When in Doubt ... Be Indecisive

Linda C. van der Gaag[1], Silja Renooij[1],
Wilma Steeneveld[2], and Henk Hogeveen[2]

[1] Department of Information and Computing Sciences, Utrecht University,
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
{linda,silja}@cs.uu.nl

[2] Department of Farm Animal Health, Utrecht University
{w.steeneveld,h.hogeveen}@uu.nl

Abstract. For a presented case, a Bayesian network classifier in essence computes a posterior probability distribution over its class variable. Based upon this distribution, the classifier's classification function returns a single, determinate class value and thereby hides the uncertainty involved. To provide reliable decision support, however, the classifier should be able to convey indecisiveness if the posterior distribution computed for the case does not clearly favour one class value over another. In this paper we present an approach for this purpose, and introduce new measures to capture the performance and practicability of such classifiers.

Keywords: Probabilistic classification, indecisiveness.

1 Introduction

Many real-life problems can be viewed as classification problems in which a case described in terms of a number of features is to be assigned to one of several distinct classes. In the management of animal health on dairy farms, for example, the problem of establishing an appropriate diagnosis for a combination of clinical signs can be viewed as a classification problem in which a cow has to be assigned to one of a number of diagnostic categories. Bayesian network classifiers have gained considerable popularity for solving such problems. These classifiers embed a Bayesian network composed of a single class variable, modelling the possible classes for the problem under study, and a set of feature variables, modelling the features that constitute the basis for distinguishing between the classes. For a presented case, this network serves to establish the posterior probability distribution over the class variable given the case's features. Based upon this distribution, the classifier assigns a single, determinate class to the case [3,4].

Bayesian network classifiers are being applied in a wide range of domains for a variety of problems; for some recent examples in the biomedical field we refer to [1,2,5,6,7]. In some applications, such as in automated spam filtering, the class value returned by the classifier conveys sufficient information to solve the problem at hand and does not require any further decisions from the user. We have noticed however, that in other applications the returned class value may not always provide a sufficient basis for reliable further decision making.

C. Sossai and G. Chemello (Eds.): ECSQARU 2009, LNAI 5590, pp. 518–529, 2009.

In our domain of animal health, for example, the actual problem is not just to establish the most likely diagnosis but, even more importantly, to control the disease patterns in a dairy herd by appropriate treatment. The diagnostic category returned by the classifier does not necessarily provide sufficient information for this purpose, as it hides the uncertainty involved in the classification result. A differential diagnosis in which two or more diagnostic categories have almost equal probabilities, for example, could call for a different treatment regime than a differential diagnosis in which one disease clearly stands out. From the classification result however, the decision maker cannot distinguish between clear-cut cases and cases which in essence are inconclusive.

In this paper, we enhance Bayesian network classifiers by allowing them to be indecisive. The basic idea is that the classifier returns a classification result for a case only if a single class value stands out convincingly in the posterior probability distribution computed over the class variable. If none of the possible class values receives sufficient support in the computed distribution, then the classifier does not return a determinate classification result but leaves the case unclassified instead. The case at hand then is left to the human decision maker, who evaluates the probabilistic information computed by the classifier in view of further decision making. For our new type of classifier we introduce measures to express its classification performance and its practicability. These measures closely resemble the well-known concept of classification accuracy, yet take into account the classifier's reduced practicability as a result of its occasional indecisiveness. We illustrate the usefulness of our new type of classifier for an example application in animal health management.

The paper is organised as follows. In Section 2 we review Bayesian network classifiers and introduce our domain of application. In Section 3, we discuss the well-known concept of classification accuracy and study its dependence on the probability thresholds commonly used by classification functions. In Section 4, we introduce the new concept of stratifying classifier and define associated measures of classification performance and practicability. We illustrate our concept of stratifying classifier and its associated measures for our domain of application in Section 5. The paper ends with our concluding observations in Section 6.

2 Preliminaries

In this section, we briefly review Bayesian network classifiers. In doing so, we restrict the discussion to naive Bayesian classifiers with binary variables only; the illustrated concepts, however, are readily extended to non-binary variables and to Bayesian network classifiers of more general topological structure. In addition, we introduce our application domain, which will serve as a running example.

2.1 Naive Bayesian Classifiers

A naive Bayesian classifier includes a designated class variable C and a set $\mathbf{F}$ of one or more feature variables F_i. If a variable V_j adopts the value *true*, we will write v_j; we use $\bar{v}_j$ to denote $V_j = \mathit{false}$. A joint value assignment to all feature variables concerned is termed a case and will be denoted by $\mathbf{f}$. The classifier's graphical structure includes arcs $C \rightarrow F_i$ which capture dependence of each feature variable on the class variable, yet mutual independence of any two feature variables given this class variable. The classifier further specifies a prior probability distribution $\Pr(C)$ over the class variable and a set of conditional distributions $\Pr(F_i \mid C)$ for each feature variable. Naive Bayesian classifiers are typically constructed by extracting the most discriminating feature variables, and their associated probability distributions, from a set of example cases.

A naive Bayesian classifier in essence allows the computation of any probability of interest over its variables. More specifically, it provides for establishing, for a presented case $\mathbf{f}$, the posterior probability distribution $\Pr(C \mid \mathbf{f})$ over the class variable given the case's features. The classifier does not return this probability distribution, but instead establishes a single, determinate class value for its output, using a classification function. For the binary class variable C, this function takes the following form:

$$\mathit{class}(C, t; \mathbf{f}) = \begin{cases} c, & \text{if} \Pr(c \mid \mathbf{f}) \geq t \\ \bar{c}, & \text{otherwise} \end{cases}$$

where t is a pre-defined threshold value. In most applications, the winner-takes-all rule is used for the model's classification function, which takes $t = 0.50$. For applications with skewed prior distributions over the class variable, however, other values of t are preferred. In general, the choice of an appropriate threshold value is domain dependent. If the classification function of a classifier returns $\mathit{class}(C, t; \mathbf{f}) = c$ for a case $\mathbf{f}$, then we say that $\mathbf{f}$ is classified as belonging to class c; analogous terminology is used for $\mathit{class}(C, t; \mathbf{f}) = \bar{c}$.

2.2 An Example Application in Dairy Science

Clinical mastitis is one of the most frequent and cost incurring diseases in a dairy herd. The disease affects the cow's udder, causing a reduction of the cow's milk production and an increased risk of the cow being culled. Clinical mastitis can be caused by a large variety of pathogens; diagnosis of the causing pathogen is done by bacteriological culturing. Bacteriological culturing takes at least three days. Yet, a timely administered treatment is important to eliminate the disease and to prevent recurrence as much as possible. Ideally, the disease is controlled with limited use of antibiotics, to reduce the risk of antibiotic contamination of the milk and to minimise the impact of treatment on antimicrobial resistance. The most appropriate treatment is highly dependent upon the specific pathogen causing the disease in the current instance, however. If a single specific pathogen is convincingly favoured over other possible causal pathogens, a narrow-spectrum antibiotic would be preferred; in case two or more pathogens are quite likely, broad-spectrum antibiotic treatment would be more appropriate. Unfortunately, a farmer will typically have to decide upon treatment in uncertainty, before the actual causal pathogen is known from bacteriological culturing.

To support a dairy farmer in his treatment decisions, we constructed a standard naive Bayesian classifier. Cases of clinical mastitis to be presented to the classifier are described by a number of features which range from the cow's mastitis history to such clinical signs as the appearance of the milk and the cow's demeanor. For a case, the classifier returns the most likely Gram-status of the pathogen causing the mastitis; this status is an important indicator for the type of antibiotics to be applied. For constructing the classifier, we had available a set of 3 833 clinical mastitis cases; in 2 706 (or 70.6%) of these cases the disease was caused by a Gram-positive pathogen, and in 1 127 cases the causal pathogen was Gram-negative. We used 2 631 cases (67%) for constructing the classifier and retained the remaining 1 202 cases for studying its performance. The constructed classifier was optimised for a classification threshold value of $t = 0.71$.

3 The Accuracy Measure and Its Threshold Dependence

The performance of a Bayesian network classifier is commonly summarised as the proportion of cases which are assigned to their true class value. In this section we review this measure of accuracy. We further argue that the accuracy of a Bayesian network classifier depends heavily on the threshold value used in its classification function. We investigate this dependence and study the effects of varying the threshold value on the classifier's accuracy.

3.1 The Measure of Accuracy

We consider a naive Bayesian classifier with the classification function *class*. We further consider a set $\mathcal{F}$ of cases for the classifier. The case set $\mathcal{F}$ is partitioned into the set $\mathcal{F}^+$ which includes p cases belonging to class c, and the set $\mathcal{F}^-$ which includes n cases belonging to $\bar{c}$, with $p + n = m$. A case belonging to class c will be termed a positive case; likewise, a case with class $\bar{c}$ is coined a negative case. The function *class* of the classifier now partitions the case set $\mathcal{F}$ into four mutually exclusive and collectively exhaustive subsets; the basic idea of this partitioning is shown in Table 1. The first subset includes all cases from $\mathcal{F}^+$ which are classified as belonging to class c by the classifier. This set is called the set of true positive cases, denoted by TP; the size of this set is denoted by tp. The cases from $\mathcal{F}^+$ which are incorrectly classified as belonging to $\bar{c}$ constitute the set of false negative cases, denoted by FN; the size of this set is fn. Note that $\mathsf{TP} \cap \mathsf{FN} = \varnothing$ and $\mathsf{TP} \cup \mathsf{FN} = \mathcal{F}^+$, and hence that $\mathsf{tp} + \mathsf{fn} = p$. Likewise, we define the set TN of true negative cases and the set FP of false positive cases; the sizes of these sets are tn and fp respectively, with $\mathsf{tn} + \mathsf{fp} = n$. The performance

Table 1. The sizes of the partition subsets resulting from the classification function

		classifier		
		c	$\bar{c}$	*total*
data	c	tp	fn	p
	$\bar{c}$	fp	tn	n
total				m

of a Bayesian network classifier now is commonly captured by its (*empirical*) *accuracy*, which is defined as the proportion of correctly classified cases for a given case set $\mathcal{F}$:

$$accuracy(\mathcal{F}) = \frac{\mathsf{tp} + \mathsf{tn}}{m}$$

The case set from which a classifier's accuracy is established, is usually omitted from the notation; we adopt this convention and from now on leave $\mathcal{F}$ implicit.

3.2 Dependency on the Classification Threshold

The measure of accuracy reviewed above pertains to a given Bayesian network classifier with a fixed classification function. The accuracy of a classifier in general depends on the threshold value t used in its classification function. More specifically, the sizes of the four partition subsets constructed from a case set $\mathcal{F}$ are threshold dependent. For example, a classification function with $t = 0$ would result in each and every case being assigned to class c, from which we would have that $\mathsf{tp} + \mathsf{fp} = m$ and $\mathsf{fn} = \mathsf{tn} = 0$; on the other hand, a threshold value $t > 1$ would distribute all cases over the two sets TN and FN, from which we would have that $\mathsf{fn} + \mathsf{tn} = m$ and $\mathsf{tp} = \mathsf{fp} = 0$. From now on, we make this dependency on the threshold value explicit in our notations, by writing $\mathsf{tp}(t)$, $\mathsf{tn}(t)$, $\mathsf{fp}(t)$, and $\mathsf{fn}(t)$ for the sizes of the four sets TP, TN, FP, and FN, respectively. The accuracy of a classifier as a function of the threshold value t then becomes

$$accuracy(t) = \frac{\mathsf{tp}(t) + \mathsf{tn}(t)}{m}$$

The above considerations show that a classifier's accuracy can be manipulated by choosing an appropriate threshold value t. We note that upon varying the value of t from zero to one, cases *migrate* from the sets TP and FP to the sets TN and FN. More specifically, upon varying t, cases from $\mathcal{F}^+$ can migrate, and migrate only, between the sets TP and FN, whereas cases from $\mathcal{F}^-$ can move only between the sets FP and TN. Despite the seeming mutual independence of the numbers of true positives and true negatives, these numbers are traded off through their dependence on the threshold value t: while $\mathsf{tp}(t)$ is non-increasing for increasing values of t, $\mathsf{tn}(t)$ is non-decreasing in t. The changes in size of the sets TP and TN upon varying the threshold value t are illustrated for our example application in Fig. 1.

4 Stratifying Classifiers

In this section we introduce the idea of stratified classification, by defining classification functions for Bayesian network classifiers with two separate threshold values. In addition, we define associated performance measures.

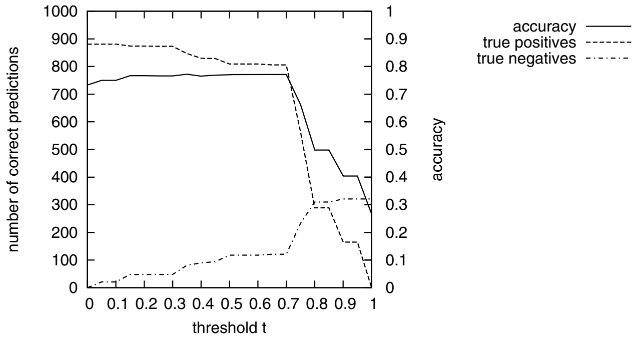


Fig. 1. The accuracy, the number of true positives, and the number of true negatives as functions of the threshold value t for our example classifier and set of 1 202 cases

4.1 Introducing Stratification

The main motivation underlying the introduction of stratifying classifiers is that Bayesian network classifiers should be able to convey indecisiveness, especially when inconclusive cases call for different further decision making than cases with a convincingly outstanding class value. In view of conveying indecisiveness, we now look upon a Bayesian network classifier as distributing a case set over different strata, based upon the computed posterior distribution over the class variable. After stratification, the classifier returns a determinate class value for the cases from some strata and leaves the cases from other strata unclassified. To distinguish between determinate and inconclusive cases, a stratifying classifier employs a partial classification function *class** with two threshold values $t^- \leq t^+$:

$$\mathit{class}^*(C, t^-, t^+; \mathbf{f}) = \begin{cases} c, \text{ if } \Pr(c \mid \mathbf{f}) \geq t^+ \\ \bar{c}, \text{ if } \Pr(c \mid \mathbf{f}) < t^- \end{cases}$$

The threshold value t^- is termed the function's lower threshold value; t^+ is called its upper threshold value. The *-notation is used to denote a function adapted to stratification. Note that for $t^- < t^+$, the stratifying classification function *class** serves to classify only those cases $\mathbf{f}$ for which either $\Pr(c \mid \mathbf{f}) \geq t^+$ or $\Pr(c \mid \mathbf{f}) < t^-$. All cases with $t^- \leq \Pr(c \mid \mathbf{f}) < t^+$ are left unclassified by the stratifying classification function. Further note that the function *class** has the standard, single-threshold classification function as a special case, with $t^- = t^+$.

At first glance, the idea of stratifying classifiers shows similarities to multiway classification and to threshold decision making. In multiway classification, the purpose of the classifier is to distinguish between more than two distinct classes. Stratification in contrast does not increase the number of class values under consideration and thus differs conceptually from multiway classification. The idea of threshold decision making, which was introduced to support physicians during the diagnostic-testing phase in patient management [8], builds upon concepts of decision analysis to establish two patient-specific threshold values, p^- and p^+, on the probability of disease $\Pr(d)$ computed for a patient. These threshold values serve to decide between withholding treatment ($\Pr(d) < p^-$), further diagnostic testing ($p^- \leq \Pr(d) < p^+$), and immediate treatment ($\Pr(d) \geq p^+$). Taking a decision of further diagnostic testing as conveying indecisiveness concerning whether or not to treat the patient, the threshold decision making model can in fact be implemented with a stratifying classifier.

4.2 The Accuracy of a Stratifying Classifier

We consider again the case set $\mathcal{F} = \mathcal{F}^+ \cup \mathcal{F}^-$ with $m = p + n$ cases. The classification function of the stratifying classifier partitions this set in five mutually exclusive and collectively exhaustive subsets. Four of these subsets match the sets TP, FP, TN, and FN introduced above; the fifth set is the set of unclassified cases. The sizes of the five sets again depend upon the threshold values used by the classification function. To capture this dependence, we observe that the stratifying classifier displays the following behaviour:

$$\begin{aligned}
\forall\ \mathbf{f} \in \mathcal{C}^+ = \{\mathbf{f} \mid \mathbf{f} \in \mathcal{F},\ \Pr(c \mid \mathbf{f}) \geq t^+\}&:\ \mathit{class}^*(C, t^-, t^+; \mathbf{f}) = c \\
\forall\ \mathbf{f} \in \mathcal{C}^- = \{\mathbf{f} \mid \mathbf{f} \in \mathcal{F},\ \Pr(c \mid \mathbf{f}) < t^-\}&:\ \mathit{class}^*(C, t^-, t^+; \mathbf{f}) = \bar{c} \\
\forall\ \mathbf{f} \in \mathcal{C}^u = \mathcal{F} \setminus (\mathcal{C}^+ \cup \mathcal{C}^-)&:\ \text{unclassified}
\end{aligned}$$

This observation shows that all cases from the set $\mathcal{C}^+$ are classified as being positive; the set thus is distributed over the two sets TP and FP. Since the size of the set $\mathcal{C}^+$ depends on the upper threshold value t^+, the sizes tp and fp of the sets TP and FP depend on the value t^+ as well; we will write $\mathsf{tp}(t^+)$ and $\mathsf{fp}(t^+)$, respectively, to express this dependence. Similarly, the set $\mathcal{C}^-$ is distributed over TN and FN. The sizes tn and fn of these sets depend on the lower threshold value t^-; we will write $\mathsf{tn}(t^-)$ and $\mathsf{fn}(t^-)$, respectively, to express this dependence.

We recall that, for a standard Bayesian network classifier with a classification function based on a single threshold value t, accuracy is defined as the proportion of cases that are assigned to their true class value:

$$\mathit{accuracy}(t) = \frac{\mathsf{tp}(t) + \mathsf{tn}(t)}{m}$$

The proportion of cases that are correctly classified by a stratifying classifier now equals

$$\mathit{accuracy}(t^-, t^+) = \frac{\mathsf{tp}(t^+) + \mathsf{tn}(t^-)}{m}$$

Since a stratifying classifier may leave some cases unclassified, one or more of the sets TP, FP, TN, and FN may decrease in size compared to those with a standard classifier. More specifically, we find that

$$\mathsf{tp}(t^+) + \mathsf{fn}(t^-) = p^* \leq p \quad \text{and} \quad \mathsf{tn}(t^-) + \mathsf{fp}(t^+) = n^* \leq n$$

where p^* is the number of actually classified cases from $\mathcal{F}^+$ and n^* is the number of classified cases from $\mathcal{F}^-$; $m - p^* - n^*$ cases are left unclassified by the stratifying classifier. Now, if the stratification results in smaller sets TP and/or TN, then some of the cases considered inconclusive after stratification would have been classified correctly, yet not convincingly, without the stratification. The accuracy of the stratifying classifier then is smaller than that of the standard classifier. On the other hand, if the stratification affects neither TP nor TN, then the accuracy of the stratifying classifier remains unchanged compared to that of the standard classifier, which indicates that all cases considered inconclusive after stratification would have been classified incorrectly without the stratification. For the stratifying classifier, the standard measure of accuracy thus still captures the proportion of correctly classified cases from *all* cases presented to the classifier, including those considered inconclusive. To capture the proportion of correctly classified cases among the cases that were actually classified, we now introduce a measure of stratified accuracy, defined by

$$accuracy^*(t^-, t^+) = \frac{\mathsf{tp}(t^+) + \mathsf{tn}(t^-)}{m^*}$$

where m^* equals $p^* + n^* = \mathsf{tp}(t^+) + \mathsf{tn}(t^-) + \mathsf{fp}(t^+) + \mathsf{fn}(t^-)$.

The measure of stratified accuracy is in essence defined in terms of the sets $\mathcal{C}^+$ and $\mathcal{C}^-$. The measure can also be related to the set $\mathcal{C}^u$ of cases that are left unclassified by the stratifying classifier. To this end, we consider the distribution of these inconclusive cases over the sets TP, FP, TN, and FN with a standard classifier. For each $\mathbf{f} \in \mathcal{C}^u$, we have that

$$class(C, t^-; \mathbf{f}) = c \quad \text{and} \quad class(C, t^+; \mathbf{f}) = \bar{c}$$

With the threshold value t^-, therefore, a standard classifier would distribute all cases from $\mathcal{C}^u$ over the two sets TP and FP, with sizes $\mathsf{tp}(t^-)$ and $\mathsf{fp}(t^-)$, respectively. Similarly, with the threshold value t^+, all cases from $\mathcal{C}^u$ would be distributed over the sets TN and FN, with sizes $\mathsf{tn}(t^+)$ and $\mathsf{fn}(t^+)$. Upon varying the threshold value from t^- to t^+, therefore, the cases $\mathbf{f} \in \mathcal{C}^u$ would migrate from the set TP to the set FN, and from FP to TN. This observation underlies the following formula:

$$accuracy^*(t^-, t^+) = \frac{\mathsf{tp}(t^-) - \Delta\mathsf{tp} + \mathsf{tn}(t^+) - \Delta\mathsf{tn}}{m - \Delta\mathsf{tp} - \Delta\mathsf{tn}}$$

where $\Delta\mathsf{tp} = \mathsf{tp}(t^-) - \mathsf{tp}(t^+)$ is the number of cases from $\mathcal{C}^u \cap \mathcal{F}^+$ that would be incorrectly classified as negative if t^+ were to be taken as the single threshold value; $\Delta\mathsf{tn} = \mathsf{tn}(t^+) - \mathsf{tn}(t^-)$ has an analogous interpretation.

The effects of stratification on the accuracy of a classifier can be studied by comparing the resulting stratified accuracy to the standard accuracy. We consider to this end a standard Bayesian network classifier with the classification function $class(C, t; \mathbf{f})$. Introducing stratification into this classifier entails choosing two threshold values t^- and t^+, $t^- \leq t \leq t^+$, and replacing the function *class* by the stratifying classification function $class^*$. If neither the set TP nor the set TN is affected by the stratification, that is, if $\mathsf{tp}(t^+) = \mathsf{tp}(t)$ and $\mathsf{tn}(t^-) = \mathsf{tn}(t)$, we find that

$$accuracy(t^-, t^+) = accuracy(t) \quad \text{and} \quad accuracy^*(t^-, t^+) \geq accuracy(t)$$

where equality holds for the formula on the right whenever $\mathsf{tp}(t^-) = \mathsf{tp}(t)$ and $\mathsf{tn}(t^+) = \mathsf{tn}(t)$. If the stratification results in a decrease in size of the sets TP and/or TN, then the standard accuracy decreases with the stratification: $accuracy(t^-, t^+) < accuracy(t)$. The stratified accuracy $accuracy^*(t^-, t^+)$, however, can be smaller than, equal to, or larger than the standard accuracy, depending on the size of the set $\mathcal{C}^u$ of inconclusive cases and the standard classifier's performance on $\mathcal{C}^u$. If $accuracy^*(t^-, t^+) < accuracy(t)$, we say that stratification results in a deterioration in the performance of the classifier. Such a deterioration indicates that, among the unclassified cases, a relatively large number were classified correctly prior to the stratification. It further means that the cases which remain incorrectly classified after the stratification, are cases for which high posterior probabilities are established for the *in*correct class value. Often, however, the introduction of stratification will result in an improvement of the performance of a classifier, that is, in $accuracy^*(t^-, t^+) > accuracy(t)$. An appropriate choice of threshold values can in fact result in extremely high stratified accuracies, possibly even equal to 1.

4.3 The Classification Percentage

In most applications, the introduction of stratification into a Bayesian network classifier will result in an increased stratified accuracy. The improvement in classification performance, however, typically comes at the price of a reduced practicability of the classifier for decision support. To capture the issue of practicability, we introduce the concept of *classification percentage*, which equals the proportion of cases that are classified:

$$classification\_percentage = \frac{m^*}{m} \cdot 100\%$$

Note that a standard classifier has a classification percentage of 100%. By introducing stratification, the classification percentage will typically decrease. When viewing a stratifying classifier as a tool for support to a decision maker in his daily practice, the classification percentage indicates, given the stratification under consideration, the percentage of cases for which the tool will actually advance the decision-making process. Alternatively, the classification percentage conveys information about the percentage of cases for which the tool will be indecisive, that is, for which the tool will leave the actual decision to the decision maker.

5 Stratification in the Example Application

In our application domain of animal health management, a dairy farmer typically has to decide upon treatment of a cow with clinical mastitis before knowing the pathogen that causes the disease. As a result, often broad-spectrum antibiotics are administered, where narrow-spectrum antibiotic treatment is preferred. The administration of narrow-spectrum antibiotics is possible, however, only if one specific pathogen is convincingly favoured over all others. Our naive Bayesian

Table 2. Predicted and actual numbers of positive and negative cases for our stratifying classifier, using threshold values $t^- = 0.30$ and $t^+ = 0.80$, with 1 202 cases

		classifier		
		+	−	*total*
data	+	289	8	297
	−	11	48	59
	total			356

classifier supports the choice of antibiotics by classifying mastitis cases according to the Gram-status of the causal pathogen. We recall that this Gram-status is an important indicator for the type of antibiotics to be used. The predicted Gram-status, however, may be quite uncertain for cases with a posterior probability close to the threshold value of the classification function. For such cases, in fact, a broad-spectrum treatment would still be preferred. In this section we use our example application to illustrate the concepts, measures and observations put forward in the previous section.

With our standard naive Bayesian classifier and with the case set of 1 202 mastitis cases, we find the following values tp and tn for the sizes of the sets TP and TN of correctly classified cases, for different threshold values t:

t	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
tp	873	830	809	809	806	289	165	0
tn	48	90	118	118	121	310	321	321

With a threshold value of $t = 0.50$ for the classification function, for example, the accuracy of our classifier equals $(809 + 118)/1\,202 = 0.77$. We now introduce stratification into our classifier, using a classification function with threshold values $t^- = 0.30$ and $t^+ = 0.80$. With the resulting stratifying classifier, a total of 356 cases from the case set are classified, giving a classification percentage of 29.6%. For the remaining 846 cases the classifier is indecisive, indicating that the dairy farmer should administer broad-spectrum antibiotics to the diseased cows. The distribution of the classified cases over the four sets TP, FP, TN, and FN is shown in Table 2. The stratifying classifier has a standard accuracy of 0.28 and a stratified accuracy of 0.95. Fig. 2 shows the effects of separately varying the two threshold values t^- and t^+, on the two accuracies. The figure clearly shows that an increasing distance between the two threshold values may result in a higher stratified accuracy, which then typically comes at the expense of a decrease in the classifier's classification percentage.

Our earlier observation that the introduction of stratification may both serve to improve and deteriorate classifier performance, is illustrated by the following example. We consider two different classification functions for our stratifying classifier: one function with the threshold values $t^- = 0.40$ and $t^+ = 0.80$ (I), and another one with the threshold values $t^- = 0.40$ and $t^+ = 1.00$ (II). The resulting classifiers both have a standard accuracy smaller than that of the standard classifier. While the standard classifier has an accuracy of $\mathit{accuracy}(0.50) = 0.77$, we find for the two stratifying classifiers:

$$\text{(I):}\quad \mathit{accuracy}(0.40, 0.80) = \frac{289 + 90}{1\,202} = 0.32, \quad \text{and}$$
$$\text{(II):}\quad \mathit{accuracy}(0.40, 1.00) = \frac{0 + 90}{1\,202} = 0.07$$

For classifier (I), we further find for its stratified accuracy:

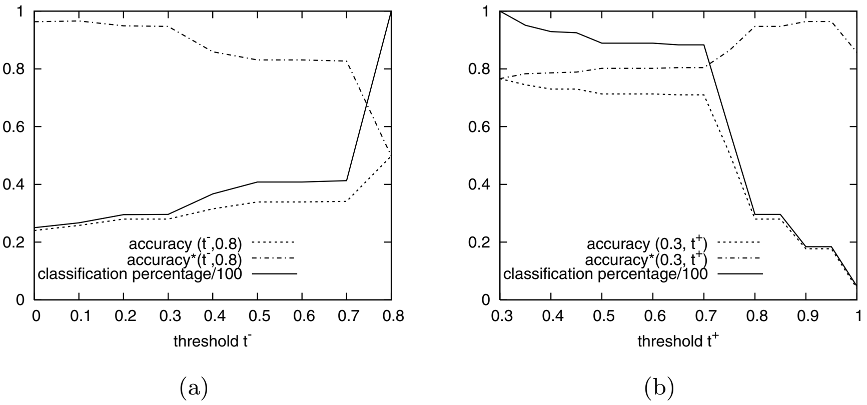
$$\mathit{accuracy}^*(0.40, 0.80) = \frac{289 + 90}{1\,202 - (830 - 289) - (310 - 90)} = 0.86$$

which shows an increase in accuracy over the standard, single-threshold classifier. This increase is explained by the observation that the decrease in the number of correct classifications compared to the standard classifier, is smaller than the relative decrease in the total number of classified cases. For classifier (II), on the other hand, we find that

$$\mathit{accuracy}^*(0.40, 1.00) = \frac{90}{1\,202 - (321 - 90) - (830 - 0)} = 0.64$$

which reveals a decrease in accuracy compared to the standard classifier. This decrease is explained by the observation that the relative decrease in the number of correct classifications now is larger than that in the number of classified cases.

To conclude, by introducing stratification into our example naive Bayesian classifier with threshold values $t^- = 0.10$ and $t^+ = 0.90$, we find a stratified accuracy of 1.00 at a classification percentage of 15.5%. A much poorer choice of threshold values is $t^- = 0.20$ and $t^+ = 1.00$, which results in a stratifying classifier which is decisive on just 4.5% of the cases and has a stratified accuracy of 0.87. The same stratified accuracy is also obtained by setting the threshold value t^+ to the smaller value of $t^+ = 0.75$. We now find a classification percentage of 58.7%!



(a) (b)

Fig. 2. The accuracy, stratified accuracy, and classification percentage for our example classifier and 1 202 cases, as functions of (a) threshold value t^-, with threshold value t^+ fixed at 0.80, and (b) threshold value t^+, with t^- fixed at 0.30

6 Conclusions

For some problems, the single class value returned by a classifier does not necessarily provide a sufficient basis for reliable further decision making. Building upon this observation, we introduced stratifying classifiers as classifiers with the ability to express indecisiveness by not classifying inconclusive cases. These stratifying classifiers are particularly appropriate for applications in which indecisiveness about the class value for a case is a usable result for the decision maker, as in our application domain. Associated with this new type of classifier, we introduced new measures of classification performance and practicability; these measures serve to give insight in the values of the two threshold values to be used with the classification function. In the future we want to extend our concept of stratification to multiway classification and to study the practicability of returning two or more class values for inconclusive cases.

References

1. Blanco, R., Inza, M., Merino, M., Quiroga, J., Larrañaga, P.: Feature Selection in Bayesian Classifiers for the Prognosis of Survival of Cirrhotic Patients Treated with TIPS. Journal of Biomedical Informatics 38, 376–388 (2005)
2. Chapman, W.W., Dowling, J.N., Wagner, M.M.: Classification of Emergency Department Chief Complaints into 7 Syndromes: A Retrospective Analysis of 527,228 Patients. Annals of Emergency Medicine 46, 445–455 (2005)
3. Friedman, N., Goldszmidt, M.: Building Classifiers Using Bayesian Networks. In: Proceedings of the National Conference on Artificial Intelligence, pp. 1277–1284. AAAI Press, Menlo Park (1996)
4. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning 29, 131–163 (1997)
5. Geenen, P.L., Van der Gaag, L.C., Loeffen, W.L.A., Elbers, A.R.W.: Naive Bayesian Classifiers for the Clinical Diagnosis of Classical Swine Fever. In: Proceedings of the 2005 Meeting of the Society for Veterinary Epidemiology and Preventive Medicine, Nairn, Scotland, pp. 169–176 (2005)
6. Kazmierska, J., Malicki, J.: Application of the Naive Bayesian Classifier to Optimize Treatment Decisions. Radiotherapy and Oncology 86, 211–216 (2008)
7. Kuncheva, L.I., Vilas, V.J.D., Rodriguez, J.J.: Diagnosing Scrapie in Sheep: A Classification Experiment. Computers in Biology and Medicine 37, 1194–1202 (2007)
8. Pauker, S.G., Kassirer, J.P.: The Threshold Approach to Clinical Decision Making. New England Journal of Medicine 302, 1109–1117 (1980)
9. Steeneveld, W., Van der Gaag, L.C., Barkema, H.W., Hogeveen, H.: Providing Probability Distributions for the Causal Pathogen of Clinical Mastitis Using Naive Bayesian Networks. Journal of Dairy Science (accepted for publication) (2009)