International Statistical Review (2001), **69**, 3, 385–398, Printed in The Netherlands

Idiot's Bayes—Not So Stupid After All?

David J. Hand[1] and Keming Yu[2]

[1]*Department of Mathematics, Imperial College, London, UK. E-mail: d.j.hand@ic.ac.uk*
[2]*University of Plymouth, UK*

Summary

Folklore has it that a very simple supervised classification rule, based on the typically false assumption that the predictor variables are independent, can be highly effective, and often more effective than sophisticated rules. We examine the evidence for this, both empirical, as observed in real data applications, and theoretical, summarising explanations for why this simple rule might be effective.

Key words: Supervised classification; Independence model; Naïve Bayes; Simple Bayes; Diagnosis.

1 Introduction

Given the measurement vectors for a sample of objects, along with the labels of classes to which those objects belong, supervised classification methods seek to construct *classification rules* which will permit new objects to be classified into classes purely on the basis of their measurement vectors. The word 'supervised' refers to the fact that the elements of the original sample (the 'design' or 'training' sample) have known class memberships—these were determined by a 'supervisor' or 'teacher'.

There are two broad paradigms through which such rules may be constructed. These have been termed the *diagnostic paradigm* and the *sampling paradigm* (Dawid, 1976). Denote the vector of measurements of an object by $\mathbf{x} = (x_1, \ldots, x_d)$ and its class by $i\,(i = 1, \ldots, c)$. Then the diagnostic paradigm focuses attention directly on the probabilities $P(i|\mathbf{x})$, the probability that an object belongs to class i given that it has measurement vector $\mathbf{x}$, and estimates these probabilities from the design sample. In contrast, the sampling paradigm uses the design sample to obtain estimates $\hat{P}(\mathbf{x}|i)$ of $P(\mathbf{x}|i)$, the distribution of $\mathbf{x}$ for objects from class i, and estimates $\hat{P}(i)$ of $P(i), i = 1, \ldots, c$, the probabilities that a member of class i will occur, and then combines these via Bayes theorem to yield estimates of the probabilities $P(i|\mathbf{x})$, $\hat{P}(i|\mathbf{x}) \propto \hat{P}(\mathbf{x}|i)\hat{P}(i)$.

Perhaps the most familiar examples of these two paradigms are logistic discriminant analysis and linear discriminant analysis, respectively corresponding to the diagnostic and sampling paradigms (although the original derivation of linear discriminant analysis was in terms of separability between the classes—that is, the diagnostic paradigm). Not surprisingly, if any assumptions made about the individual class distributions of $\mathbf{x}$ in the sampling paradigm are correct, then the extra information this implies one has about the problem means that more accurate estimates result (Efron, 1975). On the other hand, if these assumptions are incorrect, then they can lead to inaccurate estimators. We expand on this below.

There is also a third possibility: to estimate the overall joint probability, $P(i, \mathbf{x})$, from which the conditional probabilities $P(i|\mathbf{x})$ can be obtained easily. There are some subtleties associated with this approach, pointed out and illustrated by, for example, Friedman *et al.* (1997) and Kontkanen *et al.* (1999). In particular, a global criterion of goodness of fit of a model for $P(i, \mathbf{x})$ will not necessarily

be optimal for a particular purpose, which is predictive classification in our case, especially if the accuracy is assessed by a different loss function. For example, likelihood is a common loss function for fitting overall models, whereas 0/1 loss or error rate is common in assessing classification rules. Issues of criteria for assessing, selecting, and estimating parameters of classification rules are discussed in detail in Hand (1997).

In this paper we are chiefly concerned with a particular type of model based on the sampling paradigm. In order to apply the sampling paradigm, we need to estimate the multivariate class-conditional distributions of $\mathbf{x}$. Books on statistical pattern recognition (e.g. Hand, 1981; McLachlan, 1992; Ripley, 1996; Hand, 1997; Webb, 1999) describe such methods, of which there are many. At one extreme we have nonparametric methods, such as kernel approaches (e.g. Hand, 1982), which make very weak assumptions about the class-conditional distributions. Towards the other extreme, we have methods which make strong parametric assumptions about the form of the distributions, such as linear discriminant analysis which assumes that the distributions are ellipsoidal—multivariate normal being the most familiar example. In ellipsoidal distributions the overall probability function $P(\mathbf{x}|i)$ factorises into terms which involve only pairs of the X_r, so that higher order interactions are assumed to be zero. Of course, such a gross factorisation of the class-conditional distribution is not necessary, and intermediate factorisations are possible, especially if one has some prior knowledge of the form that such a factorisation might take. In this paper, however, we are concerned with a model which makes an even more severe factorisation, assuming that $P(\mathbf{x}|i)$ will factorise into a product of its univariate marginals—that the components of $\mathbf{x}$ are independent: $P(\mathbf{x}|i) = \prod_{r=1}^{d} P(x_r|i)$. Clearly this model is likely to be unrealistic for most problems. Moreover, it is easy to invent examples where this model is incorrect. Despite this, this model has a long and successful history. The aim of this paper is to look at that history, and to examine why the model so often performs well.

This model, which assumes that the measurements are independent within each class, has gone under various names. Because of the use of Bayes theorem to deduce the class membership probabilities, the word 'Bayes' often figures in the names. Names which have been used to describe it include *idiot's Bayes* (Ohmann *et al.*, 1988), *naïve Bayes* (Kononenko, 1990), *simple Bayes* (Gammerman & Thatcher, 1991), and *independence Bayes* (Todd & Stamper, 1994)—the first two in recognition of the fact that its basic simplicity, ignoring interactions which might almost always be expected to exist, might be regarded as naïve. Sometimes the model is described as assuming *conditional* independence, referring to the conditioning on the class indicator.

The early literature on machine diagnosis (as well as some which is not so early), demonstrated some misunderstanding of the necessity of the independence assumption. For example, Barker & Bishop (1970) say '... (in) statistical methods based on Bayes' theorem ... it is necessary to assume that symptoms are independent', and O'Connor & Sox (1991) comment 'implicit in Bayes theorem is the assumption of conditional independence'. Given the computational burden of including dependencies, one can perhaps have some sympathy with this mistake being made before the era of powerful computers, even though Fisher showed how to relax the assumption as early as 1936 (Fisher, 1936). One has less sympathy with more recent occurrences, though the fact that 'diagnostic tests are often used in sequence' (O'Connor & Sox, 1991) may partly explain why people make it.

In Section 2 we examine why the independence model might be expected to perform well, despite superficial theoretical arguments to the contrary. In Section 3 we look at how the model may be extended. Section 4 reviews some additional comparative studies of the method. Section 5 presents some summarising and concluding remarks.

Before commencing, however, some clarification of the role of the word 'Bayes' in this context is appropriate. In work on supervised classification methods it has become standard to refer to the $P(i)$ as the class i *prior probability*, because this gives the probability that an object will belong to class i prior to observing any information about the object. Combining the prior with $P(\mathbf{x}|i)$, as shown above, gives the *posterior* probability, after having observed $\mathbf{x}$. This combination is effected via

Bayes theorem, and it is because of this that 'Bayes' is used in the name of these methods. (Although it appears to be used widely only in methods based on the independence assumption for the $P(\mathbf{x}|i)$, which possibly explains some of the misunderstandings indicated above.) The important point for us is that no notion of subjective probability is introduced: the methods are not 'Bayesian' in the formal statistical sense. Perhaps it would be better to say 'not necessarily Bayesian in the formal sense', since there is no reason why this need be so, and in other work we are developing 'Bayesian' Bayes independence classifiers. In this paper, following almost all of the work on the idiot's Bayes method, we adopt a frequentist interpretation.

2 Why the Assumption is Not So Absurd

The fact that the assumption of independence is clearly almost always wrong (naturally occurring covariance matrices are rarely diagonal) has led to a general rejection of the crude independence model in favour of more complicated alternatives, at least by researchers knowledgeable about theoretical issues. And yet practical comparisons have often shown it to perform surprisingly well. For example, in a study of head injury data, Titterington *et al.* (1981) found that the independence model yielded the overall best result. Similarly, in a study comparing classifiers for predicting breast cancer recurrence, Mani, Pazzani & West (1997) found that the independence Bayes model did best. Moreover, in studies of heart disease (Russek, Kronmal & Fisher, 1983), thyroid disease (Nordyke, Kulikowski & Kulikowski, 1971), liver disease (Croft & Machol, 1987), abdominal pain (Gammerman & Thatcher, 1991; Todd & Stamper, 1994; Ohmann *et al.*, 1996), and dyspepsia (Fox, Barber & Bardhan, 1980) the independence model was a good choice. The phenomenon is not limited to medicine. Other studies which found that the independence Bayes method performed very well, often better than the alternatives, include Cestnik, Kononenko & Bratko (1987), Clark & Niblett (1989), Cestnik (1990), Langley, Iba & Thompson (1992), Pazzani, Muramatsu & Billsus (1996), Friedman, Geiger & Goldszmidt (1997), and Domingos & Pazzani (1997). Monte Carlo studies have also yielded good results with the independence model, even when the assumption is manifestly false (see, for example, Ott & Kronmal, 1976). There seem to be very few studies which report poor performance for the independence Bayes model in comparison with other methods. One is the STATLOG project, for which (King, Feng & Sutherland, 1995) the independence Bayes model appeared 'near the bottom of the performance table for almost all data sets.' (Even so, for the 'heart data' in that study, the independence model was the best.) The fact that this study is so out of tune with most others does require explanation. A hint might be obtained from those researchers' perspective of what the naïve Bayes model is. Michie, Spiegelhalter & Taylor (1994, p.40), in a review of the STATLOG project, say 'if an attribute takes a continuous value, the usual procedure is to discretise the interval and to use the appropriate frequency of the interval, although there is an option to use the normal distribution to calculate probabilities'. This, of course, is a crude version of the approach, and it may result in discriminating information being sacrificed relative to information in the original continuous variables. It is not necessary to discretise the variables. Of course, this is not the only study showing situations where the independence Bayes model does not perform as well as alternatives (e.g. Heckerman & Nathwani, 1992). It would be extraordinary if it were. But the surprising point is that the simple independence model performs so well so often.

Hand (1992) summarised some reasons why one might expect the independence Bayes model to do well. One important reason is that it requires fewer parameters to be estimated than alternative methods which seek to model interactions in the individual class-conditional $\mathbf{x}$ distributions. This matters because models based on fewer parameters will tend to have a lower variance for the estimates of $P(i|\mathbf{x})$. This can be seen from the following general result given by Altham (1984). Suppose that a model is based on a vector of p parameters, α, and let $\beta = \beta(\alpha)$ be a vector of q parameters, defined as functions of the parameters α, with $q < p$. (In our case, β will be the subset of α which

refers to the univariate distributions, ignoring interactions.) Let α_1 be the vector yielding L_1, the maximum likelihood for the model, and let α_2 be the vector yielding L_2, the maximum likelihood for the model under the constraints implied by the relations $\beta = \beta(\alpha)$. Then, for large sample sizes, if $f(\alpha)$ is an arbitrary function of α, $\text{var}(f(\alpha_2)) \leq \text{var}(f(\alpha_1))$.

This fact may underlie the results of Russek, Kronmal & Fisher (1983), who describe a study to predict which heart disease patients would die within six months. When a subset of six variables was used, the independence model performed poorly relative to the other methods (linear discriminant analysis and nonparametric methods), but when all 22 variables were used, it performed well. With 22 variables any method more sophisticated than assuming independence could have a large number of parameters to estimate. (We present another possible explanation for the difference between these two models below.)

Although the variance of the estimates $\hat{P}(i|\mathbf{x})$ of the $P(i|\mathbf{x})$, as different design samples are taken, may be lower in the simple independence model than the variance of a complex alternative, the independence model is also likely to yield biased probability estimates for much of the $\mathbf{x}$ space, so that $E[\hat{P}(i|\mathbf{x})] \neq P(i|\mathbf{x})$, where the expectation is over different design samples. One manifestation of this occurs in the phenomenon of 'probability overshoot' or 'multiple counting', which can be illustrated as follows.

To take an extreme case, suppose that the variables are perfectly correlated, so that, for any given class, the probability that the rth variable takes the value x is the same for all r. The true odds ratio, $P(i|\mathbf{x})/P(j|\mathbf{x})$, is then given by

$$\frac{P(i|\mathbf{x})}{P(j|\mathbf{x})} = \frac{P(i)P(\mathbf{x}|i)}{P(j)P(\mathbf{x}|j)} = \frac{P(i)P(x|i)}{P(j)P(x|j)}, \tag{1}$$

when $\mathbf{x} = (x, x, \ldots, x)$, and is undefined when the components of $\mathbf{x}$ are different (which does not matter, since this never occurs if the components of $\mathbf{x}$ are perfectly correlated).

In contrast, the independence model is based on estimating the odds ratio $P(i|\mathbf{x})/P(j|\mathbf{x})$ assuming independence of the components of $\mathbf{x}$ within each class. That is, it is based on estimating

$$\frac{P(i|\mathbf{x})}{P(j|\mathbf{x})} = \frac{P(i)P(\mathbf{x}|i)}{P(j)P(\mathbf{x}|j)} = \frac{P(i)\prod p(x_r|i)}{P(j)\prod P(x_r|j)}. \tag{2}$$

When the variables are perfectly correlated, this simplifies to

$$\frac{P(i|\mathbf{x})}{P(j|\mathbf{x})} = \frac{P(i)}{P(j)}\left[\frac{P(x|i)}{P(x|j)}\right]^d. \tag{3}$$

Comparison of (1) with (3) shows that, when the ratio $P(x|i)/P(x|j)$ is less than 1, the independence model will tend to underestimate $P(i|\mathbf{x})/P(j|\mathbf{x})$, and when the ratio $P(x|i)/P(x|j)$ is greater than 1 the independence model will tend to overestimate $P(i|\mathbf{x})/P(j|\mathbf{x})$. Furthermore, it can be seen that the phenomenon will be more marked the more variables there are (the larger is d in Equation 3). That is, the independence model will have a tendency to be too confident in its predictions and will tend to produce modes at the extremes 0 and 1 (Cornfield, 1971; Cornfield *et al.*, 1973; Ott & Kronmal, 1976; Russek, Kronmal & Fisher, 1983). Russek *et al.* (1983) also discuss theoretical explanations for this phenomenon.

Sometimes the reduction in variance resulting from the relatively few parameters involved in the independence model will more than compensate for any increase in bias, and we suspect that this often occurs in practice and is one explanation for the frequent good performance of the method. Note, however, that if simple classification is the aim, then bias may not matter: the best achievable classification results will be obtained provided $\hat{P}(i|\mathbf{x}) > \hat{P}(j|\mathbf{x})$ whenever $P(i|\mathbf{x}) > P(j|\mathbf{x})$ (see, for example, Friedman, 1997) and bias, even severe bias, will not matter provided it is in the right direction.

In summary, the independence Bayes method may beat flexible highly parameterised alternatives

when small sample sizes are involved because of the relatively low variance of its estimates of $P(i|\mathbf{x})$, and may in fact also do well when quite large sample sizes are involved because its relatively large bias may not matter over most of the possible patterns (Kohavi, 1996).

In many problems the variables undergo a selection process before being combined to yield a classification (if it is expensive to measure the variables, for example, or in personnel classification or psychiatric screening, where one wants to avoid asking too many questions). Examples are the stepwise forward, backwards, and combined methods common in classical discriminant analysis. The aim is to remove redundant variables, so that superfluous variables are not measured, and also so that unnecessary parameters do not have to be estimated. Typically (but not necessarily—see below) this process leads to the elimination of variables which are highly correlated with those already included—such variables are less likely to contribute much additional classification information. This means that, in real problems, there will be a tendency towards selecting weakly correlated variables, so that the independence model may be quite a realistic model in many real situations.

On the other hand, we note that, in contrast to the above, the variable selection process *could* also mean that some comparative studies are perhaps unfairly biased against the independence model. In particular, if the variable selection is made on the basis of performance of a model which makes some allowance for dependencies, then it is likely that the selected variables will do better with such a model than with a model which assumes independence. This is another possible explanation for the result of Russek *et al.* (1983) mentioned above, in which an independence model based on a selected six variables (in fact, selected using linear discriminant analysis) did poorly, while one based on all 22 variables did well. To make fair comparisons, the variables for the independence model should be selected independently. In fact, earlier and less sophisticated studies did do exactly this (e.g., Goldberg, 1972).

It is also possible that the independence model will yield the optimal separating surface (subject to sampling fluctuations) even if the variables are correlated. An example is given by two multivariate normal classes with equal (non-diagonal) covariance matrices, and with the vector of the difference between the means lying parallel to the first principal axis. Note, however, that even though the population decision surface is optimal, and even though the estimate of $P(i|\mathbf{x})$ may be accurate, the estimates of the probability density functions for each class separately, based on the independence assumption, may be very wrong.

In a recent paper, Domingos & Pazzani (1997) examine conditions under which one might expect the independence Bayes model to do well for categorical predictor variables. They comment (page 106) that 'the simple Bayesian classifier is limited in expressiveness in that it can only create linear frontiers'. In fact this is not true—unless they are assuming highly restricted marginal forms. For example, if on one of the variables the marginal for one population has the form of a two component mixture, with well separated means, while the other has a distribution which lies between these two components, then the decision surface will be nonlinear. Even simpler, if the two populations have normal marginals, but with different variances, then the decision surface will be quadratic. These authors also comment on the point noted above, that for *classification*, accurate probability estimates are not required—all that need be preserved is the rank order.

3 Modifications to the Basic Independence Model

The basic independence Bayes model has been modified in various ways in attempts to improve its performance (usually interpreted as classification accuracy as measured by a 0/1 loss function). Of course, these modifications sometimes mean that the elegance, computational simplicity, and other advantages (e.g. ease of handling incomplete vectors) can be lost. In this section, we examine these modifications.

Hilden & Bjerregaard (1976) suggested making some allowance for dependence by shrinking

the marginal probability estimates towards zero by raising them to a power B less than 1. This suggestion is based on the probability overshoot property described in Section 2. Titterington *et al.* (1981) used this approach in a comparison of a wide range of discrimination methods applied to a data set describing head injuries. The data were categorical. They estimated the separate marginal distributions of the x_r using multinomial estimators, but modified these slightly to take into account the number of possible response categories in the variable (see Jeffreys, 1961; Bailey, 1964; Good, 1965). Thus, for class i the probability that the rth variable X_r takes value x_r, $P(X_r = x_r|i)$, is estimated by

$$\hat{P}(X_r = x_r|i) = \frac{n_i(x_r) + 1/C_r}{N_i(r) + 1}, \tag{4}$$

so that

$$\hat{P}(\mathbf{x}|i) = \prod_{r=1}^{d} \frac{n_i(x_r) + 1/C_r}{N_i(r) + 1}, \tag{5}$$

where $n_i(x_r)$ is the number of design set elements from the ith class which fall in category x_r for variable r, C_r is the number of categories for variable r, and $N_i(r)$ is the number of design set elements which have an observation on the rth variable (the paper describes application on a real data set which has missing values). If there are no missing values, the denominator is constant in each class. This basic form of estimator was then modified by raising it to the power B

$$\hat{P}(\mathbf{x}|i) = \left\{ \prod_{r=1}^{d} \frac{n_i(x_r) + 1/C_r}{N_i(r) + 1} \right\}^B, \tag{6}$$

where B is a positive exponent less than 1. This is now a common form for cases when the predictors are categorical. This study examined several subsets of variables. In a subset of four 'weakly dependent variables with appreciable missing data' the independence models yielded the best performance. Even in a subset of four 'highly dependent variables with little missing data' the independence model was almost the best if assessed by logarithmic score. The best performance over all variable sets and methods used was produced by the independence model with $B = 0.8$.

Thornton, Lilford & Newcombe (1991) also attempted to overcome the independence assumption by adjusting the odds ratios, in their case by rescaling the odds ratio using a factor which was a function of the false negative and false positive rates. They produced tables for the risks of fetal neural tube and ventral wall defects, based on this model.

More elaborate attempts to overcome the independence assumption are based on including extra terms describing the relationship between the components of $\mathbf{x}$—see, for example, Friedman, Geiger & Goldszmidt (1997). Of course, this means that the simple form of the independence Bayes model is sacrificed. In fact, these particular authors approached things from the opposite direction, identifying weaknesses of the approach based on constructing an overall model of the joint distribution $P(i, \mathbf{x})$, and restricting themselves to models which included all the two way marginals of $(\mathbf{x}, i)$ which had the class label as one component. This was motivated by comparisons of independence Bayes methods with more elaborate Bayesian belief networks constructed using the minimum description length (MDL) to choose the model (Rissanen, 1978). They found that, in some situations, the independence Bayes method did better, while in others the more elaborate models did better. They attributed the inconsistency to the choice of MDL as the model selection criterion, arguing that this is a global measure of goodness of fit, and not one which focuses on the conditional distribution of the class given the covariates (see also Kontkanen *et al.*, 1999). This led them to explore the set of models, described above, in which the class identifier has edges in the Bayesian belief network linking it

to all components of $\mathbf{x}$. Broadly speaking, the extra complexity in these models did yield greater accuracy.

As we have noted, adding extra edges to include some of the dependencies between the predictors in this way detracts from the elegant simplicity of the basic independence Bayes model. An alternative generalisation, which remains within the mould of independence Bayes, has been explored by Ohmann *et al.* (1986), Langley & Sage (1994), Singh & Provan (1996), Kontkanen *et al.* (1999), and others, sometimes called 'selective Bayesian classifiers'. They built independence Bayes model forms, but included a variable selection stage, so that their final model need not include all of the variables, with redundant variables being eliminated. Clearly, depending on the number of raw variables and the relationships between them, this may lead to superior performance. In a similar vein, Crichton & Hinde (1988, 1989) used correspondence analysis to help to choose which variables, from a large potential set, should be included in an independence Bayes model. They went on to compare this approach with CART on a chest pain data set, finding that the independence Bayes model yielded superior performance.

We have already noted that, for classification, our ultimate interest lies in *comparing* the distributions, and not in calculating their absolute values. Thus we will classify a point $\mathbf{x}$ by choosing the largest of the $\hat{P}(i|\mathbf{x}) \propto \hat{P}(i)\hat{P}(\mathbf{x}|i)$, where we assume (the basic independence Bayes model) that $P(\mathbf{x}|i) = \prod_{r=1}^{d} P(x_r|i)$. We can extend this model by supposing that each $P(\mathbf{x}|i)$ has the form $P(\mathbf{x}|i) = g(\mathbf{x})\prod_{r=1}^{d} Q(x_r|i)$, where the g is an arbitrary factor which is common to all classes. Comparison between the $P(\mathbf{x}|i)$ is then equivalent to comparison between the $\prod_{r=1}^{d} Q(x_r|i)$, because the g factor cancels (Hilden, 1984). Indeed, we do not even need to know what the g factor is. The difficulty here is that we cannot estimate the $\prod_{r=1}^{d} Q(x_r|i)$ directly from the marginals in the design set. However, if one estimates *ratios* (c.f. the diagnostic paradigm) then g cancels. To see how this compares with the straightforward independence model in more detail, consider the following. From Equation 2, for the independence model, denoting estimates of P by $\hat{P}$, we have

$$\frac{\hat{P}(i|\mathbf{x})}{\hat{P}(j|\mathbf{x})} = \frac{\hat{P}(i)\hat{P}(\mathbf{x}|i)}{\hat{P}(j)\hat{P}(\mathbf{x}|j)} = \frac{\hat{P}(i)\prod \hat{P}(x_r|i)}{\hat{P}(j)\prod \hat{P}(x_r|j)}. \tag{7}$$

For the more general form for $\hat{P}(\mathbf{x}|i)$ given by $\hat{P}(\mathbf{x}|i) = g(\mathbf{x})\prod_{r=1}^{d} \hat{Q}(x_r|i)$ we have

$$\frac{\hat{P}(i|\mathbf{x})}{\hat{P}(j|\mathbf{x})} = \frac{\hat{P}(i)\hat{P}(\mathbf{x}|i)}{\hat{P}(j)\hat{P}(\mathbf{x}|j)} = \frac{\hat{P}(i)g(\mathbf{x})\prod \hat{Q}(x_r|i)}{\hat{P}(j)g(\mathbf{x})\prod \hat{Q}(x_r|j)} = \frac{\hat{P}(i)\prod \hat{Q}(x_r|i)}{\hat{P}(j)\prod \hat{Q}(x_r|j)}. \tag{8}$$

Note that in this formulation neither g nor the Q need be probability distributions; only their overall product is. This means that the second model includes the first (the simple independence model) as a special case. Of course, it also means that one cannot estimate the $\hat{Q}$ factors from the marginal distributions, which is one of the attractive features of the independence Bayes model. Instead an iterative approach is necessary. In fact, if one takes logarithms to expand the product as a sum, this reduces to logistic regression. This relationship between independence Bayes and logistic models is explored in more detail in Hand & Adams (2000).

Although this paper describes frequentist approaches to using the independence Bayes model, there is no reason why full Bayesian approaches should not be used. Of course, since these involve combining multiple models according to the prior distributions of the parameters, the simplicity of the basic independence Bayes model is again lost. On the other hand, as many studies have shown, this can yield improved predictive performance. Kontkanen *et al.* (2000) discuss such approaches.

In recent years several general methods have been developed for improving the performance of classification rules. One of these, *boosting* (Freund & Schapire, 1996), involves iteratively building a series of model, each time weighting those design set points which have been misclassified by the previous version of the model, and then taking the final model as a weighted sum of the separate

models. Unfortunately, because it is a combination of independence models, the final result from this procedure sacrifices the interpretable simplicity of the basic independence Bayes model and the computational simplicity of a direct estimate from the marginal distributions. However, Ridgeway *et al.* (1998) have developed a version which retains the interpretability.

Both boosting and Bayesian model averaging can be regarded as ways of combining different classifiers to yield a single model. In these cases, all of the component models have the form of independence Bayes classifiers, but other authors have experimented with combining independence Bayes models with other forms. For example, Monti & Cooper (1999) have combined finite mixture models with independence Bayes methods. Of course, this loses the simple independence Bayes form.

More generally, the independence model may be used as a base from which to introduce refinements in an attempt to produce more effective classification rules. An illustration is the work of Lowe & Webb (1990), who use the independence Bayes model as a starting point for a neural network model.

For non-categorical variables, one can model the component marginal distributions in a wide variety of ways. The simplest would be to adopt some parametric form. However, given the likely high density of points consequent on the projection from the multivariate space to the marginals, more sophisticated nonparametric methods can be used. John & Langley (1995), for example, showed that improvement resulted if marginal Gaussian estimators were replaced by kernel estimators (yielding a generalised additive model) or discretised marginals. Dougherty, Kohavi & Sahami (1995) also reduced the continuous marginals by discretising them.

4 Examples of Empirical Studies of the Independence Bayes Model

A large number of studies use the independence model in investigations of the efficacy of automated classification methods, often with particular application domains in mind. In many cases, the independence model is taken as a baseline, against which to compare the more sophisticated methods, but in many studies it does surprisingly well. This section reviews a selection of such studies. A comprehensive review would probably be impossible, especially when one takes into account the range of branches of scientific literature which have contributed to these investigations, from the obvious ones of statistics and machine learning, through application domains such as medicine and engineering.

Bailey (1964), in an early description of statistical methods for medical diagnosis, described the independence model. He also suggested ways (which have subsequently been explored, by others—see the general books on statistical pattern recognition listed in Section 1) for including dependencies between the variables.

Boyle *et al.* (1966) used independence Bayes in an early investigation of machine diagnosis applied to distinguishing between three classes of goitre. Unfortunately, the models that this paper compared were independence Bayes with and without taking into account the priors of the classes (it was a very early study), so that the only conclusions which could be drawn were about the value of the methods for matching the clinicians' predictions. The authors were aware of the independence assumption underlying their model, as well as the possibility that more than one of the diagnostic classes might occur together.

Fryback (1978) compared the performance of the independence Bayes model applied to skull fracture data with the performance that would have been obtained had the variables really been conditionally independent given the observed marginals. This seems a rather odd comparison. He then relaxed the independence assumption to permit selected pairwise dependencies—resulting in improved performance using error rate. (Derived from an ROC curve, though he appears not to have used the positions on the ROC curve that lead to minimum error rate, which requires a projection operation—see Hand (1997) for details.) One interesting observation made in this study is that

'the greater the number of variables used in the Bayesian calculations, the more the degradation of the model's performance when the data violate conditional independence'. Fryback goes on to say 'When relatively few variables are used, the independent contributions of each toward making the final Bayesian diagnosis seem to outweigh the degradation in performance due to over weighting redundant information. As the variable set increases in size, the independent contribution gained by adding yet one more variable tends to be less than the detrimental effect of multiple counting its information already conveyed by other variables'. This is the probability overshoot phenomenon mentioned in Section 2.

Crichton, Fryer & Spicer (1987) re-examined some of the data used in de Dombal's pioneering work (e.g. de Dombal *et al.*, 1972; de Dombal, 1991) with the objective of assessing the problems which arise from the independence assumption. They focused on the two class problem of distinguishing between non-specific abdominal pain and specific abdominal pain, the latter meaning one of the diagnostic classes appendicitis, perforated peptic ulcer, small bowel obstruction, cholecystitis, pancreatitis, or diverticulitis. They found a marginal improvement in the test set error rate when the number of variables was reduced from 46 to 5, using a stepwise selection procedure to maximise the proportion of design set cases correctly classified. This sort of phenomenon should be expected with any classifier, of course (see, for example, Hand (1981), Chapter 6), because of the reduced variance consequent on reducing the number of parameters to be estimated, as described above. In the present situation, however, the method may be especially beneficial since it may lead to rejection of variables which substantially contravene the independence assumption. The definition of one of the classes as a mixture of specific disease classes might cause problems. This fundamentally heterogeneous class may be less susceptible to reasonable approximation by an independence model than might the individual components.

Ohmann *et al.* (1988) remarked that 'having a limited sample, choosing the Independence Bayes model and using many variables may well provide a better classification rule than keeping fewer variables and using a sophisticated model', where they used Fisher's linear discriminant analysis as an example of a sophisticated model. They compared the basic independence Bayes method, independence Bayes with an overall shrinkage factor as in the Titterington *et al.* (1981) study, a method which grouped the variables into pairs and allowed dependence within a pair but independence between pairs, a method which allowed all pairwise dependencies, and linear logistic regression applied to the diagnosis of upper gastro-intestinal bleeding. Interestingly, they report that 'the good results reported in the literature [with the independence Bayes model] could not be achieved in our study, especially with many variables'. Their suggested explanations include the large number of pairwise interactions between the variables. They note that the predicted probabilities demonstrated the U-shaped distribution described above. Finally, 'in our study, the model could be recommended only for the sets with few variables'. Interestingly, however, 'for all sets of variables and all performance criteria, Independence Bayes with a global association factor produced the most reliable predictions'.

Sutton (1989) presented a critical review of computer-aided diagnosis of acute abdominal problems. He concluded that much of the improved patient management and outcome effects are due to the structured data collection methods and feedback which are a part of such investigations, rather than the automated diagnostic methods themselves (see also Wellwood & Spiegelhalter, 1989; Paterson-Brown *et al.*, 1989; Gunn, 1991), and he pointed out many of the weaknesses of the early studies. However, despite writing as late as 1989, he claimed that Bayes formula assumes that the 'indicants are independent'.

Gammerman & Thatcher (1991) compared the independence Bayes model with a model which included some symptom combinations and also the CART algorithm. Using error rate on an independent test set to assess the results, they found that the independence Bayes yielded best performance.

In a comparative study of diagnostic programs for abdominal pain, Todd & Stamper (1994) said

'the usual approach is to apply Bayes' theorem with the assumption of conditional independence'. They concluded that 'in this application no significant improvement in accuracy can be made by taking interactions into account, either by statistical or by knowledge-based means; independence Bayes is near optimal'.

Ohmann *et al.* (1996) remarked that 'the standard model used in computer-aided diagnosis of acute abdominal pain is Bayes theorem with the assumption of conditional independence.' They compared this model with six automatic rule induction techniques and evaluated the results on an independent test set. They found no differences in overall accuracy (except for one particular, less accurate, rule induction method) although they did find 'considerable differences with respect to specific diagnoses'. In particular, they concluded that 'machine learning techniques did not improve the results of the standard model Independence Bayes'.

Penny & Frost (1997) describe the use of a variety of classification rules to predict the class of 'level of observation' to which psychiatric patients will be assigned. The initial 512 variables were reduced to 50 by a process which would seem to favour the independence model, but were then further reduced to ten by stepwise regression. The results show independence Bayes performing better than nearest neighbour methods.

The trouble with all such empirical comparisons is that one can always refine the methods. The poor showing of the nearest neighbour methods is surprising, and might be attributed to the choice of the Euclidean metric, which is unlikely to be optimal. Perhaps a nearest neighbour method with a sophisticated adaptive metric might have done well. Moreover, the paper does not state how k in the k-nearest neighbour method was chosen. In contrast, the authors of this paper were clearly able to refine and tune their neural network algorithm—in a way others might not be able to do. The point is that, as discussed in Hand (1997), any comparison of rules should really be from the perspective of someone who is likely to use them. In this regard, the independence Bayes model has the merit that it is straightforward. Off-the-shelf application, even involving the B shrinking parameter, involves no great skills.

Penny & Frost (1997) also make the important point that sometimes the measured variables have a deterministic structure—and that this should be taken advantage of in formulating the model. The independence Bayes model ignores any such structure.

Mani, Pazzani & West (1997) described a comparative study of two tree classifiers, two rule inducers, and the independence Bayes model to predict breast cancer recurrence on the basis of six attributes. Evaluating the rules on an independent test set, they found that the independence Bayes model outperformed the other models.

Of course, the relative performance of models depends on how one compares them—see Hand (1997) for an extensive discussion. It is possible that the use of different performance criteria could lead to difference conclusions. In particular, the wider use of measures which assess the accuracy of the $\hat{P}(i|\mathbf{x})$ as estimates of the $P(i|\mathbf{x})$ (such as the Brier score, used in the study of Titterington *et al.*, 1981), rather than the common use of measures based on proportions of objects correctly classified might lead one to different conclusions, because of the probability overshoot phenomenon mentioned above. Similarly, the use of a cost matrix to weight the different kinds of misclassification properly (in those cases where this can be determined—see Adams & Hand, 1999; Kelly *et al.*, 2000) could lead to different conclusions. In general, of course, the performance criterion should be chosen to match the objectives of the problem.

5 Conclusion

It will be clear from the above that many of the practical applications of the independence model are in medical areas. We speculate that this is at least partly because of the attractive simplicity of the method. However, it has been applied in other areas as well. Willcox & Lapage (1975), Willcox,

Lapage & Holmes (1980), and Willcox *et al.* (1973), for example, apply it to bacterial identification, and Hand & Adams (2000) apply it to credit scoring. Other publications describing the model in medical contexts include Adams *et al.* (1986), Brunk *et al.* (1975), Coomans *et al.* (1983), Cornfield *et al.* (1973), Croft & Machol (1987), duBoulay *et al.* (1977), de Dombal (1991), de Dombal *et al.* (1972), Edwards & Davis (1984), Heckerman, Horvitz & Nathwani (1992), Spiegelhalter & Knill-Jones (1984), van Woerkom & Brodman (1961), and Warner *et al.* (1961).

The independence Bayes model seems often to perform surprisingly well. There are sound reasons for this: its intrinsic simplicity means low variance in its probability estimates; although these will typically be biased, this may not matter in classification situations as long as the rank order is preserved; in many situations the variables have undergone a selection process which tends to reduce their interdependencies; moreover simple extensions of the basic independence structure can be adopted which improve its performance yet further.

Although this review has focused on the *performance* of the independence Bayes model as a classification rule, one should be aware that the method also has significant other merits. In particular, missing values in both the design set and in new cases to be classified can be handled with consummate ease—one simply ignores them. (Of course, we are assuming here that the values are missing completely at random. To do otherwise gets one into rather deeper water.) Alternatively, one can easily introduce 'missing' as an extra category of a response variable, without altering the basic structure of the model. Moreover, since the model has the form of a product, by taking logs it can be converted into a sum—with the usual consequent computational advantages.

Finally, we should remark that the list of references below is not a comprehensive list of papers discussing independence Bayes methods: our aim was to produce a critical review of the key ideas, rather than a simple list of all publications which had discussed or made use of those ideas. In particular, a large number of papers include the independence Bayes method as a comparator in studies of the effectiveness of different classification tools. Despite this, we hope that the references cited below cover the major theoretical issues, and provide routes into the main branches of the literature dealing with such methods, be it statistical, machine learning, artificial intelligence, or whatever.

Acknowledgements

Keming Yu's work on this project was supported by grant number CSM3200611 from the Defence Evaluation and Research Agency. We are indebted to Andrew Webb for his advice and comments on this work, and to an anonymous referee for a helpful and detailed critique of an earlier version and for drawing our attention to publications we had missed. We would also like to express our appreciation to George Barnard for his interest in and encouragement of this work.

References

Adams, I.D., Chan, M., Clifford, P.C., *et al.* (1986). Computer-aided diagnosis of acute abdominal pain: a multicentre study. *British Medical Journal*, **293**, 800–804.

Adams, N.M. & Hand, D.J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, **32**, 1139–1147.

Altham, P.M.E. (1984). Improving the precision of estimation by fitting a model. *Journal of the Royal Statistical Society, Series B*, **46**, 118–119.

Bailey, N.T.J. (1964). Probability methods of diagnosis based on small samples. In *Mathematics and Computer Science in Biology and Medicine*, pp. 103–107. London: HMSO.

Barker, D.J.P. & Bishop, J.M. (1970). Computer analysis of symptom patterns as a method of screening patients at a special risk of hyperthyroidism. *British Journal of Preventive and Social Medicine*, **24**, 193–196.

Boyle, J.A., Greig, W.R., Franklin, D.A., Harden, R.M., Buchanan, W.W. & McGirr, E. (1966). Construction of a model for computer-assisted diagnosis: application to the problem of non-toxic goitre. *Quarterly Journal of Medicine*, **35**, 565–588.

Brunk, H.D., Thomas, D.R., Elashoff, R.M. & Zippin, C. (1975). Computer aided prognosis. In *Perspectives in Biometrics*,

Ed. R.M. Elashoff, Vol.1, New York: Academic Press.
Cestnik, B. (1990). Estimating probabilities: a crucial task in machine learning. *Proceedings of the Ninth European Conference on Artificial Intelligence*. Stockholm, Sweden: Pitman.
Cestnik, G., Kononenko, I. & Bratko, I. (1987). ASSISTANT-86: A knowledge elicitation tool for sophisticated users. In *Progress in Machine Learning*, Eds. I. Bratko and N. Lavrac. Sigma Press.
Clark, P. & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, **3**, 261–283.
Coomans, D., Broeckaert, I., Jonckheer, M. & Massart, D.L. (1983). Comparison of multivariate discrimination techniques for clinical data application to the thyroid functional state. *Methods of Information in Medicine*, **22**, 93–101.
Cornfield, J. (1971). Statistical classification methods. In *Proceedings of the Symposium on the Diagnostic Process*. University of Michigan Press, Ann Arbor, June.
Cornfield, J., Dunn, R.A., Batchlor, C. & Pipberger, H. (1973). Multigroup diagnosis of electrocardiograms. *Computers and Biomedical Research*, **6**, 97.
Crichton, N.J., Fryer, J.G. & Spicer, C.C. (1987). Some points on the use of 'independent Bayes' to diagnose acute abdominal pain. *Statistics in Medicine*, **6**, 945–959.
Crichton, N.J. & Hinde, J.P. (1988). Correspondence analysis and independent Bayes for clinical diagnosis. *Proceedings of IMA Conference on Applications of Statistics in Medicine*, pp. 235–247.
Crichton, N.J. & Hinde, J.P. (1989). Correspondence analysis as a screening method for indicants for clinical diagnosis. *Statistics in Medicine*, **8**, 1351–1362.
Croft, D.J. & Machol, R.E. (1987). Mathematical models in medical diagnosis. *Ann. Biomed. Engng.*, **2**, 69–89.
Dawid, A.P. (1976). Properties of diagnostic data distributions. *Biometrics*, **32**, 647–658.
de Dombal, F.T. (1991). The diagnosis of acute abdominal pain with computer assistance: worldwide perspective. *Ann. Chir.*, **45**, 273–277.
de Dombal, F.T., Leaper, D.J., Staniland, J.R., McCann, A. & Horrocks, J. (1972). Computer aided diagnosis of acute abdominal pain. *British Medical Journal*, **2**, 9–13.
Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, **29**, 103–130.
Dougherty, J., Kohavi, R. & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 194–202. Tahoe City, California: Morgan Kaufman.
duBoulay, G.H., Teather, D., Harling, D. & Clark, G. (1977). Improvements in computer-assisted diagnosis of cerebral tumours. *British Journal of Radiology*, **50**, 849–854.
Edwards, F.H. & Davies, R.S. (1984). Use of a Bayesian algorithm in the computer-assisted diagnosis of appendicitis. *Surg. Gynecol. Obstet.*, **158**, 219–222.
Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, **70**, 892–898.
Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.
Fox, J., Barber, D. & Bardhan, K.D. (1980). A quantitative comparison with rule-based diagnostic inference. *Methods of Information in Medicine*, **19**, 210–215.
Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning*, pp. 256–285.
Friedman, J.H. (1997). On bias, variance, 0/1-loss, and the curse of dimensionality. *Data Mining and Knowledge Discovery*, **1**, 55–77.
Friedman, N., Geiger, D. & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, **29**, 131–163.
Fryback, D.G. (1978). Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Computers and Biomedical Research*, **11**, 423–434.
Gammerman, A. & Thatcher, A.R. (1991). Bayesian diagnostic probabilities without assuming independence of symptoms. *Methods of Information in Medicine*, **30**, 15–22.
Goldberg, D.P. (1972). *The Detection of Psychiatric Illness by Questionnaire*. Maudsley Monograph 21, London: Oxford University Press.
Good, I.J. (1965). *The estimation of probabilities*. Cambridge, Massachusetts: MIT Press.
Gunn, A.A. (1991). The acute abdomen: the role of computer-assisted diagnosis. *Baillieres Clin. Gastroenterology*, **5**, 639–635.
Hand, D.J. (1981). *Discrimination and Classification*. Chichester: John Wiley and Sons.
Hand, D.J. (1982). *Kernel Discriminant Analysis*. Letchworth: Research Studies Press.
Hand, D.J. (1992). Statistical methods in diagnosis. *Statistical Methods in Medical Research*, **1**, 49–67.
Hand, D.J. (1997). *Construction and Assessment of Classification Rules*. Chichester: John Wiley and Sons.
Hand, D.J. & Adams, N.M. (2000). Defining attributes for scorecard construction. *Journal of Applied Statistics*, **27**, 527–540.
Heckerman, D.E., Horvitz, E.J. & Nathwani, B.N. (1992). Toward normative expert systems: Part I: the Pathfinder project. *Methods of Information in Medicine*, **31**, 90–105.
Heckerman, D.E. & Nathwani, B.N. (1992). An evaluation of the diagnostic accuracy of Pathfinder. *Computers and Biomedical Research*, **25**, 56–74.
Hilden, J. (1984). Statistical diagnosis based on conditional independence does not require it. *Computers in Biology and Medicine*, **14**, 429–435.
Hilden, J. & Bjerregaard, B. (1976). Computer-aided diagnosis and the atypical case. In *Decision Making and Medical Care: Can Information Science Help*, Eds. F.T. de Dombal and F. Gremy, pp. 365–378. Amsterdam: North Holland.
Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press.
John, G. & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *Proceedings of the Eleventh*

Conference on Uncertainty in Artificial Intelligence, pp. 338–345. Montreal, Canada: Morgan Kaufman.

Kelly, M.G., Hand, D.J. & Adams, N.M. (2000). *Choosing good predictive models for consumer credit data*. Technical Report TR-00-15, Department of Mathematics, Imperial College, London.

King, R.D., Feng, C. & Sutherland, A. (1995). STATLOG—comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence*, **9**, 289–333.

Kohavi, R. (1996). Scaling up the accuracy of naïve-Bayes classifiers: a decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207. Portland, Oregon: AAAI Press.

Kononenko, I. (1990). Comparison of inductive and naïve Bayesian learning approaches to automatic knowledge acquisition. In *Current trends in knowledge acquisition*, Eds. B. Wielinga *et al.*. Amsterdam: IOS Press.

Kontkanen, P., Myllymäki, P., Silander, T. & Tirri, H. (1999). On supervised selection of Bayesian networks. In *Proceedings of the 15th International Conference on Uncertainty in Artificial Intelligence*, Eds. K. Laskey and H. Prade. San Mateo, CA: Morgan Kaufmann.

Kontkanen, P., Myllymäki, P., Silander, T., Tirri, H. & Grünwald, P. (2000). On predictive distributions and Bayesian networks. *Statistics and Computing*, **10**, 39–54.

Langley, P., Iba, W. & Thompson, K. (1992). An analysis of Bayesian classifiers. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 223–228. San Jose, California: AAAI Press.

Langley, P. & Sage, S. (1994). Induction of selective Bayesian classifiers. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pp. 399–406. Seattle, WA: Morgan Kaufmann.

Lowe, D. & Webb, A.R. (1990). Exploiting prior knowledge in network optimisation: an illustration form medical prognosis. *Network Comput. Neural Sys.*, **1**, 299–323.

Mani, S., Pazzani, M.J. & West, J. (1997). Knowledge discovery from a breast cancer database. *Lecture Notes in Artificial Intelligence*, **1211**, 130–133.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley and Sons.

Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. New York: Ellis Horwood.

Monti, S. & Cooper, G.F. (1999). A Bayesian network classifier that combines a finite mixture model and a naïve Bayes model. In *Proceedings of the 15th Conference on Uncertainty in AI*, Stockholm, Sweden.

Nordyke, R., Kulikowski, C.A. & Kulikowski, C.W. (1971). A comparison of methods for the automated diagnosis of thyroid dysfunction. *Computers and Biomedical Research*, **4**, 374–389.

O'Connor, G.T. & Sox, H.C. (1991). Bayesian reasoning in medicine: the contributions of Lee B. Lusted, MD. *Medical Decision Making*, **11**, 107–111.

Ohmann, C., Künneke, M., Zaczyk, R., Thon, K. & Lorenz, W. (1986). Selection of variables using 'independence Bayes' in computer-aided diagnosis of upper gastrointestinal bleeding. *Statistics in Medicine*, **5**, 503–515.

Ohmann, C., Moustakis, V., Yang, Q. & Lang, K. (1996). Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain. *Artificial Intelligence in Medicine*, **8**, 23–36.

Ohmann, C., Yang, Q., Künneke, M., Stöltzing, H., Thon, K. & Lorenz, W. (1988). Bayes theorem and conditional dependence of symptoms: different models applied to data of upper gastrointestinal bleeding. *Methods of Information in Medicine*, **27**, 73–83.

Ott, J. & Kronmal, R.A. (1976). Some classification procedures for multivariate binary data using orthogonal functions. *Journal of the American Statistical Association*, **71**, 391.

Paterson-Brown, S., Vipond, M.N., Simms, K., Gatzen, C., Thompson, J.N. & Dudley, H.A.F. (1989). Clinical decision making and laparoscopy versus computer prediction in the management of the acute abdomen. *British Journal of Surgery*, **76**, 1011–1013.

Pazzani, M., Muramatsu, J. & Billsus, D. (1996). Syskil and Webert: identifying interesting web sites. *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pp. 54–61. Portland, Oregon: AAAI Press.

Penny, W.D. & Frost, D.P. (1997). Neural network modelling of the level of observation decision in an acute psychiatric ward. *Computers and Biomedical Research*, **30**, 1–17.

Ridgeway, G., Madigan, D. & Richardson, T. (1998). Interpretable boosted naïve Bayes classification. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, Eds. R. Agrawal, P. Stolorz and G. Piatetsky-Shapiro, pp. 101–104. AAAI Press, Menlo Park, California.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, **14**, 465–471.

Russek, E., Kronmal, R.A. & Fisher, L.D. (1983). The effect of assuming independence in applying Bayes' theorem to risk estimation and classification in diagnosis. *Computers and Biomedical Research*, **16**, 537–552.

Singh, M. & Provan, G. (1996). Efficient learning of selective Bayesian network classifiers. In *Proceedings of the 13th International Conference on Machine Learning*, pp. 453–461, Bari, Italy.

Spiegelhalter, D.J. & Knill-Jones, R.P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems. *Journal of the Royal Statistical Society, Series A*, **147**, 35–76.

Sutton, G.C. (1989). Computer-aided diagnosis: a review. *British Journal of Surgery*, **76**, 82–85.

Thornton, J.G., Lilford, R.J. & Newcombe, R.G. (1991). Tables for estimation of individual risks of fetal neural tube and ventral wall defects, incorporating prior probability, maternal serum α-fetoprotein levels, and ultrasonographic examination results. *American Journal of Obstetrics and Gynaecology*, **164**, 154–160.

Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F. & Gelpke, G.J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society, Series A*, **144**, 145–175.

Todd, B.S. & Stamper, R. (1994). The relative accuracy of a variety of medical diagnostic programmes. *Methods of Information*

in Medicine, **33**, 402–416.
van Woerkom, A.J. & Brodman, K. (1961). Statistics for a diagnostic model. *Biometrics*, **17**, 299–318.
Warner, H.R., Toronto, A.F., Veasey, L.R. & Stephenson, R. (1961). A mathematical model for medical diagnosis—application to congenital heart disease. *Journal of the American Medical Association*, **177**, 177–184.
Webb, A. (1999). *Statistical Pattern Recognition*. London: Arnold.
Wellwood, J.M. & Spiegelhalter, D.J. (1989). Computers and the diagnosis of acute abdominal pain. *British Journal of Hospital Medicine*, **41**, 564–567.
Willcox, W.R. & Lapage, S.P. (1975). Methods used in a program for computer-aided identification of bacteria. In *Biological Identification with Computers*, pp. 103–119. New York: Academic Press.
Willcox, W.R., Lapage, S.P. & Holmes, B. (1980). A review of numerical methods in bacterial identification. *Antonie van Leeuwenhoek*, **46**, 233–299.
Willcox, W.R., Lapage, S.P., Bascomb, S. & Cursin, M.A. (1973). Identification of bacteria by computer: theory and programming. *J. Gen. Microbiol.*, **77**, 317–330.

Résumé

La tradition veut qu'une règle très simple assumant l'independance des variables prédictives, une hypothèse fausse dans la plupart des cas, peut être très efficace, souvent même plus efficace qu'une méthode plus sophistiquée en ce qui concerne l'attribution de classes a un groupe d'objets. A ce sujet, nous examinons les preuves empiriques, observées sur des données réelles, et les preuves théoriques, c'est-a-dire les raisons pour lesquelles cette simple règle pourrait faciliter le processus de tri.

[*Received June 2000, accepted March 2001*]