Decision-tree induction to interpret lactation curves

D. Pietersma[1], R. Lacroix[1], D. Lefebvre[2] and K.M. Wade[1]

[1]Department of Animal Science, McGill University, 21111 Lakeshore Road, Ste. Anne de Bellevue, Quebec, Canada H9X 3V9; and [2]Department of R&D, Programme d'Analyse des Troupeaux Laitiers du Québec, 555 Boul. des Anciens Combattants, Ste. Anne de Bellevue, Quebec, Canada H9X 3R4.

Pietersma, D., Lacroix, R., Lefebvre, D. and Wade, K.M. 2002. **Decision-tree induction to interpret lactation curves**. Canadian Biosystems Engineering/Le génie des biosystèmes au Canada **44**: 7.1 - 7.13. Decision-tree induction was used to learn to interpret parity-group average lactation curves automatically in dairy farming. Three parity groups were involved consisting of cows in their first, second, or third and higher parity. A dairy-nutrition specialist analyzed 99 parity-group average lactation curves, representing 33 dairy herds, and classified these curves using 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 was induced from the entire data set with increasing sensitivity, and associated with a low, medium, and high tendency of classifying new cases as abnormal. For the other 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 performed well. 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 match the interpretation of parity-group average lactation curves closely 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. **Keywords**: machine learning, decision-tree induction, knowledge acquisition, lactation-curve analysis, dairy-cattle management.

L'induction d'arbres de décision a été utilisée pour l'apprentissage machine de l'interprétation de courbes de lactation, pour la production laitière. Trois groupes de vaches ont été formés sur la base de leur parité : première parité, deuxième parité, et trois parités et plus. Un spécialiste de l'alimentation bovine a analysé 99 courbes de lactation moyennes résultant d'une agrégation par groupe de parité, pour 33 troupeaux. Le spécialiste a classé ces courbes en fonction de critères d'interprétation prédéfinis. Sept principales tâches de classification et trois tâches secondaires, supportant les premières, ont été identifiées. Pour chaque tâche, des attributs potentiellement prédictifs ont été créés sur la base de l'information graphique et numérique qui était disponible pour le spécialiste. La performance des classifications était estimée par la validation croisée en quintuple, et les courbes caractéristiques d'opération relative (ROC) permettaient de visualiser le compromis atteint entre la spécificité et la sensibilité des classificateurs (i.e., arbres de décision). Pour cinq des sept tâches principales, l'ensemble des données a permis de générer une série de trois classificateurs avec une sensibilité croissante, et respectivement associés à une tendance basse, moyenne et élevée à classer un nouveau cas comme étant anormal. Pour deux des tâches principales, des classificateurs alternatifs performaient similairement. Les classificateurs avec une tendance moyenne ont été retenus, à cause de leur probabilité à classer comme anormaux de nouveaux cas, qui serait semblable à la prévalence de cas anormaux observée dans les données utilisées pour l'induction des classificateurs. Les arbres de décision induits pour les principales tâches de classification avaient de bonnes performances. Pour les classificateurs à tendance moyenne, la sensibilité était d'au moins 80% et la proportion de cas correctement prédits comme anormaux était d'au moins 75%. La performance des classificateurs pour les tâches secondaires n'était pas bonne, et l'expertise du spécialiste a été requise pour choisir un arbre de décision plausible parmi ceux qui ont été induits. Les arbres de décision, dont le nombre de noeuds de décision variait entre deux et sept, ont été évalués par le spécialiste et, après certains ajustements, considérés comme plausibles. Cette étude suggère que les arbres de décision induits automatiquement sont capables de reproduire correctement l'interprétation de courbes de lactation telle que pratiquée par un spécialiste du domaine. L'acquisition de connaissances basée sur l'apprentissage machine pourrait être appropriée pour des domaines d'expertise où les spécialistes ont de la difficulté à exprimer les règles de décision qu'ils emploient, comme lors de l'interprétation d'information graphique. **Mots-clés**: Apprentissage machine, induction d'arbre de décision, acquisition de connaissances, analyse de courbes de lactation, gestion des troupeaux laitiers.

Note: The following abreviations have been used. CADSS = case-acquisition and decision-support system, KBS = knowledge-based system, FP = false positive, POP = prevalence of positives, PPR = positive prediction rate, PVP = predictive value positive, ROC = relative operating characteristic, TP = true positive.

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. These 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 parity-group average lactation curves, derived from such data, has been identified as a useful tool in supporting 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 for cows grouped according to their parity (1, 2, and 3+), comparison with a standard lactation curve, and the analysis of additional explanatory data with the objective of detecting potential management deficiencies related to nutrition. Reducing feed costs can have an important impact on the profitability of dairy farms. For example, for an average herd in Quebec, with 45 cows producing 8000 kg milk per year and $0.14 feed costs per liter milk produced (Programme d'analyse des troupeaux laitiers du Québec 2001), each 1% decrease in feed costs represents a savings of $500. Such savings, if achieved by 50% of the 6900 herds enrolled with the provincial dairy herd analysis service, would represent more than $1.7 million per year. A knowledge-based system (KBS) for the analysis of parity-group average lactation curves was developed at the Texas A&M University (Fourdraine et al. 1992; Whittaker et al. 1989) to automate the preprocessing of the large amount of raw data involved and to provide dairy producers and their advisors with expert interpretation. A KBS for the analysis of parity-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 interpret the information described by the graph automatically. 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 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).

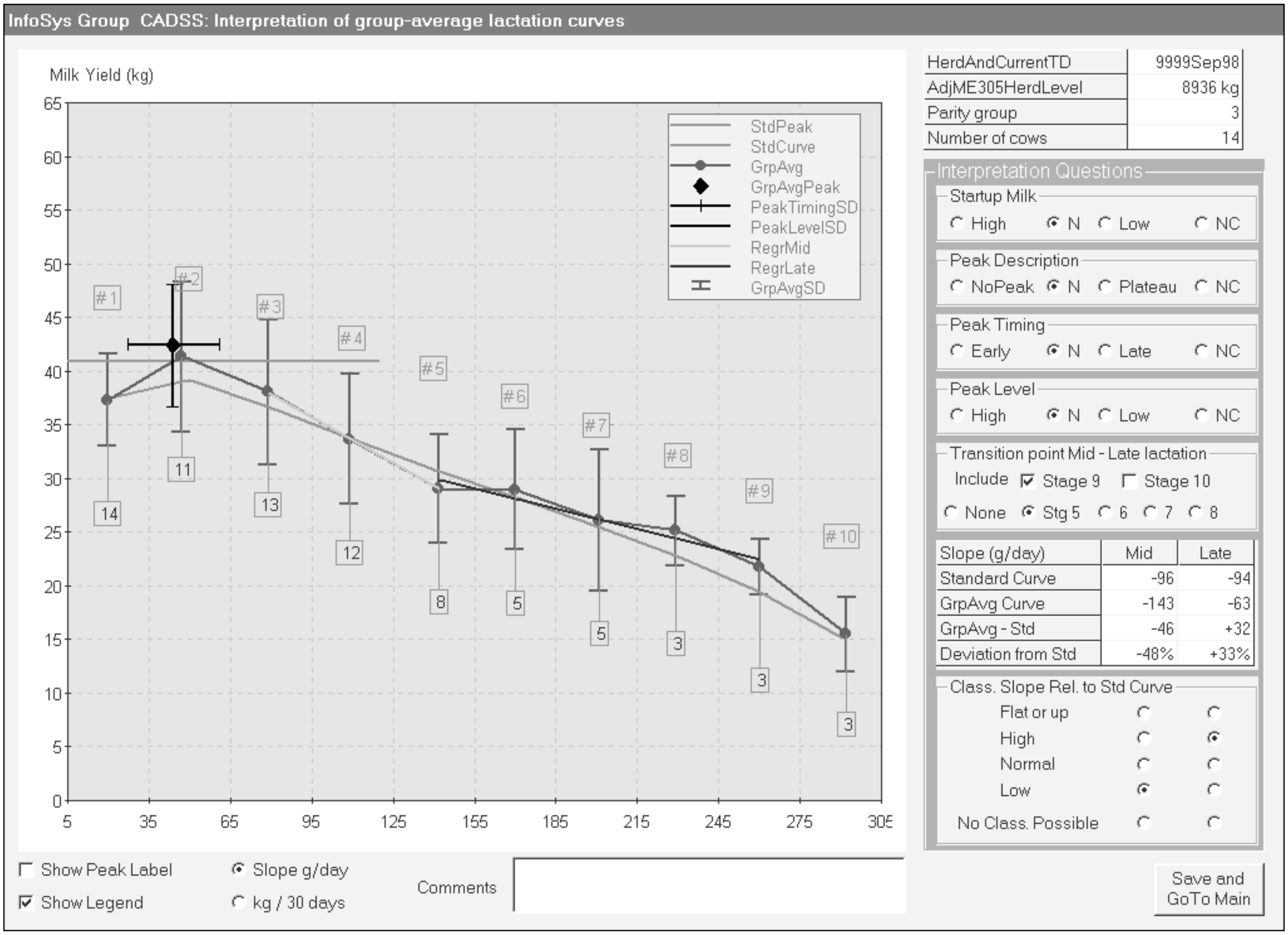
In a previous study (Pietersma et al. 2001), research was initiated to explore the usefulness of machine-learning assisted knowledge acquisition for group-average lactation-curve analysis. A large amount of consultation with two dairy-nutrition specialists was required to elicit the domain vocabulary, decompose the overall problem area into three sub-problems (removal of outlier data, interpretation of parity-group average lactation curves, and diagnosis of detected abnormalities), and to develop a case-acquisition and decision-support system (CADSS). 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 specific project presented here was to develop a knowledge-based module for implementation in the CADSS to partially automate the interpretation of parity-group average lactation curves. The objectives were: 1) to induce decision trees for each classification task involved with the interpretation of parity-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 the plausibility of the induced decision trees with the specialist.

MATERIALS and METHODS

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 were 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. 2001). The CADSS allowed the specialist to compare the lactation curves of individual cows with parity-group average lactation curves and with standard lactation curves for the Holstein breed. Furthermore, for each individual test, additional information including the milk protein-to-fat ratio and the somatic cell count could be viewed. For the 33 herds used in this study, the number of cows per herd had a median of 38 and ranged from 20 to 102.



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

Fig. 1. Screen capture of the case-acquisition software module to interpret parity-group average lactation curves.

The domain specialist identified 2.4% of the lactations of individual cows as outlier (Pietersma et al. 2002b) and 1.4% of the tests within the remaining lactations as outlier (Pietersma et al. 2002). In the second step of the overall analysis process, addressed in this research, non-outlier milk yield data were used to create, for each parity group, a parity-group average lactation curve and a group-average peak production. The median number of lactations used to create the parity-group average lactation curves was 11, 8, and 18 for parity groups 1, 2, and 3, respectively. The specialist analyzed the parity-group average lactation descriptions for the 33 herds using the CADSS, which led to a total of 99 interpretations.

Figure 1 shows a screen capture of the CADSS module for the interpretation of parity-group average lactation curves. The parity-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-shaped marker and both horizontal and vertical error bars in Fig. 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 or minus the standard deviation. In addition, the standard lactation curve and peak level are shown for the parity group in question as well as the milk production level of the herd.

With the interpretation module, six main classification tasks had already been identified by the specialists (Pietersma et al. 2001). 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 parity-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 (Fig. 1). For each classification task, a set of classes had been defined by the specialists. For example, the peak level could be classified as either "High", "Normal" ("N" in Fig. 1), "Low", or "Not Classifiable" ("NC" in Fig. 1).

Table 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	NoPeak	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

Interpretation of the 99 parity-group average lactation descriptions by the domain specialist using the CADSS resulted in a distribution of cases per class for each classification task as shown in Table 1. 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 Fig. 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.

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 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 2).

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

Table 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

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 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 class "High" represented a parity-group average lactation curve with a larger value for the slope than the standard lactation curve. For example, the slope during late lactation in Fig. 1 is considered "High". In this case the slope of the linear regression line through the group-average data was –63 g/d, while the slope for the standard curve was –94 g/d. The class "Flat" represents an extreme form of the class "High", with a group-average curve after the peak approximating a horizontal line. To enable detailed analysis of the classification performance in machine-learning experiments, the merged task was decomposed into two steps. The first step determined whether the slope should be considered "Low", "Normal", or "High+Flat". For cases considered as "High+Flat" in the first step, a second sub-task determined whether the slope should be considered as "Flat" with classes "True" and "False" (Table 2). Cases classified as "False" in the second step represented a classification as "High".

For the task "Include Stage 9", most cases were classified as "True" which was considered as the default situation (Table 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 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.

Creation of attributes

The domain specialist had access to complex graphical information to interpret the various aspects of the parity-group average lactation curves. The CADSS provided numerical data only for the slope during mid and late lactation (Fig. 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 (Pietersma et al. 2002). Thus, to make machine learning feasible with the relatively small number of example cases available for learning, considerable time was spent in developing attributes that were expected to be useful for discerning between classes.

With the CADSS, the domain specialist could compare the group-average performance with standard lactation curves and peak levels that were used as benchmarks. Thus, in order 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 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 also 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, an "expected" start-up milk yield 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 attributes were then 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 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 parity-group average lactation curve.

For some cases, certain attributes were irrelevant and their value could, therefore, 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 irrelevant situation (Pietersma et al. 2003a; Witten and Frank 2000).

Decision-tree induction algorithm

In this study decision-tree induction was performed using version 3.6 of CART 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 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 ten-fold cross-validation (explained below) 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 (Breiman et

Table 3. Listing of potentially predictive attributes for each of the classification tasks used for machine learning.

Task*	Description of attribute or attributes
SM	Absolute, relative, and number of standard deviations (numSD) deviation of the parity-group average lactation curve (GrpAvgCrv) milk at Stage 1 from the standard curve (StdCrv) milk at Stage 1
SM	Absolute, relative, and numSD deviation of GrpAvg milk at Stage 1 from prediction based on maximum GrpAvg milk at Stage 2 and 3 and the shape of StdCrv
NP, PP	Stage 1 has maximum GrpAvgCrv milk
NP, PP	Absolute, relative, and numSD deviation of GrpAvg milk at Stage 1 from maximum 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 maximum milk stage or from Stage 1 to 2
NP, PP	Deviation slope LinRegr through GrpAvgCrv from slope StdCrv for Stage 2 to 3, 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	Absolute, relative, and numSD deviation of GrpAvg peak milk from Std peak
PL	Absolute, relative, and numSD deviation of maximum GrpAvgCrv milk from maximum StdCrv milk
E9	SD and number of test GrpAvgCrv at Stage 9
E9	Absolute, relative, and numSD deviation of GrpAvgCrv milk at Stage 9 from prediction based on LinRegr through previous 2 and previous 3 stages
SS	Average SD of stages of GrpAvgCrv after peak
SS	Maximum numSD deviation of GrpAvgCrv milk from LinRegr GrpAvgCrv after peak
SS	Relative deviation of root mean squares error (RMSE) for no transition point from minimum RMSE for any transition point
SS	Maximum difference between relative 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 average RMSE for two regression lines for mid and late lactation for transition points at Stage 5, 6, and 7
TS	Rank of the average, maximum, or difference for the relative 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	Average SD of GrpAvgCrv for section
LS, FS	Absolute, relative, and numSD deviation slope GrpAvgCrv from slope StdCrv for section
FS	Slope GrpAvgCrv for section

* 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.

al. 1984), 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).

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 randomly 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 twice as many cases per test set as 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. 2003a). 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.

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 considered worse than others. For example, with a diagnostic test it may be more costly to classify a person with a serious disease as healthy than to classify a healthy person as possibly 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 and true negatives are correct classifications, a false positive is an actual negative case incorrectly predicted as positive, and a false negative is an actual positive case incorrectly predicted as negative. To allow for detailed analysis of the performance of the generated classifiers, the following four performance indices were used: 1) true positive (TP) rate, defined as the true positives as a proportion of the actual positives; 2) false positive (FP) rate, defined as the false positives as a proportion of the actual negatives; 3) predictive value positive (PVP), defined as the true 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 (Weiss and Kulikowski 1991; Witten and Frank 2000). In the literature, the TP rate is also referred to as true positive proportion (Swets 1988) or sensitivity (Weiss and Kulikowski 1991), while the FP rate is sometimes called false positive proportion (Swets 1988). In some domains the sensitivity is used in conjunction with the specificity. The specificity can be defined as the true negatives as a proportion of the actual negatives and is equivalent to 1 – FP rate (Weiss and Kulikowski 1991). 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 (POP) and can be mathematically derived from the TP and FP rates for a given prevalence level using:

$$PPR = POP \times TPrate + \left(1 - POP\right) \times FPrate \tag{1}$$

$$PVP = POP \times TPrate / PPR \tag{2}$$

Relative operating characteristic (ROC) curves (Swets 1988; Yang et al. 1999) were used to visualize the trade-off between correctly classifying normal cases and correctly classifying abnormal cases, i.e. between the sensitivity and specificity. An ROC curve consists of the TP rate plotted against the FP rate. In ROC space, the upper left point (0,100) represents perfect classification performance (0% FP rate and 100% TP rate). The lower left point (0,0) is achieved with 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. Thus, each point on the ROC curve represents a classifier with a particular trade-off between sensitivity and specificity. The closer an ROC curve approximates the lines connecting (0,0) with (0,100) and (100,100), the better the performance. 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, e.g. “High” or a “Low” peak level, 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).

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, costs of false positives and false negatives, and end-user preference. 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 data used for training (Witten and Frank 2000). Resubstitution FP and TP rates were used to verify quantitatively 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. 2003b). In addition, the final decision trees were evaluated by the domain specialist to verify qualitatively the plausibility of the induced rules. This allowed for the removal of counter-intuitive decision nodes and the use of alternative splits that were provided by the CART algorithm.

RESULTS

Classification performance

Start-up milk For the task “Start-up milk”, the ROC curve showed good classification performance, with 89% TP rate

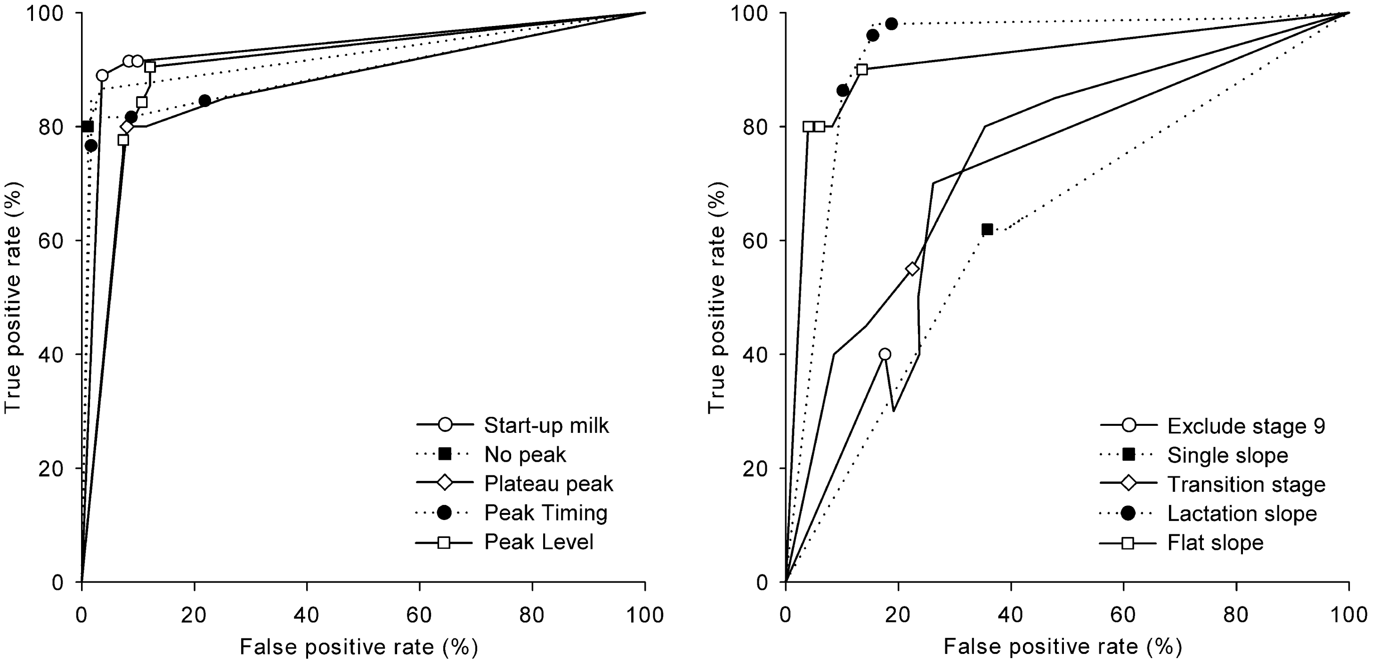


Fig. 2. Relative operating characteristic curves for 10 classification tasks with markers showing the cross-validation performance for each induced decision tree.

achieved at 3.6% FP rate (Fig. 2). Three different points along the ROC curve, each indicated with a marker in Fig. 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. The specific misclassification cost settings used with these three final decision trees were 0.19, 0.5, and 1, respectively (Table 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).

Figure 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 among 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. 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 42 cases, respectively, as either "High" or "Low" (Fig. 3).

No peak and Plateau peak The ROC curve for the "No peak" classification task showed fairly good performance (Fig. 2), with 80% TP rate achieved at 1% FP rate (Table 4). The "Plateau peak" task showed somewhat poorer performance 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 4).

Peak timing and Peak level For the "Peak timing" classification task, the ROC curve showed fairly good performance (Fig. 2). A 77% TP rate was achieved at 2% FP rate, while 85% TP rate required a relatively high FP rate of 22% (Table 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" (Fig. 2), reaching 91% TP rate at 12% FP rate. For both classification tasks, three

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

decision trees were induced from the entire data set with reasonable values for PPR and PVP (Table 4).

Exclude Stage 9, Single slope, and Transition stage The "Exclude Stage 9" classification task showed poor classification performance (Fig. 2) and a very high standard error for the TP rate estimate (Table 4). This may have been caused by the very small number of positive cases (8). For this task, the TP rate of the ROC curve actually decreased at some point with an increase in the FP rate (Fig. 2). This unusual type of response of the ROC curve can be explained by the fact that a small increase in the cost of misclassifying positive cases may trigger the decision-tree induction algorithm to induce a decision tree that results in a lower TP rate or lower FP rate for the independent test data (Pietersma et al. 2003b). With the "Exclude Stage 9" 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. Both tasks showed poor classification performance (Fig. 2), and for each task, a single decision tree was chosen in consultation with the domain specialist (Table 4). These three classification tasks are only of indirect importance for the interpretation of parity-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".

Lactation slope and Flat slope The "Lactation slope" and "Flat slope" tasks showed good classification performance, achieving TP rates higher than 80% at relatively low FP rates (Fig. 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 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.

Quantitative evaluation of the plausibility of the final decision trees

For most of the decision trees induced from the entire data set, the resubstitution performance was very similar to the

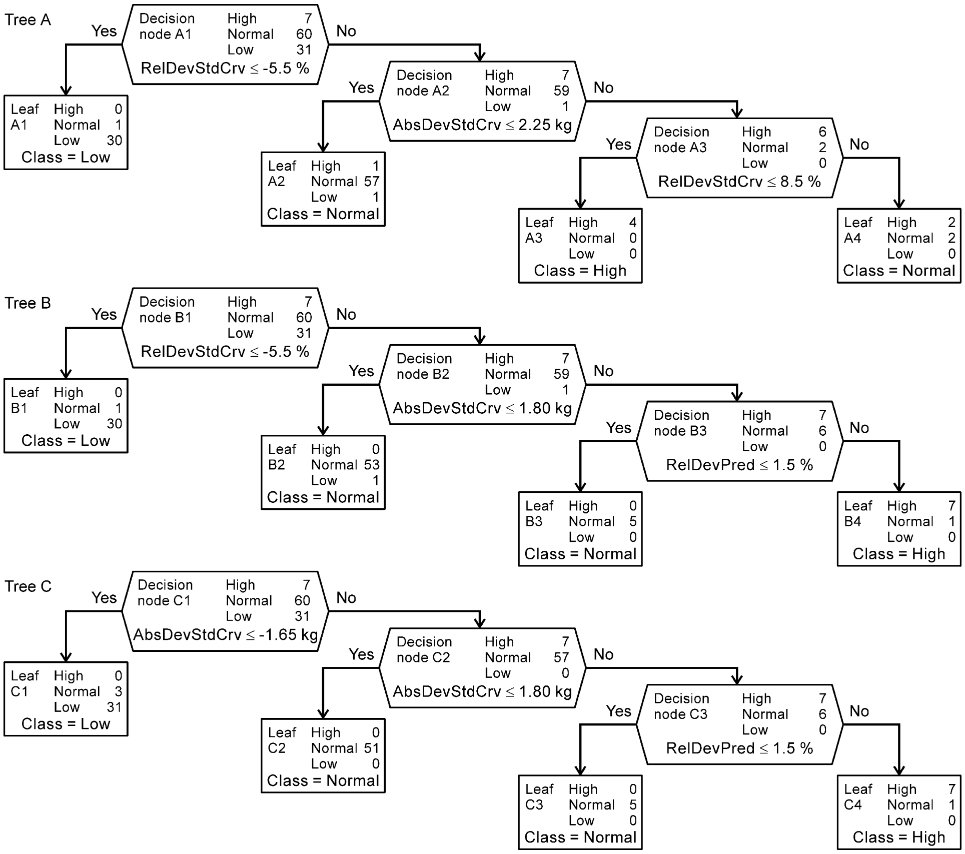

Tree A
Yes
No
Decision node A1 High 7 Normal 60 Low 31
RelDevStdCrv ≤ -5.5 %
Leaf A1 High 0 Normal 1 Low 30
Class = Low
Decision node A2 High 7 Normal 59 Low 1
AbsDevStdCrv ≤ 2.25 kg
Leaf A2 High 1 Normal 57 Low 1
Class = Normal
Decision node A3 High 6 Normal 2 Low 0
RelDevStdCrv ≤ 8.5 %
Leaf A3 High 4 Normal 0 Low 0
Class = High
Leaf A4 High 2 Normal 2 Low 0
Class = Normal
Tree B
Decision node B1 High 7 Normal 60 Low 31
RelDevStdCrv ≤ -5.5 %
Leaf B1 High 0 Normal 1 Low 30
Class = Low
Decision node B2 High 7 Normal 59 Low 1
AbsDevStdCrv ≤ 1.80 kg
Leaf B2 High 0 Normal 53 Low 1
Class = Normal
Decision node B3 High 7 Normal 6 Low 0
RelDevPred ≤ 1.5 %
Leaf B3 High 0 Normal 5 Low 0
Class = Normal
Leaf B4 High 7 Normal 1 Low 0
Class = High
Tree C
Decision node C1 High 7 Normal 60 Low 31
AbsDevStdCrv ≤ -1.65 kg
Leaf C1 High 0 Normal 3 Low 31
Class = Low
Decision node C2 High 7 Normal 57 Low 0
AbsDevStdCrv ≤ 1.80 kg
Leaf C2 High 0 Normal 51 Low 0
Class = Normal
Decision node C3 High 7 Normal 6 Low 0
RelDevPred ≤ 1.5 %
Leaf C3 High 0 Normal 5 Low 0
Class = Normal
Leaf C4 High 7 Normal 1 Low 0
Class = High

Fig. 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.

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 5). However, for five 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 5). This may have been caused by the internal 10-fold cross validation used by the CART algorithm 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, 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 generated by CART for the task and misclassification cost settings in question (Table 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 with the desired trade-off between sensitivity and specificity.

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 Fig. 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 particular 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. The alternative split resulted in one additional false negative case for the entire data set. However, the adjusted decision tree was considered as plausible by the domain specialist and expected to lead to improved classification performance when used to classify new data in a KBS to interpret parity-group average lactation curves.

DISCUSSION

The decision trees for the main classification tasks (start-up milk, no peak, plateau peak, peak timing, peak level, lactation slope, and flat slope) showed good classification performance in cross-validation experiments. 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

Table 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		FP* rate		TP rate		Leaf nodes (#)	FP rate (%)	TP rate (%)	Leaf nodes (#)	FP rate (%)	TP rate (%)
		(#)	s.e.	(%)	s.e.	(%)	s.e.						
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

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 six multi-class tasks, 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 other three 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.

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 an equivalent 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 of labeled cases, but, relying on the expertise of the domain specialist, were expected to lead to improved performance with new data. For exact estimates of the classification performance of the manually adjusted decision trees, the final classifiers would have to be tested using a completely new set of example cases interpreted by the domain specialist.

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 a 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. Adjustment of the tendency of the system to indicate a situation as abnormal was also possible with the KBS for lactation curve analysis developed at the Texas A&M University but required the user to change thresholds for the deviations from standard data (Fourdraine et al. 1992).

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. 2001). 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.

CONCLUSIONS

This research suggests that automatically induced decision trees are able to match closely the interpretation of parity-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 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.

ACKNOWLEDGEMENTS

The authors acknowledge 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 were provided by the Québec Dairy Herd Analysis Service (Programme d'analyse des troupeaux laitiers du Québec).

REFERENCES

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.

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone. 1984. *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

Dhar, V. and R. Stein. 1997. *Intelligent Decision Support Methods*. Upper Saddle River, NJ: Prentice Hall.

Durkin, J. 1994. *Expert Systems: Design and Development*. New York, NY: Macmillan.

Fourdraine, R.H., M.A. Tomaszewski and T.J. Cannon. 1992. Dairy herd lactation expert system, a program to analyze and evaluate lactation curves. In *Proceedings International Symposium on Prospects for Automatic Milking*, 331-337. Wageningen, Netherlands: Pudoc Scientific Publishers.

Kononenko, I., I. Bratko and M. Kukar. 1998. Application of machine learning to medical diagnosis. In *Machine Learning and Data Mining: Methods and Applications*, eds. R.S. Michalski, I. Bratko and M. Kubat, 389-408. Chichester, England: John Wiley and Sons.

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.

Langley, P. and H. A. Simon. 1995. Applications of machine learning and rule induction. *Communications of the Association for Computing Machinery* 38(11): 55-64.

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. In *Symposium Bovins Laitiers*, 13-39. Québec, QC: Conceil des productions animales du Québec.

McQueen, R.J., S.R. Garner, C.G. Nevill-Manning and I.H. Witten. 1995. Applying machine learning to agricultural data. *Computers and Electronics in Agriculture* 12: 275-293.

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. *Journal of Dairy Science* 81: 2962-2972.

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. *Journal of Dairy Science* 84: 730–739.

Pietersma, D., R. Lacroix, D. Lefebvre and K.M. Wade. 2002. Machine-learning assisted knowledge acquisition to filter lactation curve data. *Transactions of the ASAE* 45(5): 1637-1650.

Pietersma, D., R. Lacroix, D. Lefebvre and K.M. Wade. 2003a. Performance analysis for machine-learning experiments using small data sets. *Computers and Electronics in Agriculture* 38(1):1-17.

Pietersma, D., R. Lacroix, D. Lefebvre and K.M. Wade. 2003b. Induction and evaluation of decision trees for lactation curve analysis. *Computers and Electronics in Agriculture* 38(1):19-32.

Programme d'analyse des troupeaux laitiers du Québec. 2001. Rapport de production 2000. Ste. Anne de Bellevue, QC: Programme d’analyse des troupeaux laitiers du Québec.

Skidmore, A.L., A. Brand and C.J. Sniffen. 1996. Monitoring milk production: Decision making and follow-up. In *Herd Health and Production Management in Dairy Practice,* eds. A. Brand, J.P.T.M. Noordhuizen and Y. H. Schukken, 263-281. Wageningen, Netherlands: Wageningen Pers.

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

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

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

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*. San Mateo, CA: Morgan Kaufmann.

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. *Agricultural Systems* 31: 83-96.

Witten, I.H. and E. Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco, CA: Morgan Kaufmann.

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