Confidence in classification: a Bayesian approach

Wojtek J. Krzanowski, Jonathan E. Fieldsend, Trevor C. Bailey,
Richard M. Everson, Derek Partridge, and Vitaly Schetinin

School of Engineering, Computer Science and Mathematics,
University of Exeter, Exeter, U.K.

Abstract

Bayesian classification is currently of considerable interest. It provides a strategy for eliminating the uncertainty associated with a particular choice of classifier-model parameters, and is the optimal decision-theoretic choice under certain circumstances when there is no single "true" classifier for a given data set. Modern computing capabilities can easily support the Markov chain Monte Carlo sampling that is necessary to carry out the calculations involved, but the information available in these samples is not at present being fully utilised. We show how it can be allied to known results concerning the "reject option" in order to produce an assessment of the confidence that can be ascribed to particular classifications, and how these confidence measures can be used to compare the performances of classifiers. Incorporating these confidence measures can alter the apparent ranking of classifiers as given by straightforward success or error rates. Several possible methods for obtaining confidence assessments are described, and compared on a range of data sets using the Bayesian probabilistic nearest-neighbour classifier.

Keywords.
Accuracy-rejection plots; Bayesian classification; confidence measures; MCMC sampling; nearest-neighbour classifiers.

1 Introduction

Bayesian methods have been advocated in principle for many years (Lindley, 1965; DeGroot, 1970), but their application has been hampered in practice by the computational intractability of many of the concomitant (high-dimensional) integrals. This state of affairs has been revolutionised in recent years by the development of Markov Chain Monte Carlo (MCMC) methods (see, e.g., the review by Brooks, 1998) and their reversible-jump (RJ) extensions (Green, 1995). These methods allow samples to be drawn from posterior distributions that are known only up to a constant of proportionality, thereby sidestepping the evaluation of the difficult integrals and

replacing other integrals by straightforward averages (or related simple summary statistics) of sampled values. The sampling process must usually be run for a very long time to allow the generated Markov Chains to stabilise at the required stationary distributions, but current computing power makes light of this demand. Consequently, there has been an explosion in the use of RJMCMC methods for statistical modelling in the past ten years.

One specific area of interest in such methods is that of discriminant analysis, or supervised classification. In essence here the problem is to define a suitable function of p features $\boldsymbol{x}' = (x_1, x_2, \ldots, x_p)$ that will best distinguish between g *a-priori* groups or populations, and that can be used to classify future unidentified individuals most accurately to their correct population. A set of individuals with known population membership is generally available for deriving the function (usually termed the classifier) and assessing its performance. If this set is large enough then it can be split into two independent parts to deal with these two aspects, the first part for training the classifier and the second part for testing its efficacy, but if the set is not large then some form of data resampling (such as jackknifing or bootstrapping) must be employed for the performance assessment. This whole area has now been studied for many years and there are many possible ways of deriving classifiers and determining their efficacies (McLachlan, 1992; Hand, 1997). A full Bayesian approach has only recently become viable, for the reasons outlined above, but the appropriate technology has been rapidly developed (Denison, Holmes, Mallick and Smith, 2002).

However, although the derivation of classifiers and the estimation of their classification performance has been worked out for a range of possible models and classifier types, other important aspects have received less attention. One such aspect, namely the confidence that can be ascribed to a particular classification result, is important in general but especially so in safety-critical systems such as medical diagnosis or air-traffic collision alert systems. We therefore focus in this paper on methods for deriving confidence measures about classifications in a Bayesian context. In section 2 we summarise the main features of Bayesian classification, in section 3 we derive two possible confidence measures and compare them on a range of data sets for one

particular classifier family, in section 4 we discuss how these measures can be used to choose between competing classifiers, and some concluding remarks are made in section 5.

2 Bayesian Classification

Bayesian classification is conducted within a parametric framework, so let us assume that the classifier $C(\boldsymbol{x}, \boldsymbol{\theta})$ comes from a family of models depending on the predictors $\boldsymbol{x}$ as well as on a set of parameters $\boldsymbol{\theta}' = (\theta_0, \theta_1, \ldots, \theta_q)$. For example, a linear classifier belongs to the family $C(\boldsymbol{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$ of all linear combinations of the predictors, with coefficients and constant term comprising the set of parameters. Applying the classifier to an individual $\boldsymbol{x}$ yields the values of one or more classification scores on which the classification of $\boldsymbol{x}$ is made; frequently these scores are the posterior probabilities of group memberships for $\boldsymbol{x}$. However, in general $\boldsymbol{\theta}$ is unknown. The classical single-classifier approach replaces it by an estimate derived from the training data D, say, and assesses the resulting classifier's efficacy by finding the proportion of each group that is misclassified in the test data T, say. Different methods of estimation make different demands on the data; for example, least squares estimates make no specific assumptions about D, but simply try to match observed and predicted classes as closely as possible, while for maximum likelihood estimation it is also necessary to postulate a probability model so that the likelihood of D can be obtained and maximised.

For a Bayesian approach we need to specify a joint prior distribution $\pi(\boldsymbol{\theta})$ for the classifier parameters, form the likelihood $L(D|\boldsymbol{\theta})$ of the training data using an appropriate probability model, and hence obtain the posterior distribution of the parameters,

$$\pi(\boldsymbol{\theta}|D) = \frac{\pi(\boldsymbol{\theta})L(D|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})L(D|\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

The Bayesian classifier is then the expected value of $C(\boldsymbol{x}, \boldsymbol{\theta})$ over this posterior distribution, $C(\boldsymbol{x}|D) = \int C(\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$. This is known as the *predictive* classification score. If the classification scores are the posterior probabilities of group membership then these predictive values are often

denoted by $p(y|\boldsymbol{x}, D)$, where y is the group label variable.

Evaluating the above two integrals can in general be very difficult, particularly when the dimensionality of $\boldsymbol{\theta}$ is large, but from its definition the Bayesian classifier will obviously be well approximated by the mean of $C(\boldsymbol{x}, \boldsymbol{\theta})$ over a large sample of independent observations from $\pi(\boldsymbol{\theta}|D)$. MCMC will enable such a sample to be drawn without having to evaluate the integral in the denominator of $\pi(\boldsymbol{\theta}|D)$. We just need to ensure that the MCMC acceptance probabilities are chosen so that $\pi(\boldsymbol{\theta}|D)$ is the limiting (stationary) distribution, run the chain for a preliminary (burn-in) period to ensure stationarity has been reached, and then sample (say) every 7th value to ensure independence of observations. Each value then yields a single observation from $\pi(\boldsymbol{\theta}|D)$, so substituting them in turn into $C(\boldsymbol{x}, \boldsymbol{\theta})$ for the particular $\boldsymbol{x}$ to be classified and averaging the results produces the Bayesian classifier.

This is just an example of Bayesian averaging, which is used much more generally in modelling (Hoeting, Madigan, Raftery and Volinsky, 1999). Of course, the Bayesian approach does not preclude the choice of a single "best" classifier, as one can simply be selected from the set of classifiers generated by the sampling process; the classifier obtained from the "maximum a-posteriori" (MAP) value of $\boldsymbol{\theta}$ would be an obvious choice. However, an averaged classifier not only usually produces better overall performance than the single MAP classifier, it is also the optimal decision-theoretic choice when there is no single "true" classifier that is being sought from among the family $C(\boldsymbol{x}, \boldsymbol{\theta})$ (Denison *et al*, 2002, pp 28-29). So it is the most appropriate one to use in many practical cases. The Bayesian approach has now been implemented for many different families of classifier, and details may be found, for example, in Denison *et al* (2002); we use the nearest neighbour family in the illustrations below. Also, as this is the form most commonly encountered in practice, we will assume that the classifier delivers posterior probabilities of group membership and therefore will denote the Bayesian predictive scores by $p(y|\boldsymbol{x}, D)$; we comment on some implications of this assumption in the final section.

3 Measures of Confidence

3.1 Methodology

A traditional method of reducing the risk of misclassification is by means of the *reject option* (surveyed in Fukunaga, 1990), whereby we do not automatically accept the outcome of the classifier for all points in the sample space, but hold back any points about whose classification we have doubts with the aim of handling these points subsequently by different procedures. If the resultant cost is less than the cost of wrong classification then such a procedure will improve classification reliability. We can label points $\boldsymbol{x}$ held back in this way as having UNSURE classification, and all other points as having SURE classification. Among the latter will be ones that are classified correctly and others that are classified incorrectly by the chosen classifier, so adopting such an approach will lead to three categories of points in a test set: those whose classification is SURE and CORRECT, those whose classification is SURE but INCORRECT, and those whose classification is UNSURE.

The question is, how should the holding back of points be determined? Various possibilities have been mooted (see, e.g., Bishop, 1995), but Chow (1970) showed that theoretically the optimal rejection rule is to hold back $\boldsymbol{x}$ if its maximum posterior probability of allocation to any group is less than a threshold t. Different values of t will lead to different proportions of UNSURE points.

In practice, of course, the posterior probabilities of allocation have to be estimated. If we use the Bayesian approach they are given by the values of $p(y|\boldsymbol{x}, D)$ for each possible setting of y, so $\boldsymbol{x}$ will be held back if $\max_y\{p(y|\boldsymbol{x}, D)\} < t$. Choosing a value of t and applying the classifier to all the points in the test set will identify the points to be classified and the points to be held back, thereby generating estimated probabilities of SURE CORRECT, SURE INCORRECT and UNSURE classifications for the given populations at the chosen value of t. We will call this procedure the standard reject method.

However, there are several problems with this approach. First, the choice

of threshold t is clearly arbitrary and there are no real guidelines as regards appropriate values to choose. In two-class problems t obviously has to be greater than 0.5, but choice of a reasonable level above this value is heavily data-dependent. In multi-class problems there is the additional difficulty of catering for the many ways in which posterior probabilities can be distributed among the classes. But most tellingly, basing the UNSURE classification purely on $\max_y\{p(y|\boldsymbol{x}, D)\}$ pays no regard to actual outcomes of classifications and hence to the confidence we can place in them. For example, a classifier might consistently deliver correct classifications of points for which $p(y|\boldsymbol{x}, D)$ is around 0.6, whereas it might occasionally misclassify points for which $p(y|\boldsymbol{x}, D)$ is around 0.8. Setting t at 0.7 would reject the former points but not the latter.

So the above discussion suggests that we should prefer a method that incorporates actual classification performance and that can be interpreted in terms of familiar quantities. The Bayesian scheme is ideally suited to this framework, as we have all the individual classification scores for each classifier making up the MCMC sample. These are all plausible classifiers that *might* have been used individually from the chosen family of classifiers. Moreover, they are not just a random selection from this family as the less likely members will be downweighted in the MCMC sampling, so they carry valuable information about the process. In particular, they give some indication about the likely *variability* of classification results, so we ought to make use of this information in formulating confidence measures. One way of doing so would be to classify each point in the test data by each of these individual classifiers; any point $\boldsymbol{x}$ that is classified to the same group by more than a proportion t of classifiers could be deemed a SURE classification, otherwise the classification is UNSURE. Here we convert each classifier result into a discrete variable (group to which $\boldsymbol{x}$ is classified) and then obtain the average of the incidences in each category, so the result is still a posterior probability of allocation and hence falls within the scope of Chow's result. We will call this the envelope procedure, as there is an envelope of classifications associated with each point.

This procedure counteracts the main objections raised above to the stan-

dard reject method. It uses consistency of actual classifications, so only labels points as UNSURE if they are unreliable in their classification rather than simply if their posterior probabilities of group membership are not high. There are no extra problems encountered in multi-class problems, as we are still looking to see whether points are consistently classified to a single class or distributed over several classes in the MCMC sample. Finally, and most usefully, the probabilities are now related to long-run classification performance so can be interpreted in familiar confidence-region fashion. For example, if we set t at 0.8 then we deem a point to be UNSURE if fewer than 80% of the MCMC samples classify that point to the same group, i.e. our confidence in classification of that point is less than 80%.

3.2 Applications

To illustrate the utility of this envelope approach, we apply it to a number of data sets from the UCI Machine Learning repository. However, first we must choose a family of classifiers. Many choices are possible, but to maintain flexibility while keeping the parameter dimensionality low we focus on $k-$nearest neighbour classifiers. The standard (classical) $k-$nearest neighbour classifier is very straightforward. To classify an observation $\boldsymbol{x}' = (x_1, \ldots, x_p)$ into one of g groups $y = (1, \ldots, g)$ we:

1. define a metric in the $\boldsymbol{x}-$space (usually Euclidean distance);
2. find the k training set members closest to $\boldsymbol{x}$;
3. classify $\boldsymbol{x}$ to the majority group among these k.

The value of k can either be set by the user or chosen from D by some data-based procedure, e.g. cross-validation.

Holmes & Adams (2002) have given a probabilistic formulation of this process, and this enables a Bayesian approach to be taken. They define

$$p(y = i|\boldsymbol{x}, \beta, k) = \frac{\exp(a_i\beta/k)}{\sum_i \exp(a_i\beta/k)},$$

where a_i is the number of group i individuals among the k nearest training set neighbours of $\boldsymbol{x}$ $(i = 1, \ldots, g)$, so that a_i/k is the proportion of such

individuals, and the parameter β reflects the influence of neighbours on the group probabilities: the greater the value of β, the higher the probability of belonging to the group that has the majority of neighbours. Thus the predictive scores are given by

$$p(y = j|\boldsymbol{x}, D) = \int p(y = j|\boldsymbol{x}, \beta, k)\pi(k, \beta|D)dkd\beta,$$

where $\pi(k, \beta|D)$ is the joint posterior distribution of the parameters β, k.

To obtain the posterior distribution we need the likelihood of the training data

$$L(D|\beta, k) = \prod_{i=1}^{n} \frac{\exp(a_{ij_i}\beta/k)}{\sum_j \exp(a_{ij}\beta/k)},$$

where a_{ij} is the number of the k nearest neighbours to the ith observation that belong to group j and j_i is the group to which the ith observation belongs. We also need to formulate a prior distribution $\pi(k, \beta)$ for the two parameters. In the case of prior ignorance it is suggested that $\pi(k, \beta) = \pi(k)\pi(\beta)$ where $\pi(k)$ is a uniform distribution between 1 and $\min(n, 200)$ and $\pi(\beta)$ is a half-normal distribution with large variance. Using a standard MCMC strategy, any proposed move to a new classifier from the current parameter settings (β, k) to new settings (β', k') is accepted if u, a draw from a $U[0, 1]$ distribution, is less than $\min\left\{1, \frac{L(D|\beta',k')\pi(\beta',k')}{L(D|\beta,k)\pi(\beta,k)}\right\}$, and otherwise the current values of β and k are retained.

We now illustrate the envelope method of assessing classifier confidence, and show how it relates to the standard reject method. To do this we apply both methods to a number of data sets, but in order to avoid problems that arise with the reject method in multi-class data we consider only two-class sets here. One of these sets is a synthetic set devised by Ripley (1994) and augmented with a further Gaussian function: it thus comprises five Gaussian components, 3 contributing to one class and 2 to the other (full details are given in Fieldsend *et al*, 2003). The other four sets are from the UCI Machine Learning Repository; they are the Wisconsin, Ionosphere, Pima, and Sonar data sets respectively. The data-set details are given in Table 1 (number of predictors, p; size of training set, D; size of test set, T). Also shown in the table are the overall classification performances, i.e. the percent correct classification of the test set, for each data set.

Table 1: Data set details and overall classification performances

Data Set	p	D	T	% correct
Pima	8	512	256	77.1
Synthetic	2	250	1000	88.6
Sonar	60	138	70	87.1
Ionosphere	33	200	151	94.7
Wisconsin	9	455	228	99.6

To compare the envelope and reject methods we need to compare the proportions each method assigns to the three categories SURE CORRECT (SC), SURE INCORRECT (SI) and UNSURE (U). In order to do this we have found the proportions assigned to each of the three categories by the envelope method at each of three commonly used threshold values (0.80, 0.95 and 0.99). To standardise the two methods we have then found the assignments to the three categories for which the reject method gives the same SI proportion as the envelope method (apart from the Pima 95% region where we standardised on the SC proportions), together with the reject threshold value that achieves this assignment. In a couple of cases there was a range of such values, and in these cases we have quoted the highest value in the range. All the results are given in Table 2 for each of the five data sets.

In terms of proportions within each category (SC, U, SI), the two methods of region construction give very comparable results. Where values differ between the two methods for a category, the better (i.e. lower SI or higher SC) value is shown in bold. However, whereas the envelope regions are defined in terms of familiar "confidence coefficient" values, the best matching reject regions have unpredictable and generally very low probability thresholds. We see that if we require strong consistency of classification (99%) then success rates (SC) show a fall from the unconditional rates in Table 1, but if we are prepared to tolerate weaker consistency then there is generally a closer match between the rates.

Table 2: 80%, 95% and 99% envelope method regions and best matching reject method regions for five data sets

	Envelope Regions				Reject Regions			
Data	#	SC	U	SI	#	SC	U	SI
Pima	80%	0.7656	0.0260	0.2083	51%	0.7656	0.0260	0.2083
	95%	0.7552	0.0521	**0.1927**	53%	0.7552	0.0469	0.1979
	99%	0.7448	0.0677	0.1875	54%	**0.7552**	0.0573	0.1875
Synthetic	80%	**0.8780**	0.0160	0.1060	54%	0.8760	0.0180	0.1060
	95%	**0.8740**	0.0270	0.0990	57%	0.8710	0.0300	0.0990
	99%	**0.8700**	0.0320	0.0980	57%	0.8680	0.0340	0.0980
Sonar	80%	0.8429	0.0286	0.1286	51%	0.8429	0.0286	0.1286
	95%	0.8286	0.0429	0.1286	51%	**0.8429**	0.0286	0.1286
	99%	0.8000	0.0857	0.1143	52%	**0.8429**	0.0429	0.1143
Ionosphere	80%	0.9470	0.0066	0.0464	51%	0.9470	0.0066	0.0464
	95%	0.9470	0.0066	0.0464	51%	0.9470	0.0066	0.0464
	99%	0.9338	0.0199	0.0464	51%	**0.9470**	0.0066	0.0464
Wisconsin	80%	0.9781	0.000	0.0219	93%	0.9781	0.000	0.0219
	95%	0.9781	0.000	0.0219	93%	0.9781	0.000	0.0219
	99%	0.9781	0.000	0.0219	93%	0.9781	0.000	0.0219

4 Choosing Between Classifiers

4.1 Methodology

Traditionally, classifier system performance has been measured simply by the percentage of test-set allocations that are correct (or by its complement, the error rate, or some simple variant depending on problem-specific variation in the importance of the alternative classifications). Thus whenever a choice has to be made between competing classifiers, either the success rate or the error rate is the criterion on which the decision is based. But this statistic carries no information regarding the confidence with which the various classifications have been made. We have argued above for the use of SURE CORRECT, SURE INCORRECT and UNSURE as measures of confidence in classifications, so a better comparison between classifiers should be based on simultaneous use of all these measures. To see how this can be implemented, we draw on the work that has been done in classifier acceptance-reject rates (see, e.g., Giacinto, Roli and Bruzzone, 2000, for a summary). In particular, Battiti and Cola (1994) have shown that to compare the performance of different classifiers we need to compare their

accuracies over a range of different rejection rates (i.e. different threshold values t), and this can be done by plotting these values in the accuracy-rejection (A-R) plane. In our case the UNSURE proportions at different t values correspond to the rejection rates, while "accuracy" is reflected by either of the SURE categories. We prefer to minimise SURE INCORRECT rather than maximise SURE CORRECT, so to compare different classifiers on a data set we compare the curves each produces when SURE INCORRECT is plotted against UNSURE for a range of values of t. The classifier corresponding to the lowest curve on such a plot is the one to be chosen.

4.2 Applications

To illustrate this methodology, we first need a set of classifiers to compare. There is an almost unlimited choice available to us, but to keep within a traditional statistical modelling framework we define a nested set of $k-$nearest neighbour classifiers by providing first a simplification and then a generalization of the probabilistic classifier introduced above.

The simplified version is obtained by keeping β fixed at 1.0 throughout, and only sampling over k. Here the probability of $\boldsymbol{x}$ belonging to a particular group is directly proportional to the preponderance of this group among the k nearest neighbours of $\boldsymbol{x}$, and there is thus no possibility of skewing this probability as the balance of neighbours between groups varies. We call this version the "simple" classifier as opposed to the other "standard" one.

The generalized version is obtained by expanding the single β parameter into a matrix $\boldsymbol{M}$ of parameters to reflect scaling and rotation of the variables. This is equivalent to replacing the Euclidean metric $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^t(\boldsymbol{x}_1 - \boldsymbol{x}_2)\}^{1/2}$ in step 1 of the $k-$nearest neighbour process by an "adaptive" metric $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^t\boldsymbol{M}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\}^{1/2}$ where the (positive-definite) matrix $\boldsymbol{M}$ is chosen to optimise the classification with regard to differential scales and orientations of the variables. Various ways can be devised for achieving such an adaptive classifier (see, e.g., Myles and Hand, 1990, or Hastie and Tibshirani, 1996). Our approach is to take $\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^t$, where $\boldsymbol{\Lambda}$ is a diagonal scaling matrix and $\boldsymbol{Q} = \exp(\boldsymbol{S})$ with $\boldsymbol{S}$ a skew-symmetric rotation matrix. The proposals are generated by forming

$\boldsymbol{M}' = \boldsymbol{Q}'\boldsymbol{\Lambda}'\boldsymbol{Q}'^t$, where quantities r_i drawn independently from $N(0, 0.2^2)$ are added to the diagonal elements of $\boldsymbol{\Lambda}$ to give $\boldsymbol{\Lambda}'$, and $\boldsymbol{Q}' = \exp(\boldsymbol{S}')$ where quantities s_i are drawn independently from $N(0, 0.1^2)$ and added to elements of $\boldsymbol{S}$ to give $\boldsymbol{S}'$. Further details are given by Everson and Fieldsend (2004); we call this version the "adaptive" classifier.

It is evident that the three classifier versions are therefore nested, with the simple one being a special case of the standard one and this in turn being a special case of the adaptive one.

4.2.1 Comparison of envelope and reject methods

To make this comparison as simply and directly as possible, we compare the two methods on just the simple Bayesian $k-$nearest neighbour classifier (i.e. only one parameter k) and the standard Bayesian $k-$nearest neighbour classifier (i.e. two parameters k and β) on the five data sets used above; Figure 1 shows the accuracy-rejection plots for these data sets. The plots obtained using the envelope method are on the left, and those using the reject method are on the right; the curve obtained from the simple classifier is indicated by crosses, that from the standard classifier by open circles.

Figure 1 about here.

The first obvious difference between the envelope plots and the reject method plots is that the latter stretch across the whole $x-$axis while the former generally stop about half-way across. This is because the envelope plots are determined by the proportion of MCMC classifiers that classify to each group, and in all data sets there will be at least some points for which all classifiers allocate to one group. Such points have posterior group allocation probabilities of 1.0 so can never be categorised as UNSURE whatever the threshold value of t − even if their estimated posterior probabilities of classification are not very high. The reject method plots, on the other hand, are based directly on these estimated posterior probabilities which rarely approach 1.0 for any data points. Hence the range of possible UNSURE values is much greater for the reject method than for the envelope method, and this feature is borne out by the plots. Indeed for some envelope plots

the range of UNSURE values is either very short or nonexistent (e.g. for the Wisconsin data). Reference back to Table 2 shows that for these data sets there are either no or very few UNSURE points at the highest threshold value, so there cannot be any such points at lower threshold values.

With that proviso, it is evident that the differences between the two methods of construction are very slight, and that they both give the same qualitative conclusions regarding the comparison between the simple and the standard $k-$nearest neighbour classifiers. Since the simple classifier is nested within the standard one we would expect the latter to have better classification performance, and this is generally the case in our examples. For the Pima, Synthetic and Sonar data sets the curve for the standard classifier lies distinctly below that for the simple classifier (although there is a small reversal at the lowest UNSURE value of the Sonar data). In the cases of Ionosphere the two classifiers give indistinguishable performances, the two curves virtually coinciding over the range plotted, while the Wisconsin data (as already noted) has virtually no variability for either classifier.

We have stressed earlier the ease of application of the envelope method to multi-class data, since the basic operation is no different from that in the two-class case. We therefore selected two more data sets from the UCI repository: the Wine data with 3 classes ($p = 13, D = 89, T = 89$, 96.6% correct classification) and the Vehicle data with 4 classes ($p = 19, D = 564, T = 282$, 67.4% correct classification). The accuracy-rejection plots for these sets are shown in Figure 2 for the envelope method.

Figure 2 about here.

The two classifiers give virtually indistinguishable performances for the vehicle data, the two curves lying more or less on top of each other, but for the wine data we have the surprising result that the simple classifier has better performance than the standard classifier. However, this set of data is one with small data sets, relatively high number of variables and well-separated classes, so a difference of one or two classifications is enough to cause the result observed.

Table 3: Percentage of correct classifications in the test set for each classifier and each data set

	Classifier		
Data Set	simple	standard	adaptive
Pima	76.0	77.1	79.2
Synthetic	87.9	88.6	89.4
Sonar	85.7	87.1	84.3
Ionosphere	94.7	94.7	98.0
Wisconsin	99.6	99.6	98.7
Vehicle	63.8	67.4	77.0
Wine	98.9	96.6	98.9

4.2.2 Comparison of classifiers

We can now turn to comparison of the three versions of $k-$nearest neighbour classifier. First, we show in Table 3 the overall classification performances of each of these versions as judged by the percentage of correct classifications in the test set T of each data set.

Although there are one or two exceptions evident in the table, the broad trend of the results suggests that classifier accuracy improves on moving successively from simple to standard to adaptive, i.e. as the complexity of the k-nearest neighbour classifier increases. (Although the details are not shown here, the Bayesian-averaging classifier also generally gives better results than just the single-best MAP classifier.) However, we have argued above that such a way of judging classifier performance is too simplistic, and that we need to examine the SURE INCORRECT versus UNSURE plots of the classifiers over a range of values of t. In Figure 3 we therefore show these plots for the test data portion of each of the seven data sets. In each plot the simple classifier is indicated by crosses, the standard classifier by open circles, and the adaptive classifier by stars.

Figure 3 about here.

The picture now is less clear-cut than the error rate comparisons would suggest. The only data set in which the above trend is definitely supported is the Pima data, where the curve for the simple classifier lies completely above the curve for the standard classifier, and this in turn lies mostly above

the curve for the adaptive classifier. Although the standard classifier curve is not completely above that for the adaptive classifier, it is nevertheless so for a sufficient part of the range of UNSURE values, so that we can indeed conclude that for this data set the adaptive classifer is best, the standard classifier is next best, and the simple classifier is the poorest. We note that the test set for the Pima data is quite large (256 individuals).

The remaining data sets depart from the expected trend to a greater or lesser extent. Closest is the Synthetic data, where the simple classifier is uniformly the poorest again, but there is nothing to choose between the other two types. However, it is possible to establish the optimal Bayes error rate for Synthetic data; in this case both standard and adaptive versions are operating at close to the Bayes level, and such a large test set (1000 observations) permits accurate estimation of classification rates. For the Vehicle data there is in fact nothing to choose between all three types until near the very end of the range of UNSURE values, so by analogy with the Synthetic data result we infer that all three classifiers are operating at close to the Bayes level. We also note that the Vehicle test set is the second largest among our data sets (282 individuals). The Wisconsin data has very little variation across the whole range of classifiers, so can perhaps be discounted, while the Wine and Sonar data exhibit so many "cross-overs" of curves as to make any single conclusion meaningless. However, we note that these latter two data sets have very small test samples (89 and 70 respectively), so these "cross-overs" are a reflection of the large variability in small data sets. The one puzzling outcome is for the Ionosphere data, which show the reverse of the expected trend with the most complex classifier being the poorest until near the end of the UNSURE range. This result is opposite to the one suggested by consideration of the straightforward correct/incorrect dichotomy and would merit further investigation.

We therefore conclude from these experiments that although the addition of a confidence measure to the usual correct/incorrect assessment of classifiers is highly desirable, it carries a penalty in terms of sampling variability. The expected trends show up only generally when samples are large (particularly when test samples are large), and in small samples the picture

is considerably less clear.

5 Conclusion

We have shown that Bayesian MCMC methodology can be allied to existing knowledge on the reject option in classification to produce a quantification of the confidence that can be ascribed to particular classification outcomes. One point that can be made here is that a typical Bayesian MCMC classification task gathers a vast amount of information, much of which is thrown away without further use. The envelope method makes use of this information, so is very efficient in that it needs little more computation than is already carried out but with considerable added benefit.

Of the two methods described, the envelope approach offers some direct advantages over the standard reject approach: interpretability in terms of familiar confidence coefficient terminology, guidance in choice of threshold values, and easy applicability to all types of grouping. However, it offers a further compelling advantage that has not been mentioned so far, namely that it is applicable to those classifiers (such as the linear discriminant function, for example) that produce non-probabilistic classification scores whereas the reject method cannot be used in such cases.

It has been shown that incorporating confidence measures into a comparison of classifiers via the accuracy-rejection plots can make the comparison less clear-cut than the traditional one based solely on either success or error rates. This is related to the variability inherent in sample-based classifiers, which is often ignored when making error rate comparisons. A more realistic assessment might come from comparisons of confidence regions for error rates (see, e.g., Krzanowski, 2001), but this does not yet seem to be standard practice.

Indeed, it is surprising that classifier confidence has received so little attention, considering the emphasis placed on confidence regions in general statistical practice. The methods described here are easily built in to a standard Bayesian procedure so should be part of the general classification tool kit, especially in such areas as safety-critical applications. However,

some aspects remain to be investigated. For example, what can be done if the available data are not extensive enough to be split into a training set D and a test set T? The standard way of proceeding in the single-classifier case would be to use a data-based method of error rate estimation such as leave-one-out, but in our set-up each unit omission in effect creates a new set D for the MCMC process. So if such a scheme were to be contemplated then an efficient way of organising the computations would be essential. Similar considerations of efficiency are paramount if the variability of the results is to be established using, say, $n-$fold cross-validation on the (D, T) splits of the data.

While such aspects remain to be investigated, we nevertheless feel that use of the confidence measures described in this paper provide a distinct step forward in classifier technology.

Acknowledgement

This work has been supported by grant no. GR/R24357/01 of the Engineering and Physical Science Research Council.

References

BATTITI, R. and COLLA, A.M. (1994), "Democracy in neural nets: voting schemes for classification," *Neural Networks*, 7, 691-707.

BISHOP, C.M. (1995), *Neural Networks for Pattern Recognition*, Oxford: Clarendon Press.

BROOKS, S.P. (1998), "Markov chain Monte Carlo method and its application," *The Statistician*, 47, 69-100.

CHOW, C.K. (1970), "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, 16, 41-46.

DE GROOT, M.H. (1970), *Optimal Statistical Decisions*, McGraw-Hill.

DENISON, D.G.T., HOLMES, C.C., MALLICK, B.K. and SMITH, A.F.M. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, Chichester: John Wiley & Sons Ltd.

EVERSON, R.M. and FIELDSEND, J.E. (2004), "A variable metric probabilistic $k-$nearest neighbours classifier," *Proceedings of the Fifth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'04)*, Lecture Notes in Computer Science, vol 3177, pp 659-664, Springer.

FIELDSEND, J.E., BAILEY, T.C., EVERSON, R.M., KRZANOWSKI, W.J., PARTRIDGE, D. and SCHETININ, V. (2003), "Bayesian inductively learned modules for safety critical systems," *Proceedings of the 35th Sypposium on the Interface: Computing Science and Statistics*, pp 110-125, Salt Lake City.

FUKUNAGA, K. (1990), *Introduction to Statistical Pattern Recognition*, Morgan Kaufmann.

GIACINTO, G., ROLI, F. and BRUZZONE, L. (2000), "Combination of neural and statistical algorithms for supervised classification of remote-sensing images," *Pattern Recognition Letters*, 21, 385-397.

GREEN, P.J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711-732.

HAND, D.J. (1997), *Construction and Assessment of Classification Rules*, Chichester: John Wiley & Sons Ltd.

HASTIE, T. and TIBSHIRANI, R. (1996), "Discriminant adaptive nearest neighbor classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18, 607-616.

HOETING, J.A., MADIGAN, D., RAFTERY, A.E. and VOLINSKY, C.T. (1999), "Bayesian model averaging: a tutorial (with discussion)," *Statistical Science*, 14, 382-417.

HOLMES, C.C. and ADAMS, N. (2002), "A probabilistic nearest-neighbour method for statistical pattern recognition," *Journal of the Royal Statistical Society Series B*, 64, 1-12.

KRZANOWSKI, W.J. (2001), "Data-based interval estimation of classification error rates," *Journal of Applied Statistics*, 28, 585-595.

LINDLEY, D.V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference*, Cambridge: University Press.

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

MYLES, J.P. and HAND, D.J. (1990), "The multi-class metric problem in nearest neighbour discrimination rules," *Pattern Recognition*, 23, 1291-1297.

RIPLEY, B.D. (1994), "Neural networks and related methods for classification (with discussion)," *Journal of the Royal Statistical Society Series B*, 56, 409-456.

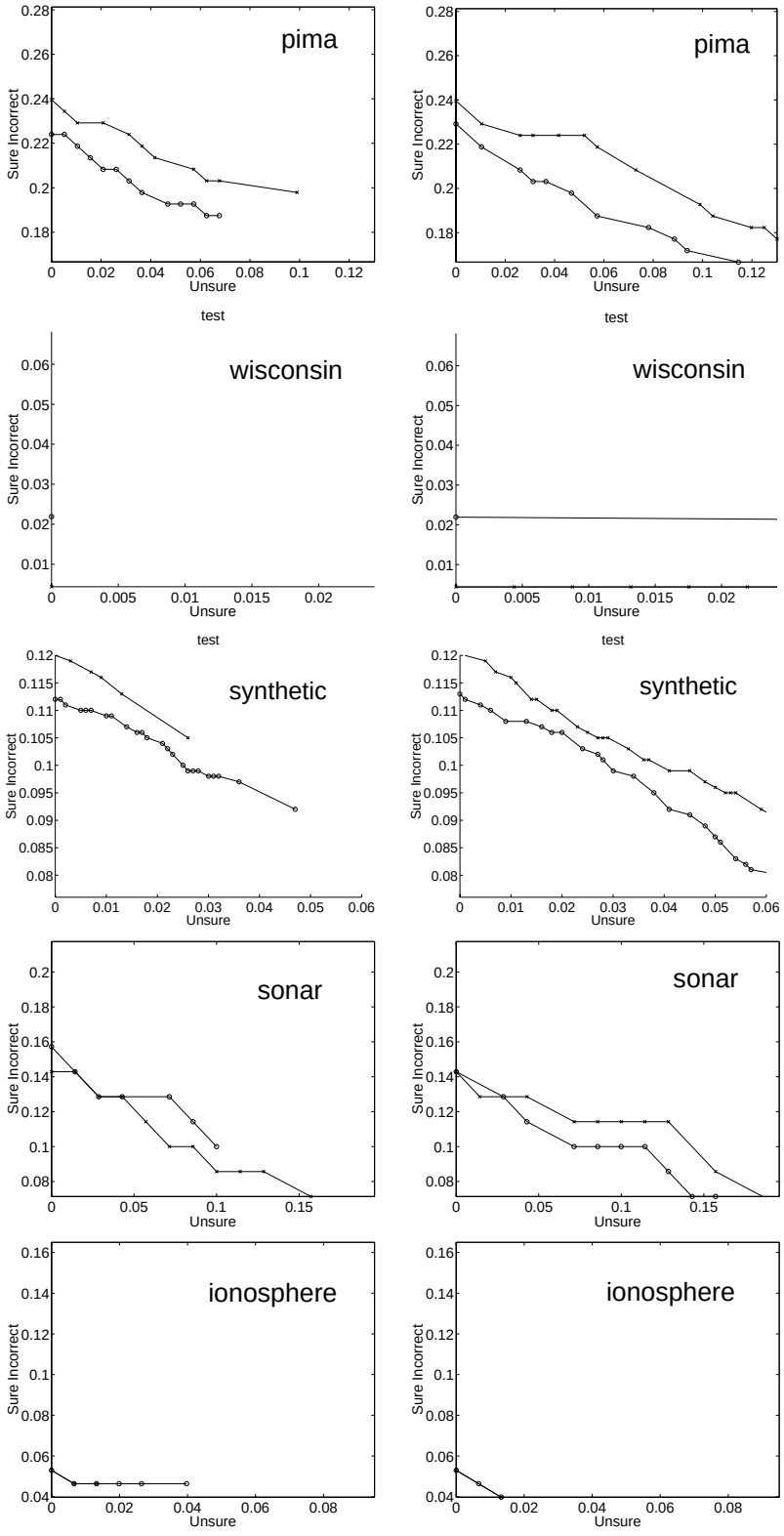

Figure 1: Accuracy-rejection plots for the 2-group data sets; envelope method on left, reject method on right, circles for general classifier and crosses for simple classifier.

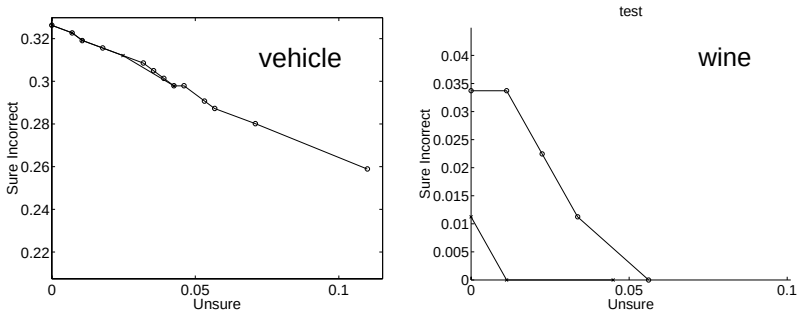

Figure 2: Accuracy-rejection plots for the multi-group data sets using the envelope method; circles for the general classifier, crosses for the simple classifier.

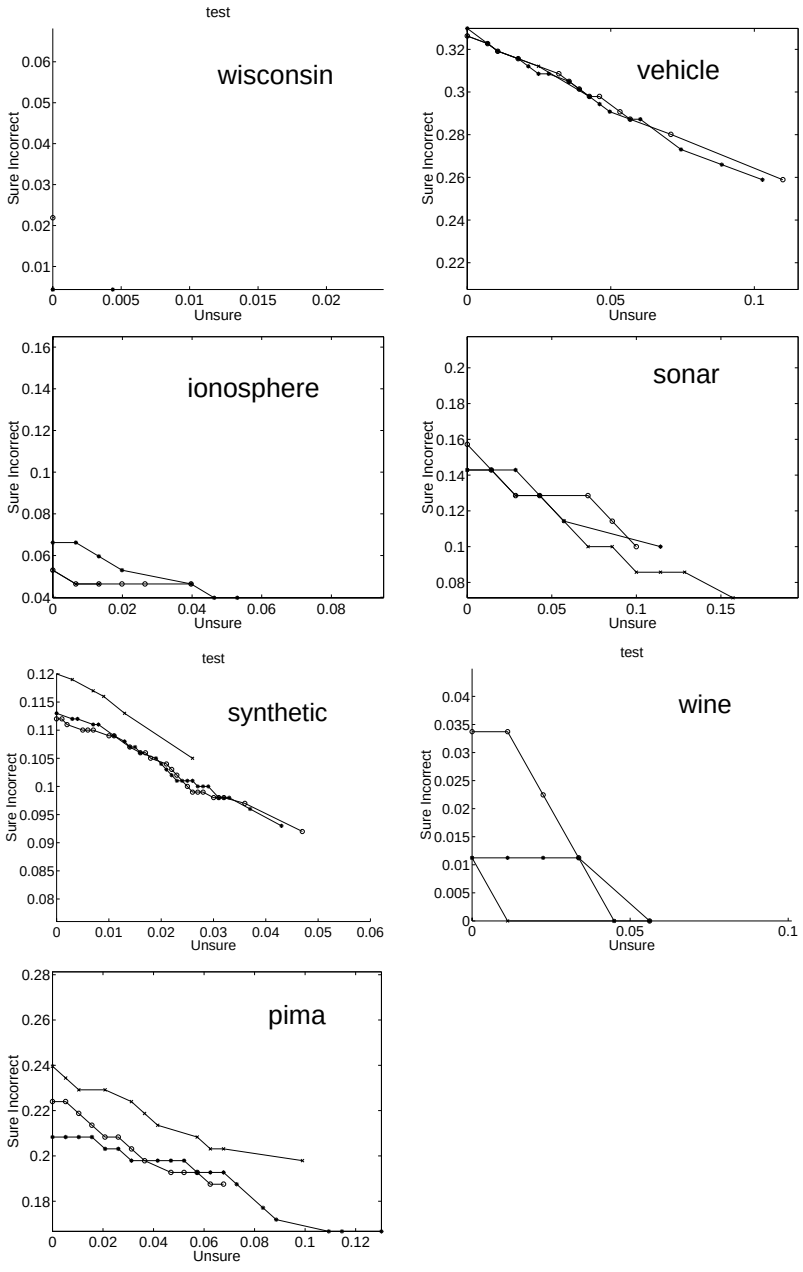

Figure 3: Accuracy-rejection plots for all data sets using the envelope method; crosses for the simple classifier, circles for the general classifier, stars for the adaptive classifier.