ELSEVIER

Artificial Intelligence 125 (2001) 209–226

Artificial Intelligence

www.elsevier.com/locate/artint

Research Note

Robust Bayes classifiers

Marco Ramoni [a,*], Paola Sebastiani [b]

[a] *Children's Hospital Informatics Program, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115, USA*

[b] *Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, USA*

Received 12 November 1999; received in revised form 24 October 2000

Abstract

Naive Bayes classifiers provide an efficient and scalable approach to supervised classification problems. When some entries in the training set are missing, methods exist to learn these classifiers under some assumptions about the pattern of missing data. Unfortunately, reliable information about the pattern of missing data may be not readily available and recent experimental results show that the enforcement of an incorrect assumption about the pattern of missing data produces a dramatic decrease in accuracy of the classifier. This paper introduces a *Robust Bayes Classifier* (RBC) able to handle incomplete databases with no assumption about the pattern of missing data. In order to avoid assumptions, the RBC bounds all the possible probability estimates within intervals using a specialized estimation method. These intervals are then used to classify new cases by computing intervals on the posterior probability distributions over the classes given a new case and by ranking the intervals according to some criteria. We provide two scoring methods to rank intervals and a decision theoretic approach to trade off the risk of an erroneous classification and the choice of not classifying unequivocally a case. This decision theoretic approach can also be used to assess the opportunity of adopting assumptions about the pattern of missing data. The proposed approach is evaluated on twenty publicly available databases. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Bayes classifier; Missing data; Probability intervals

1. Introduction

Supervised classification is the task of assigning a *class* label to unclassified *cases* described as a set of *attribute values*. This task is typically performed by first training a classifier on a set of classified cases and then using it to label unclassified cases. The

[*] Corresponding author.
E-mail addresses: marco_ramoni@harvard.edu (M. Ramoni), sebas@math.umass.edu (P. Sebastiani).

PII: S0004-3702(00)00085-0

supervisory component of this classifier resides in the training signal, which provides the classifier with a way to assess a dependency measure between attributes and classes. Naive Bayes classifiers (NBCs) [4,11] have been among the first supervised classification methods and, during the past few years, they have enjoyed a renewed interest and consideration [6]. The training step for a NBC consists of estimating the conditional probability distributions of each attribute given the class from a training data set. Once trained, the NBC classifies a case by computing the posterior probability distribution over the classes via Bayes' Theorem and assigning the case to the class with the highest posterior probability. NBCs assumes that the attributes are conditionally independent given the class and this assumption renders very efficient both training and classification. Unfortunately, when the training set is incomplete, that is, some attribute values or the class itself are reported as unknown, both efficiency and accuracy of the classifier can be lost. Simple solutions to handle missing data are either to ignore the cases including unknown entries or to ascribe these entries to an *ad hoc* dummy state of the respective variables [15]. Both these solutions are known to introduce potentially dangerous biases in the estimates, see [9] for a discussion. In order to overcome this problem, Friedman et al. [6] suggest the use of the EM algorithm [3], gradient descent [20] or, we add, Gibbs sampling [7]. All these methods rely on the assumption that data are *Missing at Random* (MAR) [13], that is, the database is left with enough information to infer the missing entries from the recorded ones. Unfortunately, there is no way to verify that data are actually MAR in a particular database and, when this assumption is violated, these estimation methods suffer of a dramatic decrease in accuracy with the consequence of jeopardizing the performance of the resulting classifier [21].

This paper introduces a new type of NBC, called *Robust Bayes Classifier* (RBC), which does not rely on any assumption about the missing data mechanism. The RBC is based on the *Robust Bayes Estimator* (RBE) [18], an estimator that returns intervals containing all the estimates that could be induced from all the possible completions of an incomplete database. The intuition behind the RBE is that, even with no information about the missing data mechanism, an incomplete data set can still constrain the set of estimates that can be induced from all its possible completions. However, in this situation, the estimator can only bound the posterior probability of the classes. The first contribution of this paper is to provide a specialized closed-form, interval-based estimation procedure for NBCs, which takes full advantage of their conditional independence assumptions. Once trained, these classifiers are used to classify unlabeled cases. Unfortunately, Bayes' Theorem cannot be straightforwardly extended from standard point-valued probabilities to interval-valued probabilities. Nonetheless, the conditional independence assumptions underlying the NBC allows for a closed-form solution for the classification task, too. The second contribution of this paper is a new propagation algorithm to compute posterior probability intervals containing all the class posterior probabilities that could be obtained from the exact computation of all possible completions of the training set. These intervals are then ranked according to a score and a new case is assigned to the class associated with the highest ranked interval. We provide two scoring methods: the first, based on the strong dominance criterion [10], assigns a case to the class whose minimum posterior probability is higher than the maximum posterior probability for all other classes. This criterion preserves the robustness of the classifier but may leave some cases unclassified and hence we provide a weaker criterion to improve the coverage. We also introduce a general decision-theoretic framework to select the most appropriate criterion by trading off accuracy and coverage. As a by-product, this decision-theoretic approach provides a principled way to asses the viability of the MAR assumption for a given training set. We also show that, when the database is complete, the RBC estimates reduce to the standard Bayesian estimates and therefore the RBC subsumes the standard NBC as a special case. This approach is evaluated on twenty publicly available databases.

2. Naive Bayes classifiers

An NBC is better understood if we regard the m attributes and the set of q mutually exclusive and exhaustive classes as discrete stochastic variables. In this way, we can depict a NBC as a Bayesian network [6]—a directed acyclic graph where nodes represent stochastic variables and arcs represent dependency relationships between variables—as shown in Fig. 1. In this network, the root node represents the set C of mutually exclusive and exhaustive classes and each attribute is a *child* node A_i. Each value c_j of the variable C is a class and each attribute A_i bears a set of s_i values $A_i = a_k$. As shorthand, we will denote $C = c_j$ by c_j and $A_i = a_k$ by a_{ik}. The graphical structure of the Bayesian network representing the NBC encodes the assumption that each attribute A_i is conditionally independent of the other attributes given the class. The classifier, therefore, is defined by the marginal probability distribution $\{p(c_j)\}$ of the variable C and by a set of conditional probability distributions $\{p(a_{ik} = c_j)\}$ of each attribute A_i given each class c_j. A consequence of the independence assumption is that all these distributions can be estimated from a training set $\mathcal{D}$, independently from each other, as follows.

Let $n(a_{ik}, c_j)$ be the frequency of cases in the training set $\mathcal{D}$ in which the attribute A_i appears with value a_{ik} and the class is c_j and let $n(c_j)$ be the frequency of cases in the training set with class c_j. When the training set $\mathcal{D}$ is complete, the Bayesian estimates of $p(a_{ik} \mid c_j)$ and $p(c_j)$ are

$$p(a_{ik} \mid c_j) = \frac{\alpha_{ijk} + n(a_{ik}, c_j)}{\sum_h [\alpha_{ijh} + n(a_{ih}, c_j)]} \quad \text{and} \quad p(c_j) = \frac{\alpha_j + n(c_j)}{\sum_l [\alpha_l + n(c_l)]}, \tag{1}$$

respectively. The quantities α_{ijk} and α_j can be regarded as frequencies of pair a_{ik}, c_j and of the class c_j, respectively, in an imaginary sample, representing the prior information about the distributions of the attributes and the classes. The size α of this imaginary sample is called global *prior precision*. Further details are in [17]. Once the classifier has been

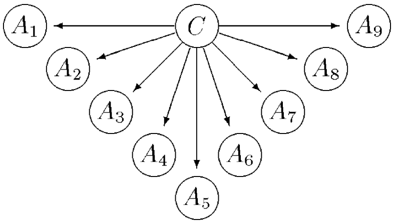

Fig. 1. A Bayesian network representing an NBC with attributes $A_1, \ldots, A_9$ and a set C of classes.

trained, we can use it for the classification of new cases. If we represent a case as a set of attribute values $e = \{a_{1k}, \ldots, a_{mk}\}$, Bayes' theorem yields the posterior probability of a class c_j given e as

$$p(c_j \mid e) = \frac{p(c_j) \prod_{i=1}^{m} p(a_{ik} \mid c_j)}{\sum_{h=1}^{q} p(c_h) \prod_{i=1}^{m} p(a_{ik} \mid c_h)} \tag{2}$$

and the case is assigned to the class with the highest posterior probability.

From the computational point of view, the training of the classifier reduces to summarizing the whole database $\mathcal{D}$ into m contingency tables T_i of dimension $(q \times s_i)$, each cell (j, k) of the table T_i collecting the frequency of the pair (a_{ik}, c_j). In this way

(i) the estimation of the q probability distributions of each attribute A_i conditional on the classes $c_1, \ldots, c_q$ can be done locally using the frequencies $n(a_{ik}, c_j)$ in the table T_i, as the frequencies $n(a_{hk}, c_j)$ in all other tables T_h are irrelevant;

(ii) the estimation of each probability distribution of the attribute A_i conditional on the class c_j can be done independently of the other classes, by using the frequencies $n(a_{ik}, c_j)$ in the row j, and

(iii) the estimation of the marginal distribution of the classes can be done in any one of the tables T_i, by using its row totals $n(c_j)$.

In other words, the estimation procedure can be performed table by table and, within each table, row by row. These properties were termed *global* and *local* parameter independence by [22] and they are the source of the computational efficiency of the training process.

When some entries in the training set $\mathcal{D}$ are missing, both accuracy and efficiency of the NBC are under threat. The reasons for this situation become clear if we regard the incomplete database as the result of a deletion process occurred on a complete (yet unknown) database. The received view on missing data [13] is based on the characterization of the deletion process. According to this approach, data are *Missing Completely at Random* (MCAR), if the probability that an entry is missing is independent of both observed and unobserved values. They are *Missing at Random* (MAR), if this probability is at most a function of the observed values in the database. In all other cases, data are *Informatively Missing*. Under the assumption that data are either MAR or MCAR, the values of the unknown entries can be estimated from the observed ones and the deletion process is called *ignorable*. This property guarantees that the available data are sufficient to train the classifier but, unfortunately, it does not enjoy any longer the properties of global and local parameter independence. Indeed, unknown entries induce three types of incomplete cases:

(i) cases in which the attribute A_i is observed and the class is missing;

(ii) cases in which the class c_j is observed and the value of the attribute A_i is missing;

(iii) cases in which both the value of the attribute A_i and the class are missing.

We denote the frequency of these cases by $n(a_{ik}, ?)$, $n(?, c_j)$ and $n(?, ?)$, respectively. Suppose now we had some estimation method able to compute the estimates in Eq. (1) by assigning a proportion of the frequencies $n(a_{ik}, ?)$, $n(?, c_j)$ and $n(?, ?)$ to the cell (j, k) in each contingency table T_i. As the reconstructed marginal frequency of each class needs to be equal in all tables, the estimation cannot be done locally any longer, and the properties of local and global parameter independence are lost. One exception arises when the class is observed in all cases.

Theorem 1. *Suppose that the class is observed in all cases of the training set $\mathcal{D}$, and that the entries are* MAR. *Then, the estimates in Eq.* (1), *with $n(a_{ik}, c_j)$ being the frequency of fully observed pairs a_{ik}, c_j and $n(c_j)$ being the class frequency, are the exact Bayesian estimates.*

The proof appears in [22]. When also some classes are missing, we can use one of the approximate methods mentioned in the Introduction to compute the estimates in Eq. (1). However, these methods require the deletion process to be ignorable. When data are informatively missing, the available entries are no longer sufficient to train the classifier. Furthermore, there is no way, yet, to check whether the deletion process responsible for the missing data is actually ignorable. These are the motivations behind the introduction of the Robust Bayesian Estimator (RBE) [18] and its application, in this paper, to the development of a robust version of the NBC.

3. Robust estimation

Recall that a NBC is trained by estimating the conditional probability distributions $\{p(a_{ik} \mid c_j)\}$ and $\{p(c_j)\}$ from an database $\mathcal{D}$. This section describes how to perform this task when the database $\mathcal{D}$ is incomplete. We need the following definitions.

Definition 1 (*Consistency*). Let $\mathcal{D}$ be an incomplete data set and let $p(x)$ be a probability that we wish to estimate from $\mathcal{D}$.

(1) A *consistent completion* of $\mathcal{D}$ is any complete data set $\mathcal{D}_c$ from which $\mathcal{D}$ is obtained via some deletion process.
(2) A *consistent estimate* of $p(x)$ is an estimate computed in a consistent completion of $\mathcal{D}$.
(3) A *consistent probability interval* for $p(x)$ is an interval $[p_{inf}(x), p_{sup}(x)]$ containing all consistent estimates. A consistent interval is *non-trivial* if $p_{inf}(x) > 0$ and $p_{sup}(x) < 1$.
(4) A consistent probability interval is *tight* when it is the smallest consistent probability interval $[\underline{p}(x), \overline{p}(x)]$ for $p(x)$.

The difference between a consistent and a tight consistent probability interval is that, in the former, the interval extreme points are lower and upper bounds for the set of consistent estimates, while in the latter, the extreme points are reached in some consistent completion of the database. The rest of this section is devoted to the construction of tight, consistent probability intervals for the quantities $p(a_{ik} \mid c_j)$ and $p(c_j)$ defining an NBC.

In order to estimate the conditional probability $p(a_{ik} \mid c_j)$ from an incomplete training set D, the RBE collects the frequencies $n(a_{ik}, ?)$, $n(?, c_j)$ and $n(?, ?)$ of incomplete cases into the *virtual frequencies* $\overline{n}(a_{ik}, c_j)$ and $\underline{n}(a_{ik}, c_j)$. These frequencies are then used to compute the extreme points of the tight consistent probability interval for $p(a_{ik} \mid c_j)$. The quantity $\overline{n}(a_{ik}, c_j)$ is the maximum number of incomplete cases (A_i, C) that can be completed as (a_{ik}, c_j) and it is given by

$$\overline{n}(a_{ik}, c_j) = n(?, c_j) + n(a_{ik}, ?) + n(?, ?). \tag{3}$$

On the other hand, the virtual frequency $\underline{n}(a_{ik}, c_j)$ is the maximum number of incomplete cases (A_i, C) that can be ascribed to c_j without increasing the frequency $n(a_{ik}, c_j)$ and it is

$$\underline{n}(a_{ik}, c_j) = n(?, c_j) + \sum_{h \neq k} n(a_{ih}, ?) + n(?, ?). \tag{4}$$

The virtual frequencies are used to compute the values $\underline{p}(a_{ik} \mid c_j)$ and $\overline{p}(a_{ik} \mid c_j)$ that are, respectively, the minimum and the maximum estimate of $p(a_{ik} \mid c_j)$ that can be found in the consistent completions of $\mathcal{D}$ and they are

$$\begin{aligned}
\underline{p}(a_{ik} \mid c_j) &= \frac{\alpha_{ijk} + n(a_{ik}, c_j)}{\sum_h [\alpha_{jh} + n(a_{ih}, c_j)] + \underline{n}(a_{ik}, c_j)}, \\
\overline{p}(a_{ik} \mid c_j) &= \frac{\alpha_{ijk} + n(a_{ik}, c_j) + \overline{n}(a_{ik}, c_j)}{\sum_h [\alpha_{ijh} + n(a_{ih}, c_j)] + \overline{n}(a_{ik}, c_j)}.
\end{aligned} \tag{5}$$

It has been shown [18] that the interval $[\underline{p}(a_{ik} \mid c_j), \overline{p}(a_{ik} \mid c_j)]$ is tight and consistent. We now consider the estimation of $p(c_j)$ and note that the virtual frequencies $\underline{n}(c_j)$ and $\overline{n}(c_j)$ are both equal to the number $n(?)$ of cases in $\mathcal{D}$ in which the class is not observed. We obtain tight consistent probability intervals for $p(c_j)$ by setting:

$$\begin{aligned}
\underline{p}(c_j) &= \frac{\alpha_j + n(c_j)}{\sum_l [\alpha_l + n(c_l)] + n(?)}, \\
\overline{p}(c_j) &= \frac{\alpha_j + n(c_j) + n(?)}{\sum_l [\alpha_l + n(c_l)] + n(?)}.
\end{aligned} \tag{6}$$

When the training set is complete, Eqs. (5) and (6) reduce to Eq. (1). Each set given by the maximum probability for the class c_j and the minimum probabilities of the other classes, say $\{\overline{p}(c_j), \underline{p}(c_h), h \neq j\}$ defines a probability distribution

$$\overline{p}(c_j) + \sum_{h \neq j} \underline{p}(c_h) = 1 \quad \text{for all } j, \tag{7}$$

so that the probability intervals $[\underline{p}(c_j), \overline{p}(c_j)]$ are *reachable*, as defined by [2]. By definition, if the probability $p(a_{ik} \mid c_j)$ is at its maximum value $\overline{p}(a_{ik} \mid c_j)$, then the virtual counter $\overline{n}(a_{ik}, c_j)$ absorbs the frequencies $n(a_{ik}, ?)$ and $n(?, ?)$ so that $p(c_j) = \overline{p}(c_j)$ and, for any other class c_h, we have that $p(a_{ik} \mid c_h) < \overline{p}(a_{ik} \mid c_h)$ and $p(c_h) = \underline{p}(c_j)$. Similarly, if the probability $p(a_{ik} \mid c_j)$ is at its minimum value $\underline{p}(a_{ik} \mid c_j)$, then for any other class c_h, we have that $p(a_{ik} \mid c_h) > \underline{p}(a_{ik} \mid c_h)$. However, if the class is always observed, the virtual frequencies $\underline{n}(a_{ik}, c_j)$ and $\overline{n}(a_{ik}, c_j)$ are both equal to $n(?, c_j)$, because $n(a_{ik}, ?) = n(?, ?) = 0$, for all i and k. In this case, the probabilities $p(a_{ik} \mid c_j)$ can vary independently and maxima and minima can be reached at the same time, for different classes c_j.

4. Robust classification

Once trained, the classifier can be used to label unclassified cases. Given a new case, an NBC performs this task in two steps: first it computes the posterior probability of each class given the attribute values, and then it assigns the case to the class with the highest posterior probability. In this section, we first show how to compute posterior probability intervals of each class and then how to rank these intervals to classify new cases.

4.1. Posterior probability intervals

Let $e = \{a_{1k}, \dots, a_{mk}\}$ be attribute values of a case e that we wish to classify. With point-valued probabilities, the expression of the posterior probability of a class c_j, given e, is given in Eq. (2). The next Theorem identifies non-trivial consistent probability intervals for the classes. The result generalizes the solution provided by [18] for Boolean classes. We then show that, when the training set $\mathcal{D}$ reports always the class, these consistent intervals are also tight.

Theorem 2. *Let $\mathcal{D}$ be an incomplete data set. Then, the probability interval $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$ with*

$$p_{sup}(c_j \mid e) = \frac{\overline{p}(c_j)\prod_{i=1}^m \overline{p}(a_{ik} \mid c_j)}{\overline{p}(c_j)\prod_{i=1}^m \overline{p}(a_{ik} \mid c_j) + \sum_{h\neq j} \underline{p}(c_h)\prod_{i=1}^m \underline{p}(a_{ik} \mid c_h)} \tag{8}$$

and

$$p_{inf}(c_j \mid e) = \frac{\underline{p}(c_j)\prod_{i=1}^m \underline{p}(a_{ik} \mid c_j)}{\underline{p}(c_j)\prod_{i=1}^m \underline{p}(a_{ik} \mid c_j) + \max\{f_g,\ g \neq j\}}, \tag{9}$$

where the set $\{f_g,\ g \neq j\}$ contains the $q-1$ quantities

$$\overline{p}(c_g)\prod_{i=1}^m \overline{p}(a_{ik} \mid c_g) + \sum_{l\neq j,g} \underline{p}(c_l)\prod_{i=1}^m \overline{p}(a_{ik} \mid c_l)$$

for $g \neq j = 1, \dots, q$, is non-trivially consistent.

Proof. To prove the theorem, we need to show that the interval $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$ contains all the posterior probabilities $p(c_j \mid e)$ that can be derived from the possible completions of the training set and that $p_{inf}(c_j \mid e) > 0$ and $p_{sup}(c_j \mid e) < 1$. The last two inequalities are a simple consequence of the property $0 < \underline{p}(a_{ik} \mid c_j) \leqslant \overline{p}(a_{ik} \mid c_j) < 1$ and $0 < \underline{p}(c_j) \leqslant \overline{p}(c_j) < 1$ enjoyed by the robust estimates. Hence, it is sufficient to show that, for each j, $p_{inf}(c_j \mid e) \leqslant p(c_j \mid e) \leqslant p_{sup}(c_j \mid e)$, the quantity $p(c_j \mid e)$ being any class posterior probability that can be computed from the consistent completions of the training set $\mathcal{D}$. From Eq. (2), we can write $p(c_j \mid e)$ as

$$f(x_j, y_j) = \frac{y_j x_j}{y_j x_j + \sum_{h\neq j} y_h x_h}, \tag{10}$$

where $y_j = p(c_j)$ and $x_j = \prod_{i=1}^m p(a_{ik} \mid c_j)$. For fixed y_j, the function $f(x_j, y_j)$ is concave, increasing in x_j and decreasing in x_h for $h \neq j$. From standard convex analysis [19], it follows that, if the variables x_j are constrained to vary in a hyper-rectangle, maxima and minima of the function are obtained in the extreme points of the constrained region. In particular, the function $f(x_j, y_j)$ is maximized by maximizing x_j and by minimizing $\sum_{h\neq j} x_h$, and it is minimized by minimizing x_j and by maximizing $\sum_{h\neq j} x_h$. This argument grounds the intuition of the proof: we will find maxima and minima of the function $f(x_j, y_j)$ in a hyper-rectangle containing the region of definition of the variables x_j, for y_j fixed, and these maxima and minima induce upper and lower bounds for the function $f(x_j, y_j)$. We then maximize and minimize these bounds with respect to y_j. The first step is to find this hyper-rectangle.

If the probabilities $p(a_{ik} \mid c_j)$ could vary independently within the intervals $[\underline{p}(a_{ik} \mid c_j), \overline{p}(a_{ik} \mid c_j)]$, then the variables x_j would vary independently in the Cartesian product $\mathcal{C}$ of the intervals

$$[\underline{x}_j\ \overline{x}_j] = \left[\prod_{i=1}^{m} \underline{p}(a_{ik} \mid c_j)\ \prod_{i=1}^{m} \overline{p}(a_{ik} \mid c_j)\right].$$

Thus, setting $x_j = \prod_{i=1}^{m} \overline{p}(a_{ik} \mid c_j)$ and $x_h = \prod_{i=1}^{m} \underline{p}(a_{ik} \mid c_h)$ yields the maximum of the function $f(x_j, y_j)$ in the hyper-rectangle $\mathcal{C}$, for y_j fixed. However, as noted in Section 3, the probabilities $p(a_{ik} \mid c_j)$ cannot vary independently so that the function $f(x_j, y_j)$ is defined in a subset of $\mathcal{C}$ and the quantity

$$f_1(y_j) = \frac{y_j \prod_{i=1}^{m} \overline{p}(a_{ik} \mid c_j)}{y_j \prod_{i=1}^{m} \overline{p}(a_{ik} \mid c_j) + \sum_{h\neq j} y_h \prod_{i=1}^{m} \underline{p}(a_{ik} \mid c_h)}$$

is only an upper bound. Now we maximize the function $f_1(y_j)$ with respect to y_j, subject to the constraint $\sum_j y_j = 1$ that is imposed by the fact that the probability intervals $[\underline{p}(c_j), \overline{p}(c_j)]$ are reachable, as shown in Eq. (7). This maximization yields the upper bound in Eq. (8). The minimum of the function $f(y_j, x_j)$ in the hyper-rectangle $\mathcal{C}$, for y_j fixed, is given by setting $x_j = \prod_{i=1}^{m} \underline{p}(a_{ik} \mid c_j)$ and by maximizing $\sum_{h\neq j} x_h$. The latter quantities is $\sum_{h\neq j} \prod_{i=1}^{m} p(a_{ik} \mid c_h)$ and it is maximized by $\sum_{h\neq j} \prod_{i=1}^{m} \overline{p}(a_{ik} \mid c_h)$, so that

$$f_2(y_j) = \frac{y_j \prod_{i=1}^{m} \underline{p}(a_{ik} \mid c_j)}{y_j \prod_{i=1}^{m} \underline{p}(a_{ik} \mid c_j) + \sum_{h\neq j} y_h \prod_{i=1}^{m} \overline{p}(a_{ik} \mid c_h)}$$

is a lower bound for $f(y_j, x_j)$. We minimize the function $f_2(y_j)$ with respect to y_j, and the minimum is given by setting $y_j = \underline{p}(c_j)$ and by maximizing the function $f_3 = \sum_{h\neq j} p(c_h) \prod_{i=1}^{m} \overline{p}(a_{ik} \mid c_h)$, subject to the constraint $\overline{p}(c_g) + \sum_l \underline{p}(c_l) = 1 - \underline{p}(c_j)$. The function f_3 is linear in the probabilities $p(c_h)$ and hence its maximum is found by evaluating it in the extreme points of the constrained region, from which lower bound in Eq. (9) follows. □

When the training set is complete, the RBE intervals reduce to the point estimates given in Section 2, and the quantities in Eqs. (8) and (9) become identical to the posterior probability in Eq. (2). The interval $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$ is consistent, as it contains all posterior probabilities $p(c_j \mid e)$ that we would obtain by applying Bayes' Theorem to all consistent estimates $p(a_{ik} \mid c_j)$ and $p(c_j)$. The proof of Theorem 2 uses the constraints imposed by the class probability intervals $[\underline{p}(c_j), \overline{p}(c_j)]$ and mixes maximum and minimum probabilities coherently. However, the probabilities $p(a_{ik} \mid c_j)$, for varying j, are minimized and maximized independently and, in general, this may produce loose bounds. Still, when the class is observed in all cases, we can prove the tightness of these bounds.

Theorem 3. *If the class c_j is reported for every case e, the probability intervals defined by*

$$\overline{p}(c_j \mid e) = \frac{p(c_j)\prod_{i=1}^m \overline{p}(a_{ik} \mid c_j)}{p(c_j)\prod_{i=1}^m \overline{p}(a_{ik} \mid c_j) + \sum_{l \neq j} p(c_l)\prod_{i=1}^m \underline{p}(a_{ik} \mid c_l)} \tag{11}$$

and by

$$\underline{p}(c_j \mid e) = \frac{p(c_j)\prod_{i=1}^m \underline{p}(a_{ik} \mid c_j)}{p(c_j)\prod_{i=1}^m \underline{p}(a_{ik} \mid c_j) + \sum_{h \neq j} p(c_h)\prod_{i=1}^m \overline{p}(a_{ik} \mid c_h)} \tag{12}$$

are tight and consistent.

Proof. If the class is always observed, we have that $\underline{p}(c_j) = \overline{p}(c_j) = p(c_j)$ and, as noted in Section 3, the probabilities $p(a_{ik} \mid c_j)$ can vary independently as j varies, so that the upper and lower bounds in Eqs. (8) and (9) are the maximum and minimum values of the function in Eq. (10). Note further that Eqs. (8) and (9) reduce to Eqs. (11) and (12). □

4.2. Ranking intervals

The previous section has shown how to compute consistent posterior probability intervals for the classes given a set e of attribute values. We can now use these intervals to assign a case to a class, by associating each interval to a score and using the following classification rule.

Definition 2 (*Interval-based classification rule*). Let e be a set of attribute values and let $s(c_j \mid e)$ be scores associated with the probability intervals $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$. Each case with attribute values e is assigned to the class associated with the largest score.

The interval-based classification rule is based on the intuition that the score $s(c_j \mid e)$ associated with the probability intervals $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$ is a "meaningful" summary of the global information contained in the probability intervals. However, this is not the unique requirement. Since the standard NBC classifies cases on the basis of the posterior probabilities of the classes given the attribute values, we require that the set of scores associated with the probability intervals $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$ defines a probability distribution, and hence

$$s(c_j \mid e) \geqslant 0 \quad \text{for all } j, \qquad \sum_j s(c_j \mid e) = 1. \tag{13}$$

Theorem 2 ensures that the interval $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$ contains all possible conditional probabilities $p(c_j \mid e)$ that can be computed from the consistent completions of the training set $\mathcal{D}$, and the variability within the intervals is due to the uncertainty about the missing data mechanism. A conservative score derived from the *strong dominance* criterion [10] provides a classification rule that does not require any assumption about the missing data mechanism.

Definition 3 (*Strong dominance score*). Given a set of q consistent posterior probability intervals $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$, we define the *strong dominance score* as:

$$s_d(c_j \mid e) = \begin{cases} 1 & \text{if } p_{inf}(c_j \mid e) > p_{sup}(c_h \mid e) \text{ for all } h \neq j, \\ 0 & \text{if } p_{inf}(c_j \mid e) \leqslant p_{sup}(c_h \mid e) \text{ for some } h \neq j. \end{cases}$$

The interval-based classification rule induced by the strong dominance score classifies a new case as c_j if and only if the probability $p_{inf}(c_j \mid e)$ is larger than the probability $p_{sup}(c_h \mid e)$, for any $h \neq j$. Strong dominance is a safe criterion since it returns the classification that we would obtain from all consistent completions of the training set $\mathcal{D}$. However, when the probability intervals are overlapping, the strong dominance score is not defined and we face a situation of undecidability. Moreover, the strong dominance score is too conservative because the condition $p_{inf}(c_j \mid e) > p_{sup}(c_h \mid e)$, for all $h \neq j$, is sufficient to yield the classification we would obtain, the complete training set being known, but it is not necessary. In order to increase the coverage of the classifier, we can weaken this criterion by making the minimal assumption that all missing data mechanisms are equally possible, thus making all values within the intervals $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$ equally likely. In this way, we summarize the interval into an average point by defining the score

$$\begin{aligned} s_u(c_j \mid e) &= p_{sup}(c_j \mid e) - k\big(p_{sup}(c_j \mid e) - p_{inf}(c_j \mid e)\big) \\ &= (1-k)p_{sup}(c_j \mid e) + kp_{inf}(c_j \mid e), \end{aligned}$$

where k is chosen so that the scores $\{s_u(c_j \mid e)\}$ satisfy the properties of Eq. (13). Hence,

$$k = \frac{1 - \sum_h p_{inf}(c_h \mid e)}{\sum_h (p_{sup}(c_h \mid e) - p_{inf}(c_h \mid e))}.$$

A consequence of the consistency of the probability intervals $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$ is that the extreme probabilities $p_{inf}(c_j \mid e)$ and $p_{sup}(c_j \mid e)$ and the probability $p(c_j \mid e)$ that we would compute, from a complete training set $\mathcal{D}$, are in the relationship $p_{inf}(c_j \mid e) \leqslant p(c_j \mid e) \leqslant p_{sup}(c_j \mid e)$. It follows that $\sum_j p_{inf}(c_j \mid e) \leqslant 1 \leqslant \sum_j p_{sup}(c_j \mid e)$ and, hence, that the quantity k is in the open interval $(0, 1)$. This last finding guarantees that the score $s_u(c_j \mid e)$ is in the interior of the interval $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$ and, consequently, that it cannot produce a classification rule that does not correspond to any *E-admissible* classification rule compatible with the intervals $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$ [12]. Note that the Hurwicz's Optimism–Pessimism criterion—the usual solution for these circumstances [14,16]—does not guarantee this property. As the score $s_u(c_j \mid e)$ always leads to a decision, we term it a *complete-admissible score*.

Definition 4 (*Complete-admissible score*). Given a set of q consistent posterior probability intervals $[p_{inf}(c_j \mid e), p_{sup}(c_j \mid e)]$, we define the quantity

$$s_u(c_j \mid e) = p_{sup}(c_j \mid e) - \frac{(p_{sup}(c_j \mid e) - p_{inf}(c_j \mid e))(1 - \sum_h p_{inf}(c_h \mid e))}{\sum_h (p_{sup}(c_h \mid e) - p_{inf}(c_h \mid e))}$$

a *complete-admissible score*.

It is worth noting that the classification based on the complete-admissible score subsumes the one based on the strong dominance score because if $p_{inf}(c_j \mid e) > p_{sup}(c_h \mid e)$, for all $h \neq j$, then $s_u(c_j \mid e) > s_u(c_h \mid e)$ for all $h \neq j$. When the condition to apply the strong dominance score does not hold, the complete-admissible score lets us classify the cases left unclassified by the strong dominance score. This strategy may result in an increased classification coverage at the price of a lower accuracy.

4.3. Which score?

Both the strong dominance and the complete-admissible score provide a sensible basis for robust classification. Strong dominance is safe at the price of leaving cases unclassified while the complete-admissible score increases the classification coverage by loosing robustness. The choice of an interval-scoring method depends on the features of the problem at hand and, in this section, we provide a principled way to choose the best interval-based classification strategy.

A classification system is typically evaluated on the basis of its classification accuracy θ and its coverage γ. The former is the probability of correctly classifying a case while the latter is the probability of classifying one case. Let θ_d and γ_d be respectively the accuracy and the coverage of an RBC with the strong dominance score (RBC$_d$). The accuracy θ_d is independent of the missing data mechanism. Similarly, let θ_u be the accuracy of the RBC with the complete-admissible score, say RBC$_u$. The accuracy θ_u of the RBC$_u$ is given by two components. The first component is the probability of correctly classifying one case when we can use the strong dominance score and, hence, it is weighted by the coverage γ_d. The second component is the probability of correctly classifying one case when we cannot use the strong dominance score and, therefore, it is weighted by $1 - \gamma_d$. Thus,

$$\theta_u = \theta_d \gamma_d + \theta_{ul}(1 - \gamma_d), \tag{14}$$

where θ_{ul} is the classification accuracy of the RBC$_u$ on the cases left unclassified by the RBC$_d$ and we term it *residual accuracy*. Residual accuracy provides a measure of the gain/loss of classification accuracy achieved by the RBC$_u$ when one relaxes the strong dominance criterion to increase coverage.

The decomposition in Eq. (14) provides a first basis to choose the scoring method. For example, a simple rule could be to adopt the complete-admissible score if θ_{ul} is greater than $1/q$, so that the cases left unclassified by the strong dominance score are classified by the complete-admissible score better than at random. The intuition behind this rule is that accuracy is more valuable than coverage and, hence, we would not prefer a method that classifies randomly just because it always classifies a case. The rationale is that we expect the consequence of a wrong classification to be worse than the inability to classify one case. This argument can be used formally to choose between the strong dominance or the complete-admissible score by introducing mis-classification costs and costs incurred for the inability to classify one case. Suppose that the cost incurred for not being able to classify a case with attribute values e is a quantity C_i, while the cost for a wrong classification is C_w. Since the former event occurs with probability $1 - \gamma_d$ and the latter occurs with probability $(1 - \theta_d)\gamma_d$, the expected cost incurred on using the RBC$_d$ is

$$C(\text{RBC}_d) = C_w(1 - \theta_d)\gamma_d + C_i(1 - \gamma_d)$$

if correct classification yields no cost. On the other hand, the expected cost incurred by an RBC_u achieving 100% coverage with accuracy θ_u is

$$C(\text{RBC}_u) = C_w(1 - \theta_u).$$

In order to minimize the cost, RBC_d is to be preferred to RBC_u when $C(\text{RBC}_d) \leqslant C(\text{RBC}_u)$. This is true if and only if $\theta_u - \theta_d\gamma_d = \theta_{ul}(1 - \gamma_d) \leqslant (1 - \gamma_d)(1 - C_i/C_w)$ and it yields the decision rule given in the next theorem.

Theorem 4. *Let C_i and C_w denote respectively the cost of a wrong classification and the cost of not being able to classify a case. The interval based classification rule which uses the strong dominance score yields minimum expected cost if and only if*:

$$\theta_{ul} \leqslant (1 - C_i/C_w)$$

where θ_{ul} is the accuracy of the RBC_u on the cases left unclassified by the RBC_d.

For example, if $C_i = C_w$, the best decision is to choose the RBC_d whenever $\theta_{ul} \geqslant 0$. Compared to the simpler rule described above, the decision now takes into account the trade-off between accuracy and coverage. In practical applications, the quantities θ_d, θ_u and γ_d can be estimated from the available data using cross validation, as shown in the next section. Suppose now the quantity θ_a is the accuracy of any other NBC_a trained on an incomplete data set under some assumption about the missing data mechanism. For example, θ_a could be the accuracy of an NBC trained on an incomplete data set under the assumption that data are MAR. We can use the same decision rule to help one decide whether the RBC with the strong dominance or the complete-admissible score yields minimum expected costs. As a by-product, the decision rule can be interpreted as an evaluation of the consequences of enforcing the MAR assumption. The comparison between the accuracy measures θ_a and θ_u is cost-independent, as we compare $C(\text{RBC}_u) = C_w(1 - \theta_u)$ and $C(\text{NBC}_a) = C_w(1 - \theta_a)$ and the minimum expected cost is achieved by the system having the highest accuracy. If we now compare the expected costs of the RBC_d and the NBC_a, and apply the decision rule in Theorem 4, we have that the NBC_a is to be preferred to the RBC_d whenever $\theta_{ul} \geqslant (1 - C_i/C_w)$ and the quantity $(1 - \gamma_d)[\theta_{ul} - (1 - C_i/C_w)]$ is the cost incurred in enforcing the assumption about the missing data mechanism. This solution can be easily extended to cases in which misclassification costs vary with the classes.

5. Evaluation

This section reports the results of an experimental evaluation of the RBC on twenty incomplete data sets. The aim of the evaluation is to compare the performance of the RBC with that of two NBCs, using the most common solutions to handle missing data [6, 15]: remove the missing entries (NBC_m) and assign the missing entries to a dummy state (NBC_*). Since all data sets always report the classes for every case, by Theorem 1 NBC_m is a faithful implementation of the MAR assumption. NBC_*, on the other hand, assumes some knowledge on the missing data mechanism since the missing data are treated as a category “other”, not reported in the observed data.

5.1. Materials and methods

The experimental evaluation was conducted on the twenty databases reported in Table 1, available from the UCI Machine Learning repository [1]. The database KDD99 consists of 4704 cases on 31 variables selected from the database used for the 1999 KDD cup. These databases offer a variety of different data types: all attributes of the database Voting record (Vote) are binary, all attributes in Breast Cancer Wisconsin (B.Cancer) and Lung Cancer (L.Cancer) and Bridge are nominal, all attributes in Hepatitis and Mushrooms are discrete, while for example Annealing, Credit, Cylinder, Horse Colic, and Sick offer a good mixture of continuous, discrete and nominal attributes. The size of these databases ranges from the 32 case on 56 attributes of Lung Cancer to the 48842 cases on 14 variables in Mushrooms. Continuous attributes were discretized by dividing the observed range into four bins with the same proportion of entries.

Following current practice [8], we compared the accuracy of classifiers by running, on each data set D, 5 replicates of a 5-fold cross validation experiment. On each database, we ran four tests: one training the NBC on a database with the missing entries removed (NBC$_m$), one assigning the missing entries to a dummy state (NBC$_*$), one using the strong dominance score (RBC$_d$) and one using the complete-admissible score (RBC$_u$). In all cases, we computed the estimates using a uniformly distributed global prior precision $\alpha = 1$. For each test, we report two values: *accuracy*—estimated as the average number of cases that were correctly classified in the test sets—and *coverage*—given by the ratio between the number of cases classified and the total number of cases in the data set. The 95% confidence limits are based on a Normal approximation of a proportion estimator [8].

5.2. Results and discussion

Table 1 reports the results. The accuracy of RBC$_d$ is overall the highest, with a gain ranging from 0.02% (Breast Cancer), in which there are only 6 missing entries in a data set of 699 cases, to 16.77% (Horse Colic), in which data are heavily missing. Except for Audiology, Breast Cancer, and Lung Cancer, the accuracy gain of RBC$_d$ is statistically significant in all cases, as shown by the non overlapping confidence intervals. This gain of accuracy is counter-balanced by a loss of coverage that can be as small as 6.51% in Horse Colic. The complete-admissible score increases the coverage to 100% at the price of reducing the accuracy, so that in Audiology, Breast Cancer, and Credit it is out-performed by the standard NBC. However, the difference in accuracy is within the sampling variability, as the associated confidence limits are roughly the same, and probably data are MAR in these data sets. On the other hand, the accuracy gain of RBC$_u$ over NBC$_m$ and NBC$_*$ is significant in all the other data sets, and reaches 10.19% in the Annealing data set, thus confirming the potential danger of wrongfully enforcing the MAR assumption.

As noted in Section 4.3, the strong dominance score partitions the data into two parts. One part comprises the cases on which there is no classification ambiguity and the accuracy is only model-dependent. The remaining part comprises those cases that cannot be classified without some assumption about the missing data mechanism. Using the notation of Section 4.3, the accuracy on these cases of the other systems achieving 100% coverage is given by the quantity

Table 1
Accuracy of NBC$_m$, NBC$_*$, RBC$_d$, RBC$_u$. Maximum values are reported in boldface

	Database	NBC$_m$	NBC$_*$	RBC$_d$		RBC$_u$
		Accuracy	Accuracy	Accuracy	Coverage	Accuracy
1	Adult	81.74 ± 0.23	81.22 ± 0.22	**86.51 ± 0.21**	81.72 ± 0.18	82.50 ± 0.20
2	Annealing	86.54 ± 2.88	80.88 ± 3.32	**97.53 ± 1.51**	49.12 ± 4.87	96.73 ± 1.24
3	Arythmia	64.40 ± 2.25	61.05 ± 2.76	**76.09 ± 3.25**	39.82 ± 2.30	66.19 ± 2.33
4	Audiology	58.34 ± 3.49	55.50 ± 3.51	**63.41 ± 5.32**	34.78 ± 3.48	55.50 ± 3.51
5	Automobile	60.48 ± 3.41	58.05 ± 3.45	**68.49 ± 3.84**	71.22 ± 3.16	61.96 ± 3.39
6	B.Cancer	97.42 ± 0.66	97.42 ± 0.66	**97.49 ± 0.67**	99.65 ± 5.23	97.23 ± 0.67
7	Bridge	67.62 ± 4.57	64.76 ± 4.66	**80.00 ± 4.78**	66.67 ± 4.60	69.52 ± 4.49
8	Credit	84.88 ± 1.30	84.88 ± 1.30	**87.48 ± 1.72**	95.40 ± 5.21	84.70 ± 1.31
9	Cylinder	73.70 ± 3.71	73.00 ± 3.74	**91.71 ± 4.14**	31.30 ± 6.97	74.26 ± 0.67
10	Echocardiogram	87.23 ± 2.94	88.54 ± 2.78	**93.58 ± 2.35**	83.21 ± 3.27	88.54 ± 2.78
11	Heart-C	54.13 ± 2.86	53.80 ± 2.86	**58.97 ± 2.89**	95.71 ± 1.16	58.07 ± 2.83
12	Heart-H	83.33 ± 2.00	81.29 ± 2.27	**85.88 ± 2.11**	86.73 ± 1.98	83.67 ± 2.11
13	Heart-S	38.29 ± 4.38	36.59 ± 4.34	**47.37 ± 11.45**	15.44 ± 3.26	42.28 ± 4.45
14	Hepatitis	85.03 ± 2.09	85.16 ± 2.08	**90.50 ± 2.84**	76.45 ± 9.73	85.55 ± 2.08
15	Horse Colic	75.79 ± 1.62	75.79 ± 1.63	**92.56 ± 0.59**	6.51 ± 2.05	77.73 ± 1.61
16	KDD99	84.68 ± 0.52	84.80 ± 0.50	**89.22 ± 0.67**	45.42 ± 0.70	84.85 ± 0.52
17	L.Cancer	43.75 ± 17.88	43.75 ± 17.88	**46.67 ± 17.85**	93.75 ± 8.66	43.75 ± 17.88
18	Mushrooms	98.53 ± 0.12	98.40 ± 0.12	**99.04 ± 0.15**	98.88 ± 1.53	98.70 ± 0.11
19	Sick	91.60 ± 0.51	90.87 ± 0.53	**97.53 ± 0.34**	86.30 ± 2.29	92.46 ± 0.49
20	Vote	90.02 ± 1.05	90.21 ± 1.04	**92.05 ± 1.75**	94.94 ± 6.47	90.21 ± 1.04

$$\theta_{al} = \frac{\hat{\theta}_a - \hat{\theta}_d \hat{\gamma}_d}{1 - \hat{\gamma}_d},$$

where $\hat{\theta}_a$ is the (estimated) accuracy of NBC$_m$, NBC$_*$ or RBC$_u$, while $\hat{\theta}_d$ and $\hat{\gamma}_d$ are the estimated accuracy and coverage of RBC$_d$. Table 2 reports these accuracy values for NBC$_m$, NBC$_*$, and RBC$_u$ in the data sets used in this experiment. The sixth column reports the maximum cost ratio C_i/C_w to make RBC$_d$ the best classification system in terms of minimum expected costs and, for reference, the last two columns note the proportion of cases left unclassified by RBC$_d$ and size of the database. If the cost ratio is higher than the reported value, then the best system is the one with the highest accuracy θ_{al}, and it is reported in bold face in the table.

In the data sets B.Cancer and Credit, RBC$_d$ is the best choice if $C_w \geqslant 4.44C_i$ and $C_w \geqslant 1.45C_i$, respectively. If these conditions are not satisfied, then NBC$_m$ or, equivalently, NBC$_*$, are the best systems. In the B.Cancer data set, the complete-admissible score

Table 2

Residual accuracy of NBC$_m$, NBC$_*$ and RBC$_u$. The sixth column reports the maximum value on the cost ratio C_i/C_w that makes RBC$_d$ the classification system with minimum expected cost. If the cost ratio C_i/C_w is superior to this value, then the system corresponding to the bold-faced accuracy is the best choice. The last two columns report the percentage of cases left unclassified by RBC$_d$ and the database size

	Database	θ_{ml}	θ_{*l}	θ_{ul}	C_i/C_w	$(1-\gamma_d)100$	Size
1	Adult	0.6040	0.5757	**0.6457**	0.3543	18.28	48842
2	Annealing	0.7593	0.6481	**0.9596**	0.0404	50.88	798
3	Arythmia	0.5666	0.5100	**0.5964**	0.4036	60.18	452
4	Audiology	**0.5564**	0.5128	0.5128	0.4436	65.22	200
5	Automobile	0.4066	0.3221	**0.4580**	0.5420	28.78	205
6	B.Cancer	**0.7749**	**0.7749**	0.2320	0.2251	0.35	699
7	Bridge	0.4286	0.3428	**0.4856**	0.5144	33.33	105
8	Credit	**0.3096**	**0.3096**	0.2705	0.6904	4.60	598
9	Cylinder	0.6448	0.6549	**0.6631**	0.3369	68.70	512
10	Echocardiogram	0.5576	**0.6356**	**0.6356**	0.3644	16.79	131
11	Heart-C	0.0000	0.0000	**0.3799**	0.6201	4.29	303
12	Heart-H	0.6667	0.5129	**0.6923**	0.3077	13.27	294
13	Heart-S	0.3663	0.3462	**0.4135**	0.5865	84.56	123
14	Hepatitis	0.6782	0.6727	**0.6948**	0.3052	23.55	155
15	Horse Colic	0.7462	0.7462	**0.7670**	0.2330	93.49	368
16	L.Cancer	0.0000	0.0000	0.0000	1.0005	6.25	32
17	KDD99	0.8090	0.8112	**0.8121**	0.1878	54.58	4704
18	Mushrooms	0.4190	0.5350	**0.6868**	0.3132	1.22	8124
19	Sick	0.4892	0.5424	**0.6052**	0.3948	15.70	2800
20	Vote	**0.5569**	0.5193	**0.5569**	0.4431	5.06	435

performs very poorly on the cases left unclassified by the strong dominance score, while the enforcement of the MAR assumption allows the standard NBC to exploit the information provided by the available data and reaches an accuracy of 77.49%. This data set has, however, only 6 cases with missing entries. In the Credit data set, NBC$_m$, NBC$_*$ and RBC$_u$ achieve an accuracy lower than 50% so that, if the the mis-classification cost is lower than $1.45C_i$, a random assignment of the cases left unclassified by RBC$_d$ is preferable. In Audiology, RBC$_d$ is the minimum expected cost system if $C_w \geqslant 2.25C_i$. When the condition on the cost ratio is not satisfied, that NBC$_m$ is the classification system to adopt. In the data set L.Cancer, the accuracy θ_{al} is null for all systems, and hence the choice of RBC$_d$ is never under discussion. This is also confirmed by the fact that the maximum value on the cost ratio C_i/C_w, which makes RBC$_d$ the system with minimum expected cost, is 1.005. Hence, RBC$_d$ is the best whenever $C_w \geqslant 0.995C_i$. As this data set is of medical nature, one can hardly imagine a situation in which not making an automatic analysis is less costly than making the wrong one. In the remaining data sets, RBC$_u$ is always the second best choice, if the cost ratio C_i/C_w is superior to the value reported in the last column of the table. In the data set Annealing, for example, if the cost for not classifying a case is smaller than 25 times—given by $1/0.0404$—the cost for a wrong classification, RBC$_u$ is the best choice and achieves an accuracy 0.9596 on the cases left unclassified by RBC$_d$. This is a gain of about 20% compared to NBC$_m$. Again this result confirms that the MAR assumption on this data set has a negative effect on the accuracy. A similar result is shown in the Mushroom data set, in which either assigning the missing entries to a dummy value or enforcing the MAR assumption yields essentially a random classification of the cases left unclassified by RBC$_d$, while the use of the complete-admissible score rises the residual accuracy to 68.68%. The Sick data set reports a similar result, while the accuracy of RBC$_u$ is only slightly superior to the NBC$_*$ in the data sets Automobile, Cylinder, Hepatitis, and Horse Colic, and is none in the Vote data set.

These results suggest that the RBC based on the strong dominance criterion delivers the highest accuracy, at risk of a decreased coverage. The use of the complete-admissible score improves coverage by decreasing accuracy, and it appears to achieve better results than standard solutions, except when the proportion of missing data is small. However, there does not seem to be a consistently superior classifier and the solution to adopt needs to take into account features of the data at hand. Nonetheless, our decision theoretic approach provides a principled way to choose the most appropriate solution.

6. Conclusions

This paper introduced the RBC: a generalization of the standard NBC which is robust with respect to the missing data mechanism. The RBC performs the training step from an incomplete data set resulting in a classification system quantified by tight consistent probability intervals. Then, the RBC classifies new cases by reasoning with probability intervals. We provided an interval propagation algorithm to identify bounds on the set of the classes posterior probabilities that can be computed from all possible completions of the data, and two scoring methods for interval-based classification. The choice of the scoring methods that best suits the problem at hand is based on a decision-theoretic rule that takes into account costs of mis-classification and cost incurred for not being able to classify a case, and can be extended to make a cost-analysis of the implications of the MAR assumption on the classification accuracy. The experimental evaluations showed the gain of accuracy that can be achieved by the RBC compared to standard solutions. However, the results also showed that there is no uniformly better classification strategy when the data are incomplete, and we expect that the principled way to choose the solution that best suits the problem at hand will become common practice in real applications.

Although the robust solution that we presented in this paper is limited to the NBC, it is straightforward to extend it to tree-structured classification systems in which attributes are binary and the classification problem is to choose between two classes. This can be done by training the classifier with the RBE and by computing bounds on the posterior probability of the classes using the 2U algorithm of [5]. The classification can be done by choosing one of the interval-based classification rules that we presented here, in the same principled way. The extension to more general classification models is the real challenge and essentially requires the development of interval propagation algorithms that returns not too loose bounds on the class posterior probability. The methods described in this paper have been implemented in the computer program [1] distributed, to date, in over 2000 copies.

Acknowledgements

Authors are very grateful to Paul Snow for his invaluable contribution to the development of the complete-admissible score, and to the anonymous referees and the editor for their helpful suggestions.

References

[1] C. Blake, E. Keogh, C.J. Merz, UCI Repository of machine learning databases, University of California, Irvine, Department of Information and Computer Sciences, 1998.
[2] L. Campos, J. Huete, S. Moral, Probability intervals: A tool for uncertain reasoning, Internat. J. Uncertainty, Fuzziness and Knowledge-Based Systems 2 (1994) 167–196.
[3] A.P. Dempster, D. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. Royal Statist. Soc. Ser. B 39 (1977) 1–38.
[4] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
[5] E. Fagiuoli, M. Zaffalon, 2U: An exact interval propagation algorithm for polytrees with binary variables, Artificial Intelligence 106 (1998) 77–108.
[6] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Machine Learning 29 (1997) 131–163.
[7] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 721–741.
[8] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proc. IJCAI-95, Montreal, Quebec, Morgan Kaufmann, San Francisco, CA, 1995, pp. 1146–1151.
[9] R. Kohavi, B. Becker, D. Sommerfield, Improving simple Bayes, in: M. van Someren, G. Widmer (Eds.), Poster Papers of the ECML-97, Charles University, Prague, 1997, pp. 78–87.
[10] H.E. Kyburg, Rational belief, Behavioral and Brain Sciences 6 (1983) 231–273.
[11] P. Langley, W. Iba, K. Thompson, An analysis of Bayesian classifiers, in: Proc. AAAI-92, San Jose, CA, AAAI Press, Menlo Park, CA, 1992, pp. 223–228.
[12] I. Levi, On indeterminate probabilities, J. Philos. 71 (1974) 391–418.
[13] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, Wiley, New York, 1987.
[14] M. Pittarelli, An algebra for probabilistic databases, IEEE Transactions on Knowledge and Data Engineering 6 (2) (1994) 293–303.
[15] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, CA, 1993.
[16] M. Ramoni, Ignorant influence diagrams, in: Proc. IJCAI-95, Montreal, Quebec, Morgan Kaufmann, San Francisco, CA, 1995, pp. 1808–1814.
[17] M. Ramoni, P. Sebastiani, Bayesian methods, in: M. Berthold, D.J. Hand (Eds.), Intelligent Data Analysis. An Introduction, Springer, New York, 1999, pp. 129–166.
[18] M. Ramoni, P. Sebastiani, Robust learning with missing data, Machine Learning (2000), to appear.
[19] R.T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, 1970.

[1] Available from Bayesware Limited (www.bayesware.com).

[20] S. Russell, J. Binder, D. Koller, K. Kanazawa, Local learning in probabilistic networks with hidden variables, in: Proc. IJCAI-95, Montreal, Quebec, Morgan Kaufmann, San Francisco, CA, 1995, pp. 1146–1151.

[21] D.J. Spiegelhalter, R.G. Cowell, Learning in probabilistic expert systems, in: J. Bernardo, J. Berger, A.P. Dawid, A.F.M. Smith (Eds.), Bayesian Statistics 4, Oxford University Press, Oxford, UK, 1992, pp. 447–466.

[22] D.J. Spiegelhalter, S.L. Lauritzen, Sequential updating of conditional probabilities on directed graphical structures, Networks 20 (1990) 157–224.