks involving two classes, true positives (TP) and true negatives (TN) are correct classifications, a false positive (FP) is an actual negative case incorrectly predicted as positive, and a false negative (FN) is an actual positive case predicted as negative. To allow for detailed analysis of the performance of the generated classifiers, the following four performance indices were used: 1) the true positive rate (TP rate), defined as TP / (TP + FN); 2) the false positive rate (FP rate), defined as FP / (FP + TN); 3) the predicted value positive (PVP), defined as TP / (TP + FP); and 4) the positive prediction rate (PPR), defined as (TP + FP) / (TP + FP + FN + TN) (Pietersma et al., 2001b; Swets, 1988; Weiss and Kulikowski, 1991). In some domains, the TP rate is referred to as the sensitivity and the FP rate as 1 – specificity. The prevalence of positive cases or prior probability of positives was estimated from the available training data as the actual positives as a proportion of all cases. The TP rate and FP rate are both independent of the prevalence of positive cases and are, thus, the characteristics of the classifier (Swets, 1988). Conversely, the PPR and PVP depend on the prevalence of positive cases and can be mathematically derived from the TP rate and FP rate for a given prevalence level using:

$$PPR = Prevalence\ of\ positives \times TP\ rate + (1 - Prevalence) \times FP\ rate$$

$$PVP = Prevalence\ of\ positives \times TP\ rate\ /\ PPR.$$

Relative operating characteristic (ROC) curves (Swets, 1988) were used to visualize the trade-off between correctly classifying normal cases and correctly classifying abnormal cases. An ROC curve consists of the TP rate plotted against the FP rate. The point (0,100) represents perfect classification performance. Thus, the closer a curve approximates a line connecting (0,0), (0,100) and (100,100), the better the performance. Each point on the ROC curve represents a classifier with a particular trade-off between sensitivity and specificity. To generate an ROC curve, a series of ten decision trees was generated using ten different settings for the CART parameters specifying the cost of mistakenly classifying an abnormal case as normal. The cost of classifying a normal case as abnormal was fixed at one.

For classification tasks consisting of three classes, the same value was used for the two misclassification cost parameters associated with misclassifying an abnormal case as normal. In addition, a very high value was used for the two cost parameters associated with misclassifying one abnormal class as the other, e.g. classifying a “High” peak level as “Low” and vice versa, to entirely avoid such misclassifications. The five-fold cross-validation provided five estimates of the FP rate and the TP rate, for each of the ten misclassification cost levels. These five estimates were averaged at each cost level, resulting in ten data points in ROC space, which were connected to get an ROC curve (Bradley, 1997).

7.2.7 Final decision trees induced from the entire data set

In a practical application, the desired trade-off between sensitivity and specificity might depend on factors such as the prevalence of positive cases for a particular herd and end-user preference regarding the number of false positives. To allow end-users to choose classifiers at different points along the ROC curve, a series of three final decision trees associated with an increasing cost of misclassifying abnormal cases was induced from the entire data set for each classification task. These three trees represented a low, medium, and high tendency of classifying new cases as abnormal. The medium tendency trees were chosen to represent a trade-off between sensitivity and specificity that would result in a PPR approximately equal to the observed prevalence of abnormal cases, given a population with that prevalence.

To evaluate the plausibility of these final decision trees, quantitative and qualitative assessments were carried out. Although the true performance with new data of a decision tree induced from the entire data set can only be estimated, the so-called resubstitution performance can be determined through testing on the training set (Witten and Frank, 2000). Resubstitution FP and TP rates were used to quantitatively verify how closely the classification performance of the decision trees induced from the entire data set resembled the performance of the cross-validated decision trees. This allowed for manual adjustment of the level of pruning of these trees to achieve the intended sensitivity versus specificity trade-off (Pietersma et al., 2001b). In addition, the final decision trees were evaluated by the domain specialist to qualitatively verify the plausibility of the induced rules. This allowed for the removal of counter-intuitive decision nodes or use of alternative splits provided by the CART algorithm.

7.3 Results

7.3.1 Classification performance

7.3.1.1 Start-up milk

Good classification performance was obtained for the task "Start-up milk", with 89% TP rate achieved at 3.6% FP rate (Figure 7.2). Three different points along the ROC curve, each indicated with a marker in Figure 7.2 and associated with a different setting for misclassification costs, were chosen to induce final decision trees from the entire data set representing a low, medium, and high tendency of classifying new cases as abnormal.

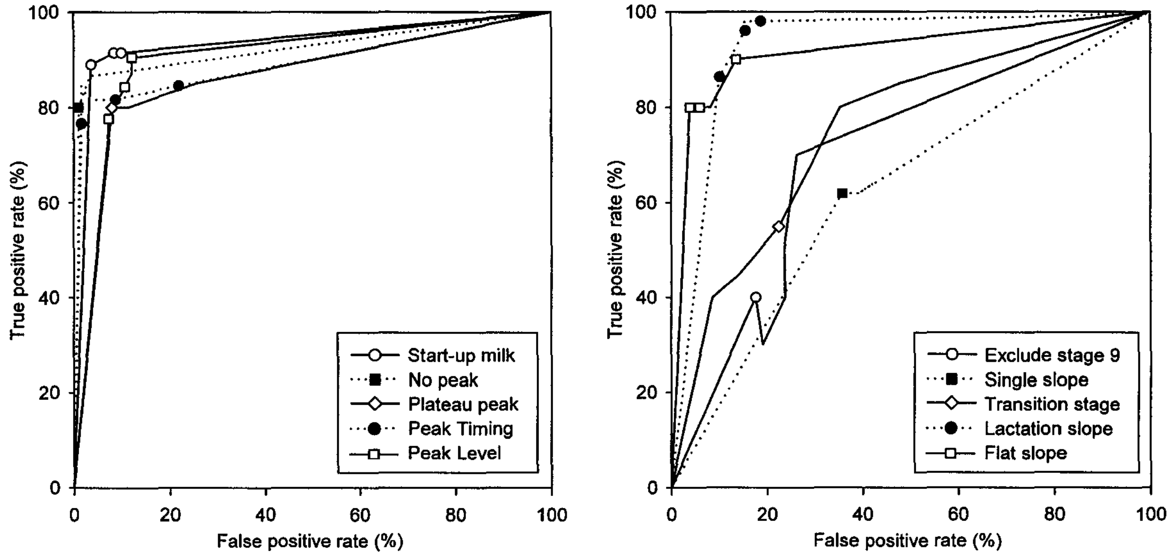


Figure 7.2 Relative operating characteristic curves for 10 classification tasks with markers showing the cross-validation performance for each induced decision tree.

The misclassification cost settings used to induce trees with a low, medium, and high tendency of classifying new cases as abnormal were 0.19, 0.5, and 1, respectively, and these final trees consisted of 3, 4, and 4 leaf nodes, respectively (Table 7.4). The cross-validation estimates of FP rate for this task ranged from 4 to 10%, and the estimates for the TP rate ranged from 89 to 91%. Given the observed 39% prevalence of abnormal classes, the three decision trees were expected to classify 37%, 41%, and 42%, respectively, of the cases as abnormal (PPR). Of those cases predicted as abnormal, 94%, 88%, and 86%, respectively, were expected to be truly abnormal (PVP).

Table 7.4 Cross-validation performance of decision trees induced for each classification task and associated with a low, medium, or high tendency of classifying new cases as abnormal.

Classification task	Preval-ence[†] (%)	Type of tree	Cost FN	Leaf nodes (#)	False positive rate (%) (s.e.)	True positive rate (%) (s.e.)	PPR (%)	PVP (%)
Start-up milk	39	Low	0.19	4[‡]	3.6 (3.6)	88.9 (8.3)	36.9	94.0
Start-up milk	39	Medium	0.5	4	8.4 (3.8)	91.4 (8.6)	40.8	87.5
Start-up milk	39	High	1	4	9.9 (4.7)	91.4 (8.6)	41.7	85.5
No peak	16	Medium	1.5	3	1.1 (1.1)	80.0 (13.3)	13.7	93.2
Plateau peak	24	Medium	1	3	8.1 (4.6)	80.0 (9.5)	25.3	75.8
Peak timing	40	Low	0.13	3	1.7 (1.7)	76.6 (5.4)	31.7	96.8
Peak timing	40	Medium	2	5	8.8 (5.0)	81.6 (5.5)	37.9	86.1
Peak timing	40	High	4	7[‡]	21.9 (8.3)	84.5 (4.8)	46.9	72.0
Peak level	33	Low	0.6	7	7.4 (2.3)	77.6 (8.7)	30.6	83.8
Peak level	33	Medium	0.8	5[‡]	10.7 (2.0)	84.3 (5.3)	35.0	79.5
Peak level	33	High	5	4	12.1 (2.9)	90.5 (3.9)	38.0	78.6
Exclude stage 9	9	Medium	5	5	17.6 (1.7)	40.0 (18.7)	19.6	18.3
Single slope	42	Medium	1	4	35.8 (8.5)	61.9 (9.0)	46.8	55.6
Transition stage	35	Medium	1.4	5	22.5 (9.8)	55.0 (9.4)	33.9	56.8
Lactation slope	64	Low	0.1	7[‡]	10.2 (6.1)	86.3 (4.5)	58.9	93.8
Lactation slope	64	Medium	0.5	5	15.5 (5.5)	96.0 (1.9)	67.0	91.7
Lactation slope	64	High	2	4	18.8 (4.0)	98.0 (1.2)	69.5	90.3
Flat slope	18	Low	0.3	2	4.0 (4.0)	80.0 (12.2)	17.7	81.4
Flat slope	18	Medium	1.5	4[‡]	6.0 (4.0)	80.0 (12.2)	19.3	74.5
Flat slope	18	High	5	2	13.6 (2.0)	90.0 (10.0)	27.3	59.3

[†] Prevalence: prevalence of abnormal class or classes; Cost FN: cost of false negatives relative to cost of false positives; PPR: positive prediction rate; PVP: predictive value positive.

[‡] Size of decision tree was manually adjusted.

Figure 7.3 shows the decision trees for the "Start-up milk" classification task, induced from the entire data set and representing a low (tree A), medium (tree B), and high (tree C) tendency of classifying new cases as abnormal. The first decision node of each tree shows the class distribution observed in the entire data set for the classes "High", "Normal", and "Low", and the attribute and threshold value considered by the decision-tree induction algorithm as being most successful to discriminate between the three classes. The two subsets resulting from the chosen split are considered as either a final leaf node, in which case the predicted class is shown, or split again using another attribute-value combination.

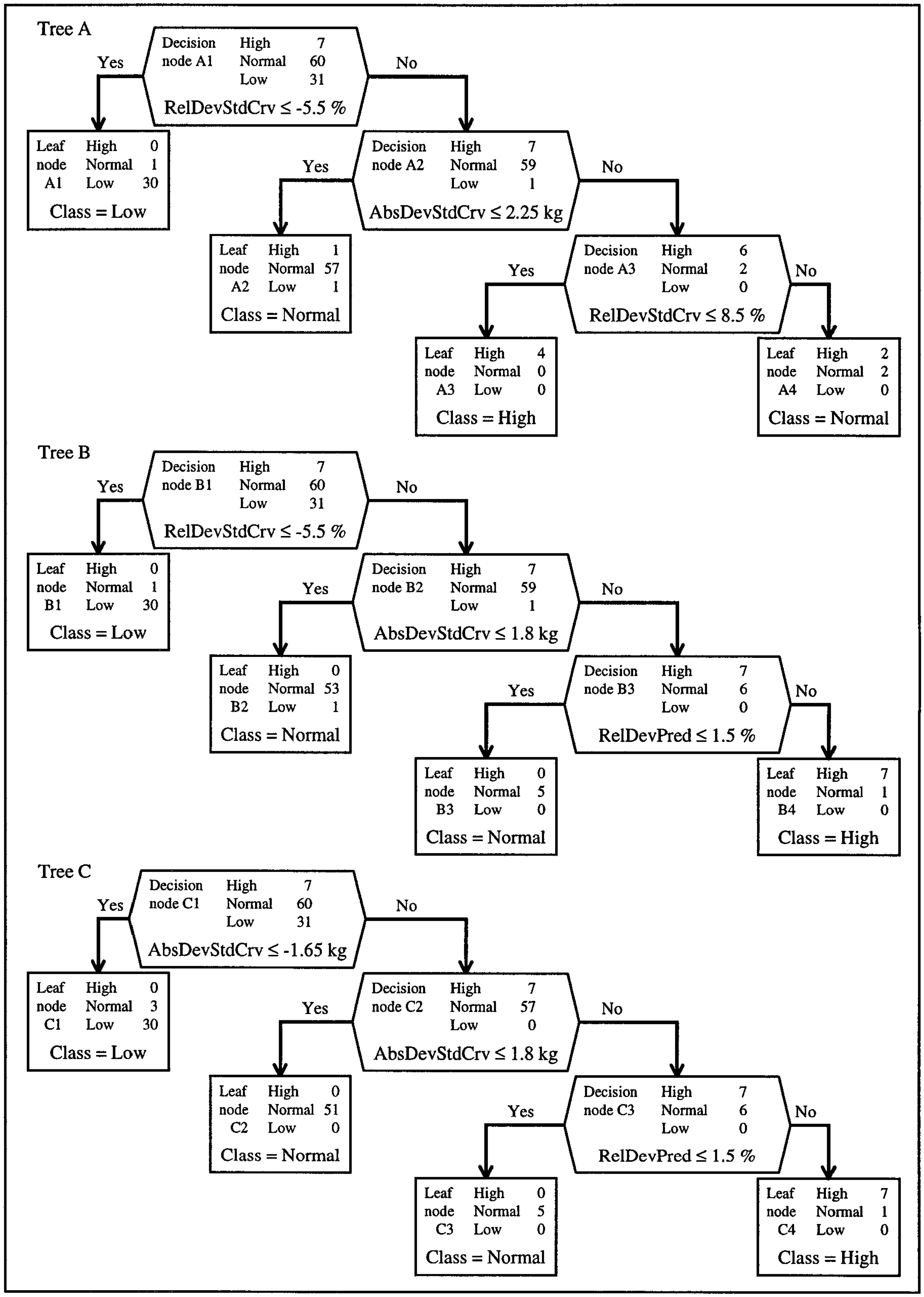

Tree A
Decision node A1 High 7 Normal 60 Low 31
RelDevStdCrv ≤ -5.5 %
Yes
No
Leaf node A1 High 0 Normal 1 Low 30
Class = Low
Decision node A2 High 7 Normal 59 Low 1
AbsDevStdCrv ≤ 2.25 kg
Yes
No
Leaf node A2 High 1 Normal 57 Low 1
Class = Normal
Decision node A3 High 6 Normal 2 Low 0
RelDevStdCrv ≤ 8.5 %
Yes
No
Leaf node A3 High 4 Normal 0 Low 0
Class = High
Leaf node A4 High 2 Normal 2 Low 0
Class = Normal
Tree B
Decision node B1 High 7 Normal 60 Low 31
RelDevStdCrv ≤ -5.5 %
Yes
No
Leaf node B1 High 0 Normal 1 Low 30
Class = Low
Decision node B2 High 7 Normal 59 Low 1
AbsDevStdCrv ≤ 1.8 kg
Yes
No
Leaf node B2 High 0 Normal 53 Low 1
Class = Normal
Decision node B3 High 7 Normal 6 Low 0
RelDevPred ≤ 1.5 %
Yes
No
Leaf node B3 High 0 Normal 5 Low 0
Class = Normal
Leaf node B4 High 7 Normal 1 Low 0
Class = High
Tree C
Decision node C1 High 7 Normal 60 Low 31
AbsDevStdCrv ≤ -1.65 kg
Yes
No
Leaf node C1 High 0 Normal 3 Low 30
Class = Low
Decision node C2 High 7 Normal 57 Low 0
AbsDevStdCrv ≤ 1.8 kg
Yes
No
Leaf node C2 High 0 Normal 51 Low 0
Class = Normal
Decision node C3 High 7 Normal 6 Low 0
RelDevPred ≤ 1.5 %
Yes
No
Leaf node C3 High 0 Normal 5 Low 0
Class = Normal
Leaf node C4 High 7 Normal 1 Low 0
Class = High

Figure 7.3 Decision trees for the “Start-up milk” classification task with a low (A), medium (B), and high (C) tendency of classifying new cases as abnormal.

These three trees illustrate how the trade-off between correctly classifying normal and abnormal cases is made by the algorithm. Decision trees A and B use the same attribute and value at the first decision node (relative deviation of start-up milk from the standard curve ≤ –5.5%), predicting 31 cases as "Low", while decision tree C uses a slightly more aggressive split (absolute deviation of start-up milk from standard curve ≤ –1.65 kg) predicting 34 cases as "Low". The second decision node for tree A (absolute deviation of start-up milk from standard curve ≤ 2.25 kg) predicts 59 cases as "Normal", while decision tree B uses a lower threshold value, 1.8 kg, for the same attribute, leading to the prediction of 54 cases as "Normal". Thus, by using slightly different attributes and threshold values, each decision tree makes a different trade-off between correctly classifying normal and abnormal cases. For the entire data set, the decision trees with a low, medium, and high tendency of classifying new cases as abnormal predicted 35, 39, and 41 cases, respectively, as either "High" or "Low" (Figure 7.3).

7.3.1.2 No peak and Plateau peak

For the "No peak" classification task the ROC curve revealed that fairly good performance was obtained (Figure 7.2), with 80% TP rate achieved at 1% FP rate (Table 7.4). Somewhat poorer classification performance was obtained for the "Plateau peak" task, with 80% TP rate achieved at 8% FP rate. For each of these two classification tasks, the PPR values of the trees with different misclassification cost settings were very similar. Thus, in both instances, only a single decision tree was generated from the entire data set. The trees for both tasks had an expected PPR fairly similar to their prevalence of positive cases. The PVP was very good (93%) for the "No Peak" task and reasonable (76%) for the "Plateau peak" task (Table 7.4).

7.3.1.3 Peak timing and Peak level

Fairly good performance was obtained for the "Peak timing" task. A 77% TP rate was achieved at 2% FP rate, while 85% TP rate required a relatively high FP rate of 22% (Table 7.4). The lowest misclassification cost level for the "Peak level" task showed 78% TP rate, which was similar to the one achieved for the "Peak timing" task, but at a much higher FP rate, 7%, instead of 2%. However, the ROC curve for "Peak level" crossed the curve for "Peak timing" (Figure 7.2), reaching 91% TP rate at 12% FP rate. For both classification tasks, three decision trees were induced from the entire data set with reasonable values for PPR and PVP (Table 7.4).

7.3.1.4 Exclude stage 9, Single slope, and Transition stage

The cross-validation experiments for the "Exclude stage 9" task resulted in poor classification performance (Figure 7.2) and a very high standard error for the TP rate estimate (Table 7.4). This may have been caused by the very small number of positive cases (8). For this classification task, several decision trees were induced from the entire data set at different misclassification cost levels and shown to the domain specialist for evaluation. The decision tree induced at cost level 5 with 5 leaf nodes was considered as most plausible. A similar situation occurred for the "Single slope" and "Transition stage" tasks. For these tasks poor classification performance was obtained (Figure 7.2), and, for each task, a single decision tree was chosen in consultation with the domain specialist (Table 7.4). These three classification tasks are only of indirect importance for the interpretation of group-average lactation curves. They allow for the calculation of a slope through linear regression for mid, late, or mid and late lactation combined, thus supporting the classification of the slope of the lactation curve after the peak. Of these tasks, determining whether "Single slope" is "True" or "False" seems most important, since the prediction of a single slope precludes the classification of mid lactation as being different from late lactation. Thus, for the task "Single slope", a plausible final tree was chosen with a relatively low tendency to predict class "True". This tree had a resubstitution FP rate of 11% with 61% TP rate, incorrectly classifying only 6 cases of the entire data set as "True".

7.3.1.5 Lactation slope and Flat slope

Good classification performance was obtained for the "Lactation slope" and "Flat slope" tasks, with TP rates higher than 80% at relatively low FP rates (Figure 7.2). For each task, a series of three decision trees was induced from the entire data set with reasonable PPR and PVP values, except for the tree for "Flat slope" with a high tendency of classifying new cases as "True" (Table 7.4). This tree showed a relatively high FP rate considering the low prevalence of positive cases, resulting in a poor value (59%) for the PVP.

7.3.2 Quantitative evaluation of the plausibility of the final decision trees

Quantitative assessment of the plausibility of the decision trees induced from the entire data set showed that for most of these trees, the resubstitution performance was very similar to the resubstitution FP and TP rates observed in the cross-validation. For example, for the "Start-up milk" task, the final tree for the medium tendency of indicating a case as abnormal had a resubstitution FP rate of 3.3% and a TP rate of 97.4%, very similar to the average resubstitution performance observed with cross-validation, with values of 2.6% and 98.4%, respectively (Table 7.5). However, for 5 of the 17 final trees induced for the main classification tasks, the resubstitution performance of the final tree was quite different from what was expected based on the cross-validation. For example, for the "Start-up milk" task, the misclassification cost setting associated with a low tendency of predicting cases as abnormal resulted in a final tree with a resubstitution FP rate of 5.0%, which was much higher than the average 1.7% FP rate of the cross-validation, and also higher than the 3.3% resubstitution FP rate of the final tree for the medium tendency (Table 7.5). This may have been caused by the internal 10-fold cross-validation used by CART to determine the appropriate level of pruning of the maximum tree, which can result in smaller or larger trees due to the differences in training data between the entire data set and the smaller cross-validation data sets. To better reflect the desired sensitivity versus specificity trade-off achieved in the cross-validation for the series of three decision trees, a larger tree with 4 instead of 3 leaf nodes, with an associated resubstitution FP rate of 1.7% was manually chosen from the decision trees provided by CART (Table 7.5). This means that one less decision node was pruned from the maximum tree than considered optimum by CART. For four additional final trees, the level of pruning of the maximum tree induced by CART was manually adjusted. Although these pruning adjustments were somewhat subjective, they were considered important to achieve a series of three final trees for each classification task, with an increasing tendency of indicating a new case as abnormal and the desired trade-off between sensitivity and specificity.

Table 7.5 Size and resubstitution performance of cross-validation decision trees and of optimal and size-adjusted decision trees induced from the entire data set.

		Cross-validation			Optimum size			Adjusted size		
Classification task	Type of tree	Leaf nodes (#) (s.e.)	FP† rate (%) (s.e)	TP rate (%) (s.e.)	Leaf nodes (#)	FP rate (%)	TP rate (%)	Leaf nodes (#)	FP rate (%)	TP rate (%)
Start-up milk	Low	3.6 (0.4)	1.7 (0.4)	96.7 (0.8)	3	5.0	94.7	4	1.7	89.5
Start-up milk	Medium	3.8 (0.2)	2.6 (0.8)	98.4 (1.0)	4	3.3	97.4			
Start-up milk	High	3.6 (0.2)	4.3 (0.7)	100.0 (0.0)	4	6.7	100.0			
No peak	Medium	3.0 (0.3)	0.6 (0.4)	95.3 (1.9)	3	0.0	93.8			
Plateau peak	Medium	2.8 (0.2)	4.4 (1.5)	85.0 (1.5)	3	3.2	85.0			
Peak timing	Low	3.8 (0.2)	0.0 (0.0)	85.7 (1.4)	3	1.7	84.6			
Peak timing	Medium	4.8 (0.4)	2.6 (2.6)	92.8 (2.9)	5	1.7	92.3			
Peak timing	High	5.6 (1.0)	9.3 (2.9)	97.5 (2.5)	9	3.4	97.4	7	13.6	97.4
Peak level	Low	5.4 (0.5)	1.9 (1.5)	90.2 (2.9)	7	1.5	84.3			
Peak level	Medium	4.0 (0.3)	5.0 (1.7)	96.4 (1.7)	7	1.5	84.3	5	6.1	96.9
Peak level	High	3.6 (0.2)	7.7 (1.1)	100.0 (0.0)	4	9.1	100.0			
Lactation slope	Low	6.4 (0.7)	0.4 (0.4)	88.1 (4.1)	8	1.7	93.1	7	1.7	88.2
Lactation slope	Medium	4.4 (0.6)	3.2 (1.2)	98.1 (0.8)	5	6.9	98.0			
Lactation slope	High	3.4 (0.2)	6.4 (0.9)	100.0 (0.0)	4	10.3	99.0			
Flat slope	Low	2.0 (0.0)	0.0 (0.0)	83.9 (4.7)	2	0.0	81.8			
Flat slope	Medium	2.4 (0.4)	0.5 (0.5)	88.3 (5.3)	2	0.0	81.8	4	2.0	100.0
Flat slope	High	2.0 (0.0)	8.4 (2.2)	100.0 (0.0)	2	9.8	100.0			

† FP rate: false positive rate; TP rate: true positive rate.

7.3.3 *Evaluation of learned knowledge by domain specialist*

The final decision trees induced from the entire data were evaluated by the domain specialist to verify their plausibility for application with new data. This resulted in the adjustment of 6 different decision nodes in 6 of the 20 decision trees: 3 decision nodes were removed and 3 decision nodes were replaced with an alternative attribute and threshold value, provided by the CART algorithm. For example, for the tree with a low tendency of classifying new cases as abnormal for the "Start-up milk" task in Figure 7.3, the decision node A3 (relative deviation of start-up milk from standard curve ≤ 8.5%) was not expected to properly classify new data. This node classifies cases with a value below this threshold as "High" and cases above this threshold as "Normal", which was considered as counter-intuitive. This may have been due to some inconsistencies in the labeling of the data, causing the algorithm to choose this split and class assignment at this section of the tree. This decision node was replaced with an alternative split provided by CART (absolute deviation from the predicted group-average start-up milk ≤ 2.4 kg) with cases below and above this threshold classified as "Normal" and "High", respectively. Although this alternative split resulted in one additional false negative case for the entire data set, the decision node and class assignment was considered as plausible by the domain specialist and expected to lead to improved classification performance on new, unseen, data.

7.4 Discussion

The cross-validation experiments resulted in good classification performance for the decision trees of the main classification tasks (start-up milk, no peak, plateau peak, peak timing, peak level, lactation slope, and flat slope). For these tasks, decision trees with a PPR similar to the prevalence of positive cases observed in the entire data set had a TP rate of at least 80% and a PVP of at least 75%, which was considered very reasonable. For the classification tasks indirectly affecting the classification of the lactation slope after the peak (exclude stage 9, single slope, and transition stage), the cross-validation performance was very poor. This may have been caused by factors such as the small number of cases in the minority class, lack of predictive attributes, and inconsistencies in the labeling by the domain specialist. For each of these three tasks, the expertise of the domain specialist was required to choose a decision tree that was expected to perform reasonably well on new data, from alternative trees generated by the decision-tree induction algorithm.

For three of the seven multi-class tasks available to the domain specialist, use of commonly employed two-class performance indices and ROC curves was possible by considering the classes other than "Normal", as "Abnormal" during performance analysis. However, the remaining four multi-class tasks had to be reformulated into a series of two- or three-class tasks. This additional task decomposition reduced the complexity for machine learning and also facilitated the induction of decision trees with a different trade-off between correctly classifying normal and abnormal cases. However, this came at the expense of additional cross-validation experiments and analyses of results of learning.

For some tasks and misclassification cost settings, the decision tree induced from the entire data set had a size and resubstitution performance very different from the average of the cross-validation trees, which was also observed in a previous study

(Pietersma et al., 2001b). For these situations, the pruning level of the maximum tree was manually adjusted to better reflect the desired trade-off between correctly classifying normal and abnormal cases for each of the three final decision trees of a classification task. However, these adjustments required time-consuming analysis of the size and resubstitution performance of both the cross-validation trees and the decision trees induced from the entire data set.

In this study, relatively small-sized decision trees, two to seven leaf nodes, were induced. This was likely due to the detailed decomposition of the problem into classification tasks with relatively low complexity. Less decomposition might have been possible as well, but would have involved more complex class descriptions, such as "High peak, Low slope mid lactation, and High slope late lactation". However, due to the increased complexity, such an approach was expected to require many more example cases to achieve the same classification performance.

Evaluation of the final trees by the domain specialist was considered an important step in the overall process. Several counter-intuitive decision nodes, which tended to occur at the end of the decision trees with limited data in the parent nodes, were manually removed or replaced with a substitute. These adjustments reduced the resubstitution performance on the entire data set, but, relying on the expertise of the domain specialist, were expected to lead to improved performance with new data.

For five of the seven main classification tasks, the machine-learning approach to knowledge acquisition allowed for the induction of a series of classifiers with an increasing tendency to classify a new case as abnormal. Implementation of these alternative decision trees for each classification task in the final KBS for group-average lactation-curve analysis allows end-users to move along the ROC curve and use the classifiers with the desired sensitivity versus specificity trade-off. For dairy herds with many abnormalities, the user may want to focus on the most obvious problems and use decision trees with a low tendency of indicating abnormal situations. Conversely, for dairy herds with few abnormalities, use of decision trees with a high tendency of indicating abnormal situations would support the user to find more subtle deviations, although at the expense of an increased probability of false positives.

In the final KBS, the classifications of the group-average lactation curves are used in a subsequent software module to determine potential management deficiencies (Pietersma et al., 2001a). Given the good classification performance of the decision trees, end-users of the final KBS might rely on the interpretations made automatically and skip the interpretation module to move directly to the final module for the diagnosis of detected abnormalities. However, end-users are able to view the group-average lactation curves and, if necessary, override the classifications made automatically and choose classes different from those suggested.

Although machine-learning assisted knowledge acquisition proved to be a very feasible approach to support the development of KBS in this research, several limitations were encountered. First of all, since previously classified example cases were not available, case-acquisition functionality had to be added to a KBS prototype to enable a domain specialist to analyze and classify a substantial number of lactation curves efficiently (Pietersma et al., 2001a). Secondly, the preprocessing of acquired example cases, including the creation of potentially predictive attributes, the learning experiments to tune algorithm parameters and to determine the expected performance with new data, and the evaluation of the learned knowledge proved to be quite time-consuming. Finally, although machine learning automated part of the knowledge acquisition process, a large amount of interaction between system developer and domain specialists remained necessary. Input from the specialist was required to decompose the overall problem into sub-problems, to identify classification tasks and their classes, to analyze and classify example cases, to support the creation of potentially predictive attributes, and for the qualitative evaluation of the plausibility of the results of learning. Thus, as with traditional interview-based knowledge acquisition, the ability of the system developer to communicate with the domain specialist was considered a critical success factor in machine-learning assisted KBS development.

In this study, the problem domain involved analysis of graphical performance representations, such as the slope after the peak, and the interpretation of new performance indices, such as the description of the lactation curve around peak production. Both aspects make it very difficult for a domain specialist to provide exact rules describing how to interpret the data. The interpretation of graphical performance representations tends to be difficult to translate into rules using numeric performance indices for use in a computer system. Also, when dealing with a novel approach to analyzing data, the domain knowledge is poorly formalized and new knowledge must be created to solve the problem (Weiss and Kulikowski, 1991). Thus, for problem areas involving interpretation of graphical performance representations or novel performance indices, machine-learning assisted knowledge acquisition is expected to be more useful than the traditional, interview-based, approach to KBS development.

7.5 Conclusions

This research suggests that automatically induced decision trees are able to closely match the interpretation of group-average lactation curves as performed by a domain specialist. However, considerable effort can be required for data preprocessing, for machine-learning experiments to determine the expected classification performance, and for evaluation of the learned knowledge. In addition, the interaction between system developer and domain specialist remains essential to achieve successful results. The induction of a series of three decision trees for each classification task allowed end-users of a final KBS to select classifiers with the appropriate tendency of classifying aspects of the lactation curve as abnormal. The machine-learning assisted approach to knowledge acquisition is expected to be appropriate in other areas of agriculture as well, especially when the problem domain involves analysis of graphical performance representations or a novel approach to data analysis.

8 General Discussion

The main goal of this research was to explore the use of machine learning to support the development of knowledge-based systems in dairy farming. The preceding five chapters reported on the investigations carried out to achieve this goal. This chapter addresses several main points of discussion that emerged from these investigations. These points focus on the method to estimate classification performance, the advantages and limitations of machine-learning assisted knowledge acquisition, and suggestions for further system development and research.

8.1 Method of performance estimation

In this research, the cross-validation approach to training and testing with five or ten folds was used to estimate the performance of classifiers generated with machine learning. With this approach, five or ten independent performance estimates were obtained, which allowed for the calculation of a mean and standard error of the performance indices used. However, in this research, a large amount of variability was observed for the estimates of the true positive rate and, to a lesser degree, also for the mean true positive rate covering a range of false positive rates (TP*). A high standard error indicates that the mean performance estimate is not very precise: the true performance on new data could be substantially higher or lower than estimated.

For the final decision trees generated in this research, the standard error of the true positive rate estimates was on average 7.7% and ranged from 1.2 to 18.7%. Figure 8.1 illustrates how this standard error varied with the number of positive example cases available per fold. These results suggest that the standard error is negatively correlated with the number of positive cases per fold. The highest values for the standard error of the true positive rate occurred for the tasks "Exclude stage 9", "No peak", and "Flat lactation" with, on average, only 1.6, 3.2, and 2.2 positive cases available per fold. The lowest standard errors occurred for the task "Slope lactation", with 20.4 positive cases per fold. These results suggest that especially with less than five positive cases per fold, the standard error is likely to be very high.

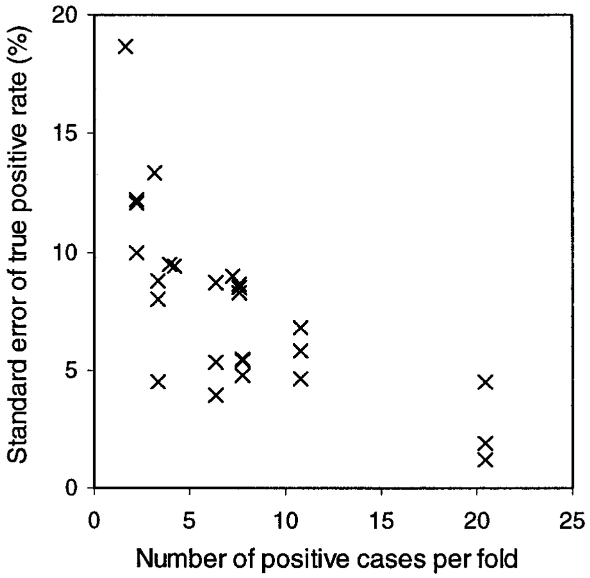


Figure 8.1 Variation of the cross-validation estimates of the standard error of the true positive rate with the number of positive examples available per cross-validation fold.

For the filtering of lactations of individual cows, with only 34 positive cases available out of 1428 lactations, ten-fold cross-validation was used (on average, 3.4 positive cases per fold). For this task, five-fold cross-validation might have been a better choice, likely reducing the fairly high standard errors of the true positive rate (8.8%, 8.0%, and 4.5% for the low, medium, and high filtering intensity levels, respectively). For the interpretation tasks, five-fold cross-validation was used, with, for some tasks, only two or three positive cases per fold. However, reducing the number of folds to less than five was considered inappropriate. With the number of folds below five, the individual folds have less than 80% in common with the entire data set, resulting in performance estimates that may not be very representative of the performance of the classifier generated from the entire data set.

This study did not include machine-learning experiments for the diagnosis module of the decision-support system described in Chapter 4. Analysis and classification of the 33 herds by the domain specialist would result in 33 example cases for each of the identified classification tasks of this module. Such a low number of cases is clearly not enough for reliable performance estimates using k-fold cross-validation. For these tasks, alternative approaches to training and testing could be used, such as bootstrapping and leave-one-out (Weiss and Kulikowski, 1991). However, with these approaches, only a single relative operating characteristic curve can be calculated, which does not allow for a direct estimate of the variability associated with such a curve or use of analysis of variance with machine-learning experiments. A more basic problem with very small data sets is that machine-learning techniques may not be able to generate a plausible knowledge description from the few example cases available, except for very simple classification tasks.

8.2 Achieved classification performance

For the main classification tasks of the interpretation of the group-average lactation curves (excluding the tasks supporting the interpretation of the slope after the peak), much better performance was achieved than for the tasks to filter lactations and tests of individual cows, even though a much smaller number of example cases was available. For the interpretation tasks, the size of the final decision trees (on average 4.3 leaf nodes) was smaller than the size of trees induced for the filtering tasks (on average 8.0 leaf nodes). These results suggest that the interpretation tasks were not as complex as the filtering tasks. The interpretation tasks were the result of detailed decomposition, whereas with the filtering tasks, many different patterns had to be learned from the data, i.e., many different reasons existed for considering a lactation or test to be an outlier.

Further investigation is required to determine if the classifiers generated for the filtering tasks have adequate classification performance with new data. Removing outliers is only of indirect importance to the analysis of group-average lactation curves. The correct removal of relatively minor outliers may not have a significant effect on the interpretation of the group-average lactation curves. Thus, removal of only the most obvious outliers through the implementation of classifiers with a low tendency of indicating a case as outlier, may be sufficient for this application. Improving the classification performance, if deemed necessary, may be achieved through re-evaluation of misclassified example cases by the domain specialist or classification of additional example cases by the specialist, followed by another iteration of machine learning and performance analysis.

8.3 Deployment of the knowledge-based system

This study explored the use of machine learning to assist the development of a KBS in dairy farming. First, a prototype decision-support system for the analysis of group-average lactation curves was developed. Knowledge-based components to remove outliers and interpret lactation curves were then created through machine learning and implemented in the decision-support system. Further research will be required to develop a final KBS that can be deployed to potential end-users.

To improve the adoption of a final KBS, end-users should be involved at several critical stages throughout the development process (Parker, 1999). Further development of a KBS for the analysis of group-average lactation curves should thus involve input from potential end-users of the system, such as dairy advisors employed by the Québec dairy herd improvement agency. The prototype system could be tested by a selected group of advisors with data from dairy herds they are familiar with. This could be followed by a consultation session with these advisors to elicit the perceived usefulness and limitations of the different parts of the system and to what extent they trust the automated filtering and interpretation of the lactation curves.

In this study, a decision-support module was developed for herd-level diagnosis of detected abnormalities with the group-average lactation curves, but knowledge-based modules for the identified classification tasks of this module were not created. However, the current prototype KBS with automated removal of outliers and interpretation of the lactation curves is expected to be already very useful to support the analysis of group-average lactation curves. Further investigation will be required to determine the usefulness of software components to automatically perform the diagnosis tasks.

Once a final KBS is implemented, the next step in the life cycle of the system involves maintenance to keep the knowledge captured in the system up-to-date. Retaining the case-acquisition functionality in the final KBS would allow end-users to store interesting new cases that the system mistakenly classified. These example cases could periodically be reviewed by a domain specialist and added to the case base. The additional example cases could then be used for the learning of new, improved, decision trees. However, this would require adoption of the machine-learning approach to KBS development by the organization that maintains the system.

Further development of a KBS for group-average lactation curve analysis might focus on the incorporation of machine-learning capabilities into the software deployed to end-users, leading to a self-learning or adaptive KBS (Schmoldt, 1997). Such a system would need to acquire automatically classified cases that the user agrees with and cases whose classification was manually adjusted. Automated learning from these newly acquired cases would allow these systems to adjust their knowledge to the specific type of data they have to deal with and to the preferences of the end-user.

8.4 Advantages and limitations of machine-learning assisted knowledge acquisition

In this research, machine-learning proved to be a feasible approach to support the development of a KBS in dairy farming. However, several limitations were encountered. First of all, the machine-learning approach can only work if a substantial number of example cases is available for learning and performance estimation. The minimum number of examples required for learning depends to a large extent on the complexity of classification task: many cases will be required for tasks with many different patterns for each class in order to cover each pattern sufficiently. The class distribution also affects the number of required examples, with the number of cases available for the minority class being most critical.

Secondly, using machine learning for knowledge acquisition was found to be time-consuming. Since previously classified example cases were not available, case-acquisition functionality had to be added to a decision-support system to facilitate the analysis and classification of a substantial number of cases by a domain specialist. In addition, considerable effort was required for 1) the preprocessing of acquired example cases (to assign cases to training and testing sets and to treat missing attribute values); 2) the construction of new attributes with potentially predictive value; 3) the execution of machine-learning experiments (to tune algorithm parameters and to determine the expected performance with new data); and 4) the evaluation of the learned knowledge. However, the procedure and software that were established can be reused to learn from data sets re-evaluated by the specialist or from larger data sets, and can be adjusted to develop KBS for related problem areas.

Finally, although machine learning automated part of the knowledge acquisition process, a large amount of interaction between system developer and domain specialists remained necessary. Input from the specialists was required for problem decomposition, development of a case-acquisition and decision-support system, to analyze and classify example cases, to create potentially predictive attributes, and to evaluate the plausibility of the learned knowledge. Thus, it was found to be of critical importance for the system developer to have sufficient knowledge of the field of application to facilitate the required communication with the domain specialist.

Direct comparisons between the machine-learning and the interview-based approach to KBS development have suggested that the machine-learning approach requires less effort and leads to a more accurate representation of the knowledge of the domain specialist (Ben-David and Mandel, 1995; Michalski and Chilausky, 1980). Such a direct comparison was not part of the scope of this project. However, this research suggested two additional advantages of machine-learning assisted over interview-based KBS development: 1) the ability to generate a series of classifiers with increasing probability of classifying a problem situation as abnormal from a single labeled data set; and 2) the ability to deal with problem situations where the domain specialist is expected to have great difficulty providing decision rules.

With the machine-learning approach to KBS development, a series of classifiers can be generated from a single data set classified by a specialist, with increasing probability of classifying a case as abnormal. These classifiers, each with a different trade-off between sensitivity and specificity, can be generated through the use of different misclassification costs during learning, and provide a variable decision boundary between the classes of the classification task. Implementation of such a series of classifiers provides end-users with the ability to control the ratio between false positives and false negatives depending of the specific data they are dealing with and their preferences (Kubat et al., 1998). For example, for a herd with many outliers or many abnormal aspects of the group-average lactation curve, use of a classifier with medium tendency of classifying new cases as abnormal may remove too many tests and lactations of individual cows from further analysis and indicate an overwhelming amount of abnormalities with the group-average lactation curves. In that case the user may want to focus on the most important deviations from normal performance and use classifiers with a low tendency of classifying new cases as abnormal (i.e., a low sensitivity and high specificity). Conversely, for herds with apparently few problems, the user of the KBS may choose to use classifiers with a high tendency of predicting new cases as abnormal to find less obvious deviations from normal performance, albeit with a high probability of false positives.

Problem areas where the domain specialist may have great difficulty expressing decision rules in interviews include the analysis of graphical performance representations and the use of novel approaches to data analysis. Both aspects were involved in the analysis of group-average lactation curves in this study. Graphical representations often help human beings to interpret the available data, but are difficult to translate into rules with numeric performance indices for use in a computer system. In addition, when dealing with a novel approach to analyzing data that includes new performance indices, the domain specialist may have the theoretical background to support proper interpretation, but does not have the practical experience of analyzing real-world example cases with that new approach. In such situations, the domain knowledge is poorly formalized and new knowledge must be created to solve the problem (Weiss and Kulikowski, 1991). For these problem areas, the development of a case-acquisition and decision-support system followed by machine learning of identified classification tasks is expected to be more useful than the traditional interview-based approach to knowledge acquisition. However, the domain specialist may have to analyze the data multiple times to adjust classifications made earlier based on the experience gained after analyzing all the available example data. Machine learning can help in this process by identifying cases that are often misclassified with different classifiers and which are, therefore, potentially mislabeled (Brodley and Friedl, 1996). Those cases could then be submitted to the domain specialist for re-evaluation, followed by machine learning using the improved data.

8.5 Machine-learning assisted knowledge acquisition in dairy farming

The analysis of group-average lactation curves can be considered as part of a tactical level decision-making activity within the spheres of dairy nutrition and milk production. Other areas and levels of decision making in dairy farming may also be suitable for machine-learning assisted KBS development. However, although KBS were expected to have great potential to support all levels of decision making in dairy farming (Doluschitz, 1990; Spahr et al., 1988), adoption of decision-support systems for strategic planning purposes has been slow and the usefulness of these often complex systems has been questioned (Kuhlmann and Brodersen, 2001). In addition, at low levels of decision making, specifically, from the regulatory to the operational level, decision making tends to be fairly simple and it may be easy for a domain specialist to provide decision rules. Thus, machine-learning assisted knowledge acquisition may be most useful for KBS that support decision making between the operational and tactical levels.

The analysis of milk urea nitrogen and related data to support nutrition management at the operational and tactical levels seems an attractive domain for the application of machine-learning assisted knowledge acquisition. The relationship between dairy nutrition and milk urea nitrogen is currently an active area of research. Machine learning may be a useful approach to formalize the existing practical knowledge in this domain, based on example cases analyzed and classified by dairy nutrition specialists with experience in dealing with this type of data. Another area of application may be the analysis of daily milk production records collected with milk yield sensors to support operational management. Example cases labeled by a domain specialist could be used to develop a KBS to indicate potential problems in the health, nutrition, and environmental spheres, based on the observed fluctuations in milk yield and additional data. Finally, machine-learning assisted knowledge acquisition could be applied to areas for which the traditional approach to KBS development has been used, such as the analysis of somatic cell count data (Allore et al., 1995), the allocation of dairy cows to feeding groups (Grinspan et al., 1994), and the ranking of cows for culling purposes (Strasser, 1997).

8.6 Further research in machine-learning assisted knowledge acquision

In this study, machine learning proved to be a feasible approach to KBS development in dairy farming. Interesting areas for further research related to machine-learning assisted knowledge acquisition include learning with multiple specialists and application of recently developed approaches to machine learning.

The knowledge-based components developed in this research were based on the expertise of a single domain specialist. However, for some applications it may be important to include the expertise of multiple specialists. The machine-learning assisted approach to KBS development may offer new ways to deal with this multi-specialist situation. One option would involve having each specialist classify the same set of example cases, followed by re-evaluation of cases for which different classes were assigned by the specialists, until consensus is reached. This approach would lead to high quality data for machine learning, but could be time consuming for the specialists. A second approach would rely on the ability of machine-learning algorithms to deal with the inconsistencies in the data. The classified example cases from each specialist could be merged into a large data set, with many duplicate and some inconsistent cases. With a third approach, a separate classifier would be generated with machine learning for each specialist involved. The predictions from the classifiers could then be processed into a single classification through, for example, a voting scheme with user-adjustable weights.

Additional research might also focus on the machine-learning algorithms used for knowledge acquisition. Further improvement of the selective naïve-Bayes algorithm implemented for this research should focus on improving the attribute selection process and developing techniques to help visualize the learned knowledge. Recent developments in machine learning that may be useful in the context of knowledge acquisition include the automatic learning of the structure and parameters of Bayesian networks (Heckerman, 1996) and the induction of fuzzy decision trees (Janikow, 1998). Both approaches deal with uncertainty in decision making and generate a graphical representation to express relations among the attributes used for classification. These approaches may improve the classification performance and also the understandability of the learned knowledge compared to naïve-Bayes classification and crisp decision-tree induction, but require further investigation regarding their applicability to real-world problems.

9 Summary and Conclusions

Knowledge-based systems (KBS) may help dairy producers and their advisors in dealing with the increasing amounts of available data and increasing complexity of decision making. However, the development of these systems has proven to be non-trivial. Thus, the main goal of this research was to explore the use of machine learning to support the development of KBS in dairy farming.

First, a framework was developed to support the creation of computerized information systems for use in dairy farming. This framework described the virtual part of dairy farming in terms of a management and control system consisting of a network of management and control activities, which process and exchange information. These activities were classified according to the level of decision making at which they take place, the sphere of dairy farming they belong to, and how and where they are performed. Decision-support systems may help dairy producers and their advisors with the proper interpretation of the large amounts of information available. Implementation of knowledge-based components into such systems would allow for the partial or complete automation of time-consuming and complex analysis tasks. The framework was used to identify promising areas for the creation of KBS, and analysis of group-average lactation curves was chosen as the application area for machine-learning assisted KBS development. This type of analysis involves evaluation of both graphical and numerical information, and can be seen as part of a management and control activity within the sphere of nutrition and at the tactical level.

In order to support the overall process of machine-learning assisted KBS development, a process model was developed. This model involved the following eight steps: problem analysis and formulation, case-acquisition tool development, classification of example cases by a domain specialist using the case-acquisition tool, analysis and preprocessing of the acquired example cases, machine-learning algorithm selection and configuration, training and testing, analysis of the results of machine learning, and deployment of the KBS. Several iterations of the first three steps of the process model were required to develop a case-acquisition and decision-support system (CADSS) in consultation with two dairy nutrition specialists. The overall problem was decomposed into three sub-problems: removal of outlier tests and lactation curves of individual cows, interpretation of group-average lactation curves, and diagnosis of detected abnormalities at the herd level through the identification of potential management deficiencies. For each sub-problem, a software module was developed allowing the user to analyze both graphical and numerical performance representations. In addition, classification tasks and their classes were identified. The example-based approach to CADSS development proved to be very useful, facilitating the communication between system developer and domain specialists, and allowing the specialists to explore the appropriateness of the various prototypes developed. The resulting software represented a formalization of the approach to group-average lactation-curve analysis, elicited from the two domain specialists. In addition, the CADSS enabled domain specialists to analyze and classify example cases of the analysis of group-average lactation curves in an efficient manner.

A dairy nutrition specialist used the CADSS to analyze and classify the milk-recording data from 33 Holstein dairy herds enrolled with the Québec dairy herd analysis service. This resulted in 1428 lactations and 7684 tests of individual cows, classified by the specialist as either outlier or non-outlier, and 99 interpretations of group-average lactation curves. Ten-fold cross-validation was used to estimate the performance of classifiers induced for the classification tasks to filter lactations and test-day data of individual cows. For the interpretation tasks, the number of folds were reduced to five due to the limited availability of data. Indices of classification performance, used in this research, included the true positive rate, false positive rate, predictive value positive, and positive prediction rate. Relative operating characteristic (ROC) curves were used to visualize the trade-off between correctly classifying positive cases and correctly classifying negative cases. Classification tasks with three or more classes were reformulated, during performance analysis, into “Normal” versus “Abnormal” classification or decomposed into a series of sub-tasks.

Analysis of variance was used to support the analysis of the results of machine-learning experiments in which the effects of different approaches to data preprocessing, attribute availability, machine-learning algorithm, and configuration of algorithm parameters were investigated. To enable the use of analysis of variance based on the information represented by ROC curves, a single performance index, called TP*, was developed. It was defined as the mean true positive rate of the ROC curve for the range of false positive rate values of interest. This index makes use of domain expertise to limit the performance analysis to false positive rate values that are considered reasonable.

Considerable effort was required to derive potentially predictive attributes for machine learning from the initial set of attributes available to the domain specialist. Experiments with the filtering of tests within lactations showed that adding derived attributes to the initial data resulted in a significant improvement of the performance.

Machine-learning experiments showed that tuning of the parameter configuration of the CART decision-tree induction algorithm greatly improved the classification performance compared to the default configuration. For the filtering of tests within lactations, the selective naïve-Bayes classifier performed significantly better than CART. Both machine-learning approaches allowed for inspection of the plausibility of individual pieces of learned knowledge. Decision nodes and attributes that were expected not to properly classify new data were removed or modified. However, the domain specialist considered the decision trees as more transparent than the knowledge generated with naïve Bayes.

For most classification tasks, it was possible to generate a series of three classifiers from the entire data set, representing a low, medium, and high tendency of classifying new cases as abnormal. Implementation of these alternative classifiers for each classification task in the final KBS for group-average lactation-curve analysis allows end-users to choose the classifier with the desired tendency of classifying new cases as abnormal.

The decision trees for the main interpretation tasks showed good classification performance in the cross-validation experiments. For the filtering of lactations and tests of individual cows the performance was fairly poor: a high true positive rate could only be achieved at the expense of many false positives. Improvement of the classification performance may be achieved through re-evaluation of misclassified example cases by the domain specialist, to reduce the number of mislabeled cases, followed by machine learning with the improved data. The acquisition and use of additional example cases for learning may also improve the performance.

The decision-tree induction approach resulted in relatively small decision trees that were easy to understand by the domain specialist. In several cases, counter-intuitive decision nodes, which tended to occur at the end of the decision trees with limited data in the parent nodes, were manually removed or replaced with a substitute suggested by the CART algorithm. Relying on the expertise of the domain specialist, these adjustments were expected to lead to improved performance with new data. For classification tasks with very poor performance in cross-validation experiments, a series of potentially useful decision trees were induced, and the most plausible one was chosen by the domain specialist. Thus, evaluation of the classifiers by the domain specialist was considered an important step in the overall process of machine-learning assisted knowledge acquisition.

The induced decision trees were implemented as knowledge-based components in the CADSS program to perform the removal of outliers and the interpretation of group-average lactation curves automatically. Before it can be deployed, the program needs to be further developed and tested in the field with, for example, a small group of dairy advisors. This end-user input should contribute significantly to its advancement.

In this research, machine-learning assisted knowledge acquisition proved to be a feasible approach to support the development of a KBS in dairy farming. However, several limitations were encountered: a substantial number of labeled example cases had to be acquired from a domain specialist, the overall process proved to be quite time-consuming, and a large amount of interaction between system developer and domain specialists was necessary. In dairy farming and agriculture in general, machine-learning assisted KBS development is expected to be especially useful for problem domains in which the specialist may have great difficulty expressing decision rules, such as tactical-level decision making and the interpretation of graphical information.

References

Adriaans, P. W. 1997. Industrial requirements for ML application technology. Pages 7-11 *in* Proc. workshop Machine learning applications in the real world: methodological aspects and implications. Int. Conf. Machine Learning, Nashville, TN, July 12, 1997. Available: http://www.aifb.uni-karlsruhe.de/WBS/ICML97/proceedings.html.

Aha, D. W. 1992. Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms. Int. J. of Man-Machine Studies 36:267-287.

Allen, B. P. 1994. Case-based reasoning: business applications. Commun. ACM 37(3):40-42.

Allore, H. G., L. R. Jones, W. G. Merrill, and P. A. Oltenacu. 1995. A decision support system for evaluating mastitis information. J. Dairy Sci. 78:1382-1398.

Bailey, T. L., W. D. Whittier, J. Murphy, and J. F. Currin. 1998. Using records to evaluate milk production. Veterinary Medicine 1998(December):1083-1093.

Ben-David, A. and J. Mandel. 1995. Classification accuracy: machine learning vs. explicit knowledge acquisition. Machine Learning 18:109-114.

Boehlje, M. D., and V. R. Eidman. 1984. Farm Management. Wiley, New York.

Booch, G. 1994. Object-oriented Analysis and Design with Applications. 2nd ed. Benjamin-Cummings, Redwood City, CA.

Boulesteix, I, B. Balvay, R. Champy, and E. Rehben. 1996. Possible consequences for French dairy farming of the ISO standards for electronic data interchange. Pages 155-158 *in* Performance Recording of Animals. Proc. 30th Bienn. Session Int. Comm. Anim. Recording. Eur. Assoc. Anim. Prod. Publ. No. 87. Wageningen Pers, Wageningen, Netherlands.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30:1145-1159.

Brand, A., H. Folkerts, W.J.A. Hanekamp, W. D. de Hoop, and G.M.A. Verheijen. 1995. Information model for dairy farms. Agric. Telematics Centre, Wageningen, Netherlands.

Bratko, I., B. Cestnik, and I. Kononenko. 1996. Attribute-based learning. Artif. Intell. Commun. 9(1):27-32.

Breiman, L., J. H. Friedman, R. A. Olshen, C. J. Stone. 1984. Classification and regression trees. Wadsworth, Pacific Grove, CA.

Briscoe, G. and T. Caelli. 1996. A compendium of machine learning volume 1: symbolic machine learning. Ablex Publ. Corp., Norwood, NJ.

Brodley, C. E. and M. A. Friedl. 1996. Identifying and eliminating mislabeled training instances. Pages 799-805 *in* Proc. 13th Nat. Conf. Artificial Intelligence, Am. Assoc. for Artificial Intelligence, Menlo Park, CA.

Brodley, C. E. and P. Smyth. 1997. Applying classification algorithms in practice. Statistics and Computing 7:45-56.

Carbonell, J. G. 1989. Introduction: paradigms for machine learning. Artif. Intell. 40:1-9.

Cohen, P. R. 1995. Empirical methods for artificial intelligence. MIT Press, Cambridge, MA.

Crosse, S. 1991. Development and implementation of a computerised management information system (DAIRYMIS II) for Irish dairy farmers. Comput. Electron. Agric. 6:157-173.

Davis, G. B., and M. H. Olson. 1985. Management Information Systems: Conceptual Foundations, Structure, and Development. 2nd ed. McGraw-Hill, New York.

De Hoop, D. W. 1988. Management processes in dairy and pig farming and the construction of systems. Pages 77-86 *in* S. Korver and J.A.M. Arendonk, Eds., Modelling livestock production systems. Proc. Seminar Eur. Commun. Progr. Coordination Agric. Res., Brussels, Belgium. Kluwer Acad. Publ., Dordrecht, Netherlands.

DeLorenzo, M. A., T. H. Spreen, G. R. Bryan, D. K. Beede, and J.A.M. van Arendonk. 1992. Optimizing model: insemination, replacement, seasonal production, and cash flow. J. Dairy Sci. 75:885-896.

Devir, S., J. A. Renkema, R.B.M. Huirne, and A. H. Ipema. 1993. A new dairy control and management system in the automatic milking farm: basic concepts and components. J. Dairy Sci. 76:3607-3616.

Dhar, V. and R. Stein. 1997. Intelligent decision support methods; the science of knowledge work. Prentice Hall, Upper Saddle River, NJ.

Dietterich, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10:1895-1923.

Doluschitz, R. 1990. Expert systems for management in dairy operations. Comput. Electron. Agric. 5:17-30.

Domecq, J. J., R. L. Nebel, M. L. McGilliard, and A. T. Pasquino. 1991. Expert system for evaluation of reproductive performance and management. J. Dairy Sci. 74:3446-3453.

Duda, R. O. and P. E. Hart. 1973. Pattern classification and scene analysis. Wiley, New York.

Durkin, J. 1994. Expert systems: design and development. Macmillan, New York.

Enevoldsen, C., J. T. Sørensen, I. Thysen, C. Guard, and Y. T. Gröhn. 1995. A diagnostic and prognostic tool for epidemiologic and economic analyses of dairy herd health management. J. Dairy Sci. 78:947-961.

Esslemont, R. J. and M. E. Williams. 1992. Assessment of the use of a decision support system to manage insemination through routine milk progesterone analysis. Pages 472-477 *in* A. H. Ipema, A. C. Lippus, J. H. M. Metz, and W. Rossing, Eds., Proc. Int. Symp. Prospects automatic milking, Wageningen, Netherlands, November 23-25, 1992. EAAP publication 65, Pudoc Scientific Publ., Wageningen, Netherlands.

Fayyad, U. M. 1996. Data mining and knowledge discovery: making sense out of data. IEEE Expert 11(5):20-25.

Feigenbaum, E. A. 1979. Themes and case studies of knowledge engineering. Pages 3-25 *in* D. Michie, Ed., Expert systems in the micro-electronic age. Edinburgh University Press, Edinburgh, UK.

Feng, C. and D. Michie. 1994. Machine learning of rules and trees. Pages 50-83 *in* D. Michie, D. J. Spiegelhalter, and C. C. Taylor, Eds., Machine learning, neural and statistical classification. Ellis Horwood, Hemel Hempstead, UK.

Fourdraine, R. H., M. A. Tomaszewski, and T. J. Cannon. 1992a. Dairy herd lactation expert system, a program to analyze and evaluate lactation curves. Pages 331-337 *in* A. H. Ipema, A. C. Lippus, J. H. M. Metz, and W. Rossing, Eds., Proc. Int. Symp. Prospects automatic milking, Wageningen, Netherlands, November 23-25, 1992. EAAP publication 65, Pudoc Scientific Publ., Wageningen, Netherlands.

Fourdraine, R. H., M. A. Tomaszewski, and T. J. Cannon. 1992b. A computer program to analyze herd management using DHI test day information. Pages 478-482 *in* A. H. Ipema, A. C. Lippus, J. H. M. Metz, and W. Rossing, Eds., Proc. Int. Symp. Prospects automatic milking, Wageningen, Netherlands, November 23-25, 1992. EAAP publication 65, Pudoc Scientific Publ., Wageningen, Netherlands.

Frost, A. R., C. P. Schofield, S. A. Beaulah, T. T. Mottram, J. A. Lines, and C. M. Wathes. 1997. A review of livestock monitoring and the need for integrated systems. Comput. Electron. Agric. 17:139-159.

Gauthier, L., and R. Kok. 1989. Integrated farm control software: I. Functional requirements and basis design criteria. Artif. Intell. Applic. Natural Resource Management 3(1):27-37.

Gillies, D. 1996. Artificial intelligence and scientific method. Oxford University Press, Oxford, UK.

Goldberg, D. E. 1994. Genetic and evolutionary algorithms come of age. Commun. ACM 37(3):113-119.

Gonzalez, A. J. and D. D. Dankel. 1993. The engineering of knowledge-based systems. Prentice Hall, Englewood Cliffs, NJ.

Grinspan, P., Y. Edan, H. E. Kahn, and E. Maltz. 1994. A fuzzy logic expert system for dairy cow transfer between feeding groups. Trans. ASAE 37:1647-1654.

Guan, B. T. and G. Gertner. 1991. Machine learning and its possible role in forest science. Artif. Intell. Applic. Natural Resource Management 5(2):27-36.

Heald, C. W., T. Kim, W. M. Sischo, J. B. Cooper, and D. R. Wolfgang. 2000. A computerized mastitis decision aid using farm-based records: an artificial neural network approach. J. Dairy Sci. 83:711-720.

Heald, C. W., T. O. Kim, J. B. Cooper, and M. A. Foster. 1995. A knowledge-based mastitis evaluation tool for dairy management advisors. Pages 269-274 *in* A. J. Udink ten Cate, R. Martin-Clouaire, A. A. Dijkhuizen and C. Lokhorst, Eds., Artificial Intelligence in Agriculture, Proc. 2nd IFAC/IFIP/EurAgEng Workshop, Wageningen, Netherlands, May 29-31, 1995. Pergamon, Oxford, UK.

Heckerman, D. 1996. A tutorial on learning with Bayesian networks. Microsoft research technical report MSR-TR-95-06. Microsoft, Redmond, WA.

Henery, R. J., 1994. Methods for comparison. Pages 107-124 *in* D. Michie, D. J. Spiegelhalter, and C. C. Taylor, Eds., Machine learning, neural and statistical classification. Ellis Horwood, Hemel Hempstead, UK.

Hogeveen, H., E. N. Noordhuizen-Stassen, J. F. Schreinemakers, and A. Brand. 1991. Development of an integrated knowledge-based system for management support on dairy farms. J. Dairy Sci. 74:4377-4384.

Hogeveen, H., M. A. Varner, D. S. Bree, D. E. Dill, E. N. Noordhuizen-Stassen, and A. Brand. 1994. Knowledge representation methods for dairy decision support systems. J. Dairy Sci. 77:3704-3715.

Howarth, M. S., J. R. Brandon, S. W. Searcy, and N. Kehtarnavaz. 1992. Estimation of tip shape for carrot classification by machine vision. J. Agric. Engng. Res. 53(2):123-139.

Huirne, R. B. M. 1990. Basic concepts of computerized support for farm management decisions. Eur. Rev. Agric. Econ. 17:69-84.

Janikow, C. Z. 1998. Fuzzy decision trees: issues and methods. IEEE Transactions on Systems, Man, and Cybernetics 28(1):1-14.

Jones, L. R. 1992. Monitoring cow performance using lactation curves. Pages 497-501 *in* A. H. Ipema, A. C. Lippus, J. H. M. Metz, and W. Rossing, Eds., Proc. Int. Symp. Prospects automatic milking, Wageningen, Netherlands, November 23-25, 1992. EAAP publication 65, Pudoc Scientific Publ., Wageningen, Netherlands.

Julien, B., S. J. Fenves, and M. J. Small. 1992. Knowledge acquisition methods for environmental evaluation. Artif. Intell. Applic. Natural Resource Management 6(1):1-20.

Kalter, R. J., A. L. Skidmore, and C. J. Sniffen. 1992. Distributed intelligence and control: The new approach to dairy farm management. Pages 171-176 *in* Proc. 4th Int. Conf. Computers in Agric. Ext. Programs, Orlando, FL.

Kay, R. D. 1986. Farm Management: Planning, Control, and Implementation. 2nd ed. McGraw-Hill, New York.

Kim, T. and C. W. Heald. 1999. Inducing inference rules for the classification of bovine mastitis. Comput. Electron. Agric. 23:27-42.

Klein, M. R., and L. B. Methlie. 1995. Knowledge-based decision support systems with applications in business. 2nd ed. Wiley, Chichester, UK.

Kohavi, R. and G. H. John. 1997. Wrappers for feature subset selection. Artificial Intelligence 97: 273-324.

Kok, R., and R. Lacroix. 1993. An analytical framework for the design of autonomous, enclosed agro-ecosystems. Agric. Systems 43:235-260.

Kolodner, J. L. 1993. Case-based reasoning. Morgan Kaufmann, San Mateo, CA.

Kononenko, I., I. Bratko, and M. Kukar. 1998. Application of machine learning to medical diagnosis. Pages 389-408 *in* R. S. Michalski, I. Bratko, and M. Kubat, Eds., Machine learning and data mining: methods and applications. Wiley, Chichester, UK.

Kroeze, G. H., C. Lokhorst, H. J. van de Beek, and F. de Vries. 1996. Computerized central farm analysis based on the farm comparison of uniform-management. Pages 610-616 *in* C. Lokhorst, A. J. Udink ten Cate, and A. A. Dijkhuizen, Eds., Information and communication technology applications in agriculture, Proc. 6th Int. Congr. Computer Technology Agric., Wageningen, Netherlands. VIAS, Wageningen, Netherlands.

Kubat, M., R. C. Holte, and S. Matwin. 1998. Machine learning for the detection of oil spills in satellite radar images. Machine Learning 30:195-215.

Kuhlmann, F. and C. Brodersen. 2001. Information technology and farm management: developments and perspectives. Comput. Electron. Agric. 30:71-83.

Lacroix, R., and K. M. Wade. 1996. Expected benefits from the use of the information superhighway in dairy herd improvement. Pages 175-178 *in* Performance Recording of Animals. Proc. 30th Bienn. Session Int. Comm. Anim. Recording. Eur. Assoc. Anim. Prod. Publ. No. 87. Wageningen Pers, Wageningen, Netherlands.

Lacroix, R., J. Huijbers, R. Tiemessen, D. Lefebvre, D. Marchand, and K. M. Wade. 1998. Fuzzy set-based analytical tools for dairy herd improvement. Appl. Eng. Agric. 14:79-85.

Lacroix, R., K. M. Wade, R. Kok, and J. F. Hayes. 1995. Prediction of cow performance with a connectionist model. Trans. ASAE 38:1573-1579.

Lacroix, R., M. Strasser, K. M. Wade, and A. Fournier. 1997. L'autoroute de l'information en agriculture… sur la bonne voie! Le producteur de lait québécoise. 18(1):39-42.

Langley, P. 1996. Elements of machine learning. Morgan Kaufmann, San Francisco.

Langley, P. and H. A. Simon. 1995. Applications of machine learning and rule induction. Commun. ACM 38(11):55-64.

Langley, P. and S. Sage. 1994. Induction of selective Bayesian classifiers. Pages 399-406. *in* R. Lopez de Mantaras and D. Poole, Eds., Proc. 10th Conf. on Uncertainty in Aritificial Intelligence. Morgan Kaufmann, Seattle, WA.

Lefebvre, D., D. Marchand, M. Léonard, C. Thibault, E. Block, and T. Cannon. 1995. Gestion de la performance du troupeau laitier: des outils à exploiter. Pages 13-39 *in* Choix d'aujourdhui, lait de demain, 19e Symp. bovins laitiers, St. Hyacinthe, Quebec, Canada, October 26, 1995. Conseil des Productions Animales du Québec.

Leigh, J. R. 1992. Control Theory, A Guided Tour. Inst. Electr. Engineers Control Eng. Ser. No. 45. Peter Peregrinus, London, UK.

Lévesque, P., G. Proulx, P. Guillemette, S. Doré, R. Pellerin, and M. Rousseau. 1994. Distribution automatique des fourrages. Pages 159-180 *in* Savoir profiter de ses atouts, 18e Symp. bovins laitier, St. Hyacinthe, Quebec, Canada, October 27, 1994. Conseil des Productions Animales du Québec.

Littell, R. C., G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. 1996. SAS system for mixed models. SAS Institute Inc., Cary, NC.

Lokhorst, C., G. H. Kroeze, J. V. van den Berg, B. Blok, and F. de Vries. 1999. Data mining for knowledge discovery in Dutch dairy databases. Pages 533-540 *in* G. Schiefer, R. Helbig, U. Rickert, Eds., Proc. 2nd EFITA conf., Bonn, Germany, September 27-30, 1999. Univ. Bonn – Inst. Landwirtschaftliche Betriebslehre, Bonn, Germany.

Mainland, D. D. 1994. A decision support system for dairy farmers and advisors. Agric. Syst. 45:217-231.

Mangina, E. E., Q. Shen, and C. P. Yialouris. 1999. A knowledge-based system for low technology tomato-greenhouse management. Pages 99-108 *in* G. Schiefer, R. Helbig, U. Rickert, Eds., Proc. 2nd EFITA conf., Bonn, Germany, September 27-30, 1999. Univ. Bonn – Inst. Landwirtschaftliche Betriebslehre, Bonn, Germany.

Marchand, R. 1995. Méthodes inductives d'intelligence artificielle comme aide à la décision en insémination artificielle. Mémoire de la Maîtrise en informatique de gestion, Université du Québec à Montréal.

McQueen, R. J., S. R. Garner, C. G. Nevill-Manning, and I. H. Witten. 1995. Applying machine learning to agricultural data. Comput. Electron. Agric. 12:275-293.

Michalski, R. S. and R. L. Chilausky. 1980. Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. Int. J. of Policy Analysis and Information Systems 4(2):125-161.

Michie, D., D. J. Spiegelhalter, and C. C. Taylor. 1994. Machine learning, neural and statistical classification. Ellis Horwood, Hemel Hempstead, UK.

Microsoft. 1997. Visual Basic Guide to Data Access Objects. Microsoft Corporation, Redmond, WA.

Mitchell, R. S., R. A. Sherlock, and L. A. Smith. 1996. An investigation into the use of machine learning for determining oestrus in cows. Comput. Electron. Agric. 15:195-213.

Mitchell, T. M. 1997. Machine learning. McGraw-Hill, New York.

National Research Council, 1989. Pages 89-115 *in* Nutritional Requirements of Dairy Cattle. 6th rev. ed. Natl. Acad. Sci. Washington, DC.

Nielen, M., Y. H. Schukken, A. Brand, S. Haring, and R. T. Ferwerda-Van Zonneveld. 1995. Comparison of analysis techniques for on-line detection of clinical mastitis. J. Dairy Sci. 78:1050-1061.

Parker, C. 1999. A user centered design method for agricultural DSS. Pages 395-403 *in* G. Schiefer, R. Helbig, U. Rickert, Eds., Proc. 2nd EFITA conf., Bonn, Germany, September 27-30, 1999. Univ. Bonn – Inst. Landwirtschaftliche Betriebslehre, Bonn, Germany.

Pazzani, M. J. 2000., Knowledge discovery from data ? IEEE Intelligent systems 15(2), 10-13.

Pellerin D., R. Levallois, G. St-Laurent, and J.-P. Perrier. 1994. LAIT-XPERT VACHES: An expert system for dairy herd management. J. Dairy Sci. 77:2308-2317.

Pietersma, D., R. Lacroix, and K. M. Wade. 1998. A framework for the development of computerized management and control systems for use in dairy farming. J. Dairy Sci. 81:2962-2972.

Pietersma, D., R. Lacroix, D. Lefebvre, E. Block, and K. M. Wade. 2001a. A case-acquisition and decision-support system for the analysis of group-average lactation curves. J. Dairy Sci. 84:730–739.

Pietersma, D., R. Lacroix, D. Lefebvre, and K. M. Wade. 2001b. Performance analysis of decision trees induced through machine learning for lactation-curve analysis. Submitted to Comput. Electron. Agric.

Plant, R. E. and N. D. Stone. 1991. Knowledge-based systems in agriculture. McGraw-Hill, New York.

Power, J. M. 1993. Object-oriented design of decision support systems in natural resource management. Comput. Electron. Agric. 8:301-324.

Programme d'analyse des troupeaux laitiers du Québec. 2000. Rapport de production 1999. Programme d’analyse des troupeaux laitiers du Québec, Ste. Anne de Bellevue, QC, Canada.

Provost, F. and R. Kohavi, Eds., 1998. Machine learning: special issue on applied research in machine learning. Machine Learning 30(2/3).

Provost, F., T. Fawcett, and R. Kohavi. 1998. The case against accuracy estimation for comparing induction algorithms. Pages 445-453 *in* J. Shavlik, Ed., Proc. 15th Int. Conf. Machine Learning, July 24-27, Madison, WI. Morgan Kaufmann, San Francisco.

Rossing, W., and P. H. Hogewerf. 1997. State of the art of automatic milking systems. Comput. Electron. Agric. 17:1-17.

Rumelhart, D. E., B. Widrow, and M. A. Lehr. 1994. The basic ideas in neural networks. Commun. ACM 37(3):87-92.

Salehi, F., R. Lacroix, and K. M. Wade. 2000. Development of neuro-fuzzifiers for qualitative analyses of milk yield. Comput. Electron. Agric. 28:171-186.

Schmisseur, E., and M. J. Gamroth. 1993. DXMAS: An expert system program providing management advice to dairy operators. J. Dairy Sci. 76:2039-2049.

Schmoldt, D. L. 1997. Adding learning to knowledge-based systems: taking the "artificial" out of AI. Artif. Intell. Applic. Natural Resource Management 11(3):1-7.

Simon, H. A. 1960. The New Science of Management Decision. Harper and Row, New York.

Skidmore, A. L., A. Brand, and C. J. Sniffen. 1996. Monitoring milk production: decision making and follow-up. Pages 263-281 *in* A. Brand, J.P.T.M. Noordhuizen, Y. H. Schukken, Eds., Herd health and production management in dairy practice. Wageningen Pers, Wageningen, Netherlands.

Spahr, S. L. 1993. New technologies and decision making in high producing herds. J. Dairy Sci. 76:3269-3277.

Spahr, S. L., and H. B. Puckett. 1986. Electronics in livestock production. Illinois Res. 28(1):22-25.

Spahr, S. L., D. E. Dill, J. B. Leverich, G. C. McCoy, and R. Sagi. 1993. Dairybase: an electronic individual animal inventory and herd management system. J. Dairy Sci. 76:1914-1927.

Spahr, S. L., L. R. Jones, and D. E. Dill. 1988. Expert systems - their use in dairy herd management. J. Dairy Sci. 71:879-885.

Steel, R.G.D. and J. H. Torrie. 1980. Principles and procedures of statistics: a biometrical approach. 2nd ed. McGraw-Hill, New York.

Steinberg, D. and P. Colla. 1997. CART -- Classification and Regression Trees. Salford Systems. San Diego, CA.

Steinmetz, V., M. J. Delwiche, D. K. Giles, and R. Evans. 1994. Sorting cut roses with machine vision. Trans. ASAE 37:1347-1353.

Strasser, M. 1997. The development of a fuzzy decision-support system for dairy cattle culling decisions. M.Sc. Thesis. McGill University, Montreal, Canada.

Strasser, M., R. Lacroix and K. M. Wade. 1998. Applying ActiveX technologies to create Internet-based decision-support systems for dairy cattle management. J. Dairy Sci. 81(Suppl. 1):266. (Abstr.).

Svennersten-Sjaunja, K., L. -O. Sjaunja, J. Bertilsson, and H. Wiktorsson. 1997. Use of regular milking records versus daily records for nutrition and other kinds of management. Livestock Prod. Sci. 48(3):167-174.

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. Science 240:1285-1293.

Tomaszewski, M. A. 1993. Record-keeping systems and control of data flow and information retrieval to manage large high producing herds. J. Dairy Sci. 76:3188-3194.

Tufte, E. R. 1997. Visual explanations: images and quantities, evidence and narrative. Graphics Press, Cheshire, CT.

Verdenius, F., A.J.M. Timmermans, and R. E. Schouten. 1997. Process models for neural network applications in agriculture. Artif. Intell. Applic. Natural Resource Management 11(3):31-45.

Wagner, P., and F. Kuhlmann. 1991. Concept and implementation of an integrated decision support system (IDSS) for capital-intensive farming. Agric. Econ. 5:287-310.

Weiss, S. M. and C. A. Kulikowski. 1991. Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems. Morgan Kaufmann, San Mateo, CA.

Whittaker, A. D., M. A. Tomaszewski, J. F. Taylor, R. Fourdraine, C. J. van Overveld, and R. G. Schepers. 1989. Dairy herd nutrition analysis using knowledge systems techniques. Agric. Systems 31:83-96.

Wilmink, J.B.M. 1987. Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. Livest. Prod. Sci. 16:335-348.

Witten, I. H. and E. Frank. 2000. Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco.

Yang, X. Z., R. Lacroix, and K. M. Wade. 1999. Neural detection of mastitis from dairy herd improvement records. Trans. ASAE 42:1063-1071.