PERGAMON

Pattern Recognition 32 (1999) 1139–1147

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

Comparing classifiers when the misallocation costs are uncertain

N.M. Adams, D.J. Hand*

Department of Statistics, The Open University, Watton Hall, Milton Keynes, MK7 6AA, UK

Received 9 January 1998: received in revised form 28 October 1998: accepted 28 October 1998

Abstract

Receiver Operating Characteristic (ROC) curves are popular ways of summarising the performance of two class classification rules. In fact, however, they are extremely inconvenient. If the relative severity of the two different kinds of misclassification is known, then an awkward projection operation is required to deduce the overall loss. At the other extreme, when the relative severity is unknown, the area under an ROC curve is often used as an index of performance. However, this essentially assumes that *nothing* whatsoever is known about the relative severity – a situation which is very rare in real problems. We present an alternative plot which is more revealing than an ROC plot and we describe a comparative index which allows one to take advantage of anything that may be known about the relative severity of the two kinds of misclassification. © 1999 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: ROC curve; Error rate; Loss function; Misclassification costs; Classification rule; Supervised classification

1. Introduction

An 'optimum' classification rule is one which minimises the expected future loss when it is used to classify objects. The amount of loss will depend on the costs of the different kinds of misclassification and the probabilities with which they occur. In this paper we restrict ourselves to just two classes and assume that the processing costs involved in producing a classification are the same for all objects. Without loss of generality, we can take this constant cost to be zero. We further assume that those objects which are correctly classified incur no additional costs, and denote the additional costs associated with *mis*classifying a class 0 object by c_0 and a class 1 object by c_1. The overall expected future loss is then simply

$$L = \pi_0 f_0 c_0 + \pi_1 f_1 c_1, \tag{1}$$

where f_i is the probability of misclassifying a class i object, and π_i is the probability that an object comes from class i. Parameter estimation, model selection, and performance assessment can be based on minimising this overall loss. So also can choice of classification threshold. This is the value t such that the object is to be classified into class 0 whenever $\hat{p}(0|x) > t$. Here $\hat{p} = \hat{p}(0|x)$ is the estimated probability that an object with measurement vector x will belong to class 0. A standard result shows that minimum loss is achieved by choosing the classification threshold such that points are classified into class 0 if $\hat{p} > t = c_1/(c_0 + c_1)$.

*Corresponding author. Tel.: +44 1908 655974; Fax: +44 1908 65 2140; E-mail: d.j.hand@open.ac.uk

PII: S0031-3203(98)00154-X

All of this is fine in principle, but in practice it is rare for the costs of misclassifying objects to be known precisely. This might be because it is simply difficult to quantify them (in medical diagnostic problems, for example), or it might be because costs evolve over time (in banking applications, for example, where the cost incurred by an incorrect classification depends on external economic conditions [6]). This paper describes a measure of performance for situations in which the costs are not known exactly.

Error rate or misclassification rate is often used as a comparison criterion for comparing classifiers [4]. Superficially this requires no choice of costs, but in fact it merely makes the implicit assumption that the costs of the two types of misclassification are equal. This is just as much an assumption as any other. Indeed, it is perhaps worse than many other assumptions, since, as we have argued elsewhere [1], we believe that the assumption of equal misclassification costs is seldom realistic.

When costs cannot be specified, an alternative strategy for measuring performance is to compare the distribution of $\hat{p}$ for objects from class 0 with the distribution of $\hat{p}$ for objects from class 1. We denote these two distributions by $f(\hat{p}|0)$ and $f(\hat{p}|1)$, respectively. In fact, of course, we do not know these distributions and must estimate them from a test set (or by some other more sophisticated means; since such issues are peripheral to the purpose of this paper, we shall simply adopt an independent test set) yielding estimates $\hat{f}(\hat{p}|0)$ and $\hat{f}(\hat{p}|1)$. The greater the separation between these two distributions the more likely it is that, for an arbitrary choice of classification threshold, good classification performance will be achieved, though this is, of course, not guaranteed. Many statistical measures of the separability between distributions have been defined. They include the Kolmogorov–Smirnov statistic, the t-statistic, Chernoff distance, Lissack–Fu distance, and the Wilcoxon two group test statistic. The last of these is an estimate of the probability that a randomly chosen member of class 0 will have a smaller value of $\hat{p}$ than a randomly chosen member of class 1. (It is defined as the sum of the ranks associated with the test set objects in class 0 when the test set elements are ranked in terms of $\hat{p}$.) We will be especially concerned with this measure because of this attractive interpretation. However, by very virtue of the fact that it compares the overall distributions of $\hat{f}(\hat{p}|0)$ and $\hat{f}(\hat{p}|1)$, it is equivalent to aggregating losses over all possible choices for the classification threshold, and hence over all possible choices for the costs. As well as being an advantage, this is also a weakness.

The fact is that, although precise costs may not be known in any given real application, it is very likely that *something* will be known about the cost ratio. For example, it would be a rare problem for which one could not assert that c_0/c_1 was either not 0 or not ∞. It follows that, in any real application, it is unlikely that nothing whatsoever would be known about the classification threshold – hence it is inappropriate to summarise over *all* possible values of this threshold. As a measure of separability suitable for classification rules, then, the Wilcoxon statistic is too loosely defined.

The Wilcoxon test statistic is closely related to a popular graphical display for showing performance of classification rules. A *receiver operating characteristic*, or ROC curve (see [2] for a comprehensive overview) plots the proportion of class 0 points correctly classified into class 0 against the proportion of class 1 points (incorrectly) classified into class 0 as the classification threshold is varied. That is, it uses the cumulative distributions $\hat{F}(\hat{p}|0)$ and $\hat{F}(\hat{p}|1)$ as vertical and horizontal axes, respectively and is parameterised by $\hat{p}$. Fig. 1 illustrates such a curve for the Pima Indians data to be described below. This curve has the characteristic shape of ROC curves, starting from the bottom left of the diagram (when no objects are classified into class 0) and ending at the top right (when all objects are classified into class 0). A classification rule which is no better than chance would produce a curve lying along the diagonal of the ROC square – then $\hat{F}(\hat{p}|0)$ and $\hat{F}(\hat{p}|1)$ would be identical. A perfect classification rule would produce a curve which followed the left-hand side and top edges of the square – $\hat{F}(\hat{p}|0)$ would reach 1 while $\hat{F}(\hat{p}|1)$ remained at 0. (In this case the value of $\hat{p}$ at the top left-hand corner of the ROC square would define a threshold t such that all of the $\hat{p}$ values for class 0 would be below t while none of the $\hat{p}$ values for class 1 would be below t.) The area of the square beneath the ROC curve (denoted AUC) is a popular measure of separability of two distributions – and

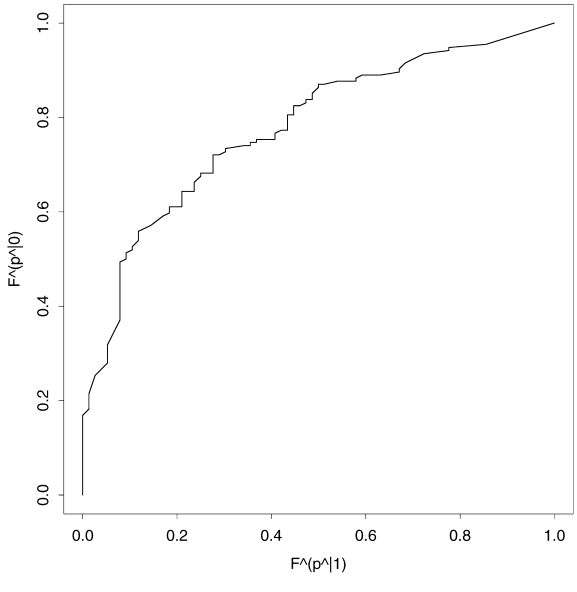

Fig. 1. An example of an ROC curve.

hence of the performance of a classification rule. Formally, the AUC is given by

$$\mathrm{AUC} = \int \hat{F}(\hat{p}\,|\,0)\, d\hat{F}(\hat{p}\,|\,1) = \int \hat{F}(\hat{p}\,|\,0)\, \hat{f}(\hat{p}\,|\,1)\, d\hat{p}.$$

It is a simple transformation of the *Gini* index (twice the area between the curve and the diagonal [5]) and is an estimate of the probability that a randomly drawn member of class 0 has $\hat{p}$ value less than that of a randomly drawn member of class 1 – as is the Wilcoxon two sample test statistic described above.

The ROC curve also permits one to see the loss associated with any choice of classification threshold and any ratio of costs. For cost ratio c_0/c_1 the loss associated with any point (x, y) on the ROC curve is given by the length of the projection of that point onto the line through (0,1) as origin and with slope $-\pi_1 c_1/\pi_0 c_0$. For a given cost ratio, the value of the classification threshold which minimises the loss corresponds to that position on the ROC curve for which this projection is the smallest.

This property of the ROC curve immediately reveals why AUC is not an ideal measure of performance. If the ROC curves of two classifiers cross, then one classifier will be superior for some values of the cost ratio and the other classifier will be superior for other values. If it were known that the actual cost ratio for a problem led to a classification threshold which lay to one side or the other of the crossing point, even if one did not know the precise position of the actual cost ratio, then a summary measure integrating over all possible values of the cost ratio would be worthless. It could tell us that one classifier was superior while, in fact, for the *relevant* values of the cost ratio, the other one was better.

In this paper we describe an alternative to the AUC which takes advantage of what is known about the likely values of the cost ratio.

An in-depth discussion of performance measures of classification rules is given in [1] where several different types of measure are identified. Measures based on loss are termed 'inaccuracy' measures and those based on overall comparisons of the distributions of $\hat{f}(\hat{p}\,|\,0)$ and $\hat{f}(\hat{p}\,|\,1)$ are termed 'resemblance' measures.

2. Loss difference plots

We commented in the preceding section that it is unlikely that nothing would be known about the relative costs of the two kinds of misclassification. To say that nothing is known means that one cannot even assert that one kind of misclassification is more serious than the other. Although there are two misclassification costs associated with two class problems, since the threshold which minimises loss in Eq. (1) is given by $t = c_1/(c_0 + c_1) = [1 + (c_0/c_1)]^{-1}$, only the *ratio* of misclassification costs is relevant. Furthermore, in our experience, acquiring cost values from domain experts is naturally done in terms of the ratios – how much more serious one type of misclassification is than the other. Thus an expert will often be happy to say that the cost of misclassifying a class 0 object is, say, roughly three times or five times as serious as the reverse. In what follows, we go further and require the expert to specify an interval within which he or she is confident that the true or future cost ratio will lie, even though they are unable to specify it precisely. Thus, for example, an expert may be able to say that misclassifying a class 1 object is between 2 and 10 times as serious as the reverse – that is, that the cost ratio c_0/c_1 lies in the interval [0.1,0.5]. In the extreme case in which all that can be said is that misclassifying a class 0 object is less serious than misclassifying a class 1 object, the expert would report that the ratio lies within the interval [0.0,1.0]. Likewise, if the expert was prepared only to say that misclassifying a class 0 object was *more* serious than the reverse, the interval would be $[1.0, \infty]$.

The asymmetry of these intervals, for what are essentially equivalent converse situations, is rather troubling. Moreover, when we come to produce plots below, it is, to say the least, inconvenient having a parameter which goes to ∞. For these reasons, although we elicit cost information in terms of the cost ratio, we will produce plots in terms of $c_1/(c_0 + c_1) = [1 + (c_0/c_1)]^{-1}$. The intervals [0.0,1.0] and $[1.0, \infty]$ above are mapped to the symmetric intervals [0.5, 1] and [0, 0.5] by this transformation. In general, $c_0/c_1 = a$ maps to $1/(1 + a) = b$ (say) while $c_0/c_1 = 1/a$ maps to $a/(1 + a) = 1 - b$. Since only the ratio of c_0 to c_1 is relevant to $c_1/(c_0 + c_1)$, we can arbitrarily rescale c_0 and c_1 without affecting the result. In particular, we can choose that rescaling such that $(c_0 + c_1) = 1$.

We could now plot loss, as calculated in Eq. (1) above, against c_1 (with c_1 ranging from 0 to 1). However, the losses due to different cost pairs (c_0, c_1) are not comparable: it does not make sense to claim that the loss due to cost pair (5,1) is greater than that due to the pair (4,3) for example. On the other hand, if one classifier has a loss curve which lies above that of another classifier in certain ranges of c_1, then the first classifier is worse, in terms of loss, in those ranges. For this reason, rather than plotting loss against c_1, to compare classifiers we simply plot the sign of the difference in losses due to the classifiers against c_1. This leads to a partition of the [0,1] range of c_1 into segments in which the classifiers alternate in terms of superiority.

3. Eliciting costs

We have already noted that in our experience domain experts find it convenient to specify costs in terms of the ratio c_0/c_1, and that they can usually give an interval I in

which they are confident this ratio must lie. Further, they can also usually specify a value they consider to be most likely. We use these three points – the ends of the interval I and the most likely value – to specify a belief distribution for likely values of c_0/c_1. For simplicity, we take the form of this distribution to be triangular, with support I on the c_1-axis, and with apex occurring at the point corresponding to the most likely value. Thus, if the interval of feasible values for the cost ratio c_0/c_1 is given as $I = [a, b]$ and the most likely value for the cost ratio is m, then the distribution is taken to be a triangle with end points at $c_1 = 1/(1 + a)$ and $c_1 = 1/(1 + b)$ and with apex at $c_1 = 1/(1 + m)$. The height of the triangle is $h = 2(1 + a)(1 + b)/(b - a)$, determined such that its area is 1. We denote this distribution on c_1 by $D(c_1)$.

One might object to the arbitrary choice of a triangular form for this distribution – although perhaps agreeing that a unimodal form is almost always appropriate. Likewise, one might also argue that a triangular form defined on the cost ratio axis c_0/c_1 rather than on the c_1-axis would be more natural, since the relative costs have been elicited in terms of this axis. If one accepts this, then a triangular distribution on the c_0 axis would be inappropriate by virtue of the nonlinearity of the transformation $c_1 = [1 + (c_0/c_1)]^{-1}$. We are sympathetic to these arguments but suggest that general uncertainty over the relative probabilities for the different values of c_0/c_1 render such arguments irrelevant. The shape of the distribution, whether on c_0/c_1 or on c_1, cannot be determined (in almost all situations, anyway) with sufficient precision to distinguish between a triangular form on c_1 and the transformation of a triangular form on c_0/c_1. To put it another way, while we accept that our proposal is not ideal, it seems to us that the popular alternatives described in Section 1 (of assuming that the cost ratio is known precisely or that nothing whatsoever is known about the cost ratio) are more extreme and even less ideal.

4. The LC index

We now have all the elements to define the new index, which we shall call the *LC index* (for loss comparison). In comparing two classifiers A and B, we know on which parts of the c_1 interval each is superior. Define a function $L(c_1)$ taking the value $+1$ in regions of the c_1 [0,1] interval for which A is superior (meaning that its loss plot lies beneath that of B) and -1 for regions in which B is superior (the loss plot of A lies above that of B). We also know how confident we are that any particular value of c_1 will occur, given by $D(c_1)$. The integral of the product $L(c_1)D(c_1)$ over [0,1] thus gives us a measure of confidence that A will yield a smaller loss than B. This is the LC index. It ranges from -1 to $+1$, taking positive values when classifier A is more likely to lead to a smaller loss than classifier B and negative values when the converse applies. A value of $+1$ means that A is certain to be the superior classifier, since A is superior for all feasible values of c_1.

5. Examples

In this section we compare the LC index with the AUC for four examples, two from medicine (obtained from the UCI repository of machine learning databases [3] and two from the area of financial credit. These examples have been chosen to illustrate the variety of index values which can be produced, and the interpretation which should be put on them. The first two examples have been chosen since they are widely used as comparative data sets in supervised classification problems. By virtue of this we have no real cost intervals available. We therefore chose some which we thought appropriate for the problem, and which would illustrate particular aspects of LC index interpretation. The third and fourth examples arose from our own work analysing banking data sets. The cost intervals we used were obtained in consultation with the bankers who provided us with the data.

5.1. Pima indians data

The data consist of 768 observations on people belonging to a particular tribe of Indians, 268 who tested positive for diabetes (class 1) and 500 (class 0) who tested negative. The values of eight variables are recorded for each. The data were randomly divided into a design set of 538 and a test set of 230. (Questions of whether such a division is an ideal thing to do are not relevant here, since our aim is merely to illustrate how classification rules built using these data can be compared.)

Fig. 2 shows ROC curves for quadratic discriminant analysis and a neural network classifier. The AUCs are 0.7781 and 0.7793, respectively, suggesting that, if anything, the neural network classifier is superior, but that there is little to choose between the two methods. However, a glance at the figure shows that things are not this simple. The two ROC curves cross. For some values of costs one classifier will be superior, while for other costs the other will be superior. Moreover, since, in order to see which is superior for a given cost pair it is necessary to carry out the awkward projection, it is not easy to see what these costs are. Suppose that one is confident that the ratio of the costs of the two kinds of misclassification, c_0/c_1, lies in the interval $[0.1, 0.25]$, so that misclassifying a diabetic as non-diabetic is thought to be between 4 and 10 times as serious as the reverse. Suppose also that it is thought most likely that misclassifying a diabetic is 7 times as serious as the reverse. This corresponds to a ratio of 0.14 on the c_0/c_1 scale. These three values correspond to a c_1 interval of $[0.80, 0.91]$ with a most likely value of 0.875. Fig. 2 is not very enlightening.

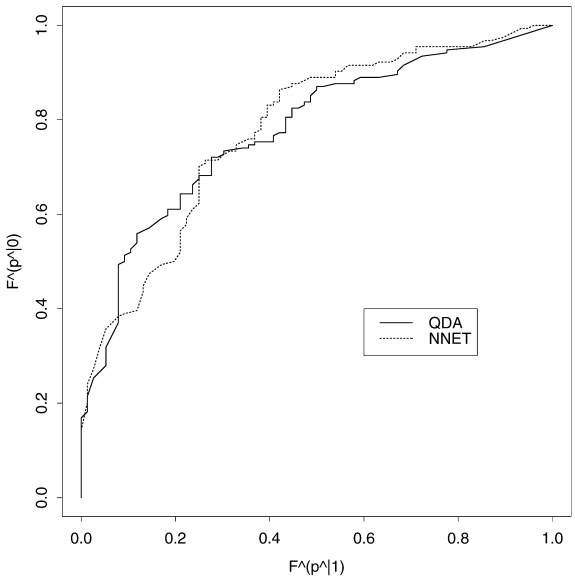

Fig. 2. ROC curves for Pima Indians data.

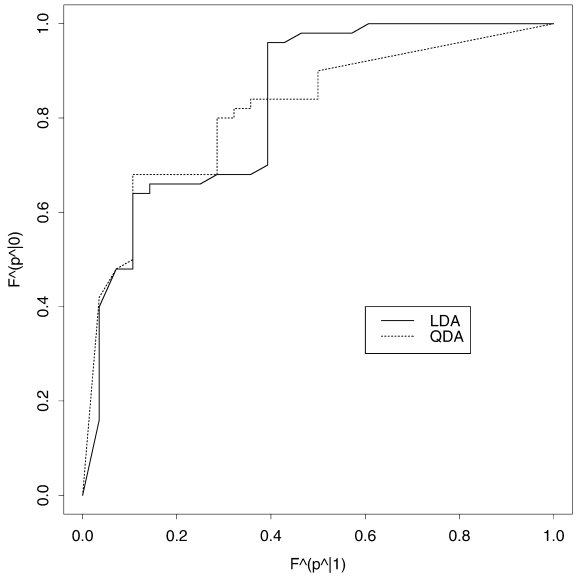

Fig. 4. ROC curves for Hungarian heart disease data.

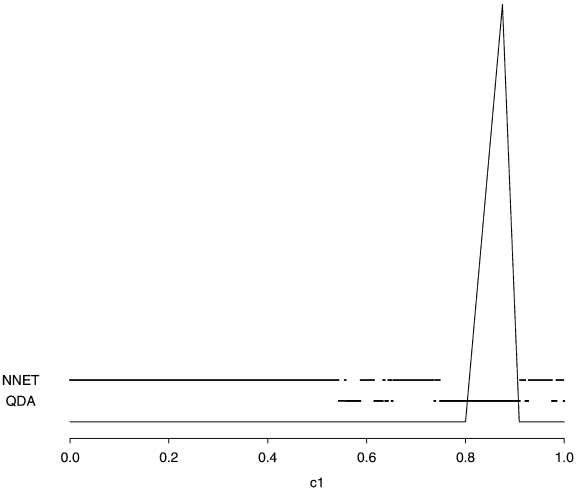

Fig. 3. The cost distribution and relative superiority of the two classifiers applied to the Pima Indians data.

Fig. 3 is a much more useful representation. The horizontal axis shows c_1, ranging from 0 to 1. The bold line segments above this axis show the intervals of c_1 in which the neural network method is superior and the intervals in which quadratic discriminant analysis is superior. The triangle standing on this axis has interval of support [0.80,0.91] and apex occurring at $c_1 = 0.875$. The LC index, computed as described above, is -1.0. This suggests that the quadratic method is superior. This thus gives a message in the opposite direction to the AUC, and much more definitively.

5.2. Hungarian heart disease data

These data were collected at the Hungarian Institute of Cardiology, Budapest, by Andros Janosi, MD, and are available from the heart disease directory of the UCI data archive. They consist of 261 subjects, 163 of whom did not have heart disease (class 0) and 98 had heart disease (class 1), each measured on 14 variables. They were randomly divided into a design set of 183 subjects and a test set of 78 subjects.

Fig. 4 shows the ROC curves for linear discriminant analysis and quadratic discriminant analysis for these data. The AUCs are 0.8311 and 0.8143 respectively. Again the superficial inference that linear discriminant analysis should be adopted is cast into doubt by the fact that neither curve dominates the other. The 'better' method will depend on the unknown costs.

Suppose, however, that our domain expert tells us that the cost of misclassifying a diseased subject is between 2 and 20 times as serious as misclassifying a healthy subject and that the most likely value is 16 times as serious. These values correspond to c_1 ranging between 0.667 and 0.952, with a most likely value of 0.937.

Our corresponding plot is shown in Fig. 5. The bars show regions in which each of quadratic and linear discriminant analysis is superior. The triangle shows the belief distribution for c_1. The corresponding value of the LC index is 0.89. This suggests that quadratic discriminant analysis is the better method, contradicting the conclusion reached using the standard AUC measure.

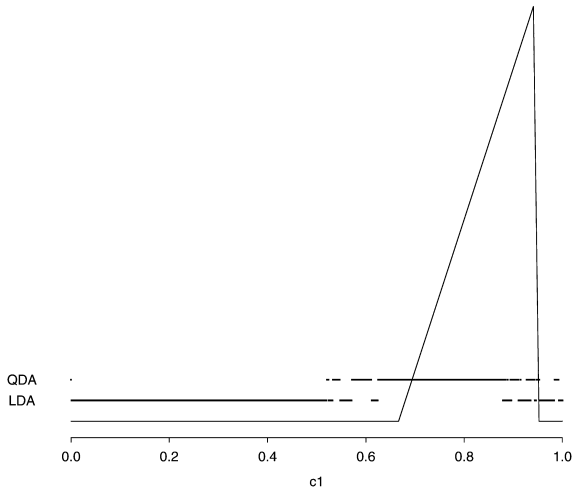

Fig. 5. The cost distribution and relative superiority of the two classifiers applied to the heart disease data.

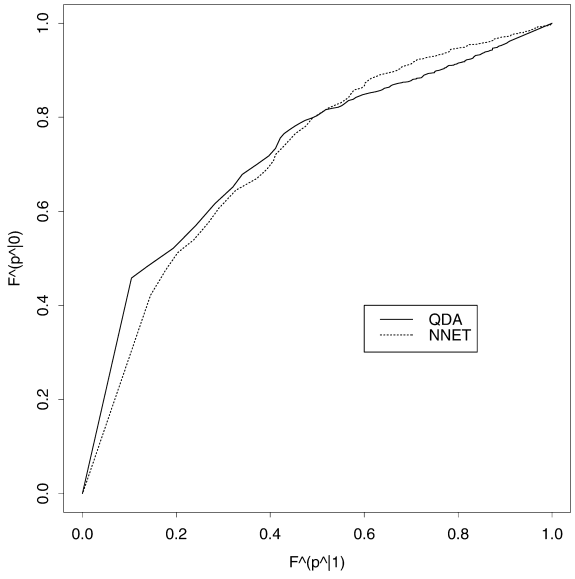

Fig. 6. ROC curves for credit card marketing data.

5.3. Marketing strategies

This data set was provided by a major UK credit card company, and is concerned with classifying customers according to their likely response to a promotional scheme. In particular, the company is interested in classifying people into one of two classes according to the pattern of interest payments they are likely to make in the future (since the details are commercially sensitive, we have avoided defining the classes more precisely here). We shall denote the classes as 1 and 0, where the former is the class thought likely to return a profit. The classes had priors of $\pi_0 = 0.87$ and $\pi_1 = 0.13$. The prediction into likely future classes is to be made on the basis of 25 variables, mainly describing earlier credit card transaction behaviour. 8000 records were available, and we split them equally into training and test sets.

In collaboration with the banking experts, we developed a model for the costs of the two kinds of misclassification, based on such factors as cost of manufacture and distribution of marketing material, cost due to irritation caused by receiving junk mail, and loss of potential profit by not mailing a potential member of class 1. All of the factors in the model were given as intervals, and from these was derived an interval of possible values for the overall cost ratio c_0/c_1 as being [0.065, 0.15], with the most probable value 0.095.

In this example, we compare quadratic discriminant analysis (QDA) and a neural network classifier. The latter had 13 nodes in its single hidden layer, and used weight decay to avoid overfitting. Both network architecture and penalty term were chosen by cross-validation. The ROC curves for the test data are shown in Fig. 6. The AUCs for these curves are 0.7244 and 0.7102 for the quadratic and neural network classifiers respectively. Thus the AUC suggests that quadratic discriminant analysis is slightly preferable. Fig. 7 shows the corresponding plot produced by our method. From this, the LC index is -0.4. This also suggests that the quadratic method is to be preferred. That is, when one places more appropriate emphasis on the likely values of the cost ratio, and de-emphasises those values thought to be inappropriate, the method suggests more clearly that the quadratic classifier should be used. Although the evidence is perhaps stronger than the AUC method provides, it is still

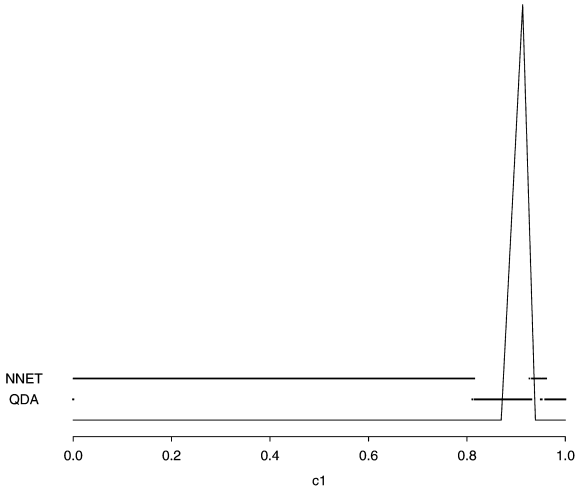

Fig. 7. The cost distribution and relative superiority of the two classifiers applied to the credit card marketing data.

not overwhelming – an absolute value of 0.4 is far from one of 1.0.

5.4. Banking data

These data were provided by a large UK bank. They refer to an aspect of current account usage and consist of 1292 good risks (class 0) and 208 bad risks (class 1). The training set had 1000 and the test set had 500 observations. ROC curves for linear discriminant analysis and the nine-nearest-neighbour method are presented in Fig. 8. The AUC for linear discriminant analysis is 0.7116

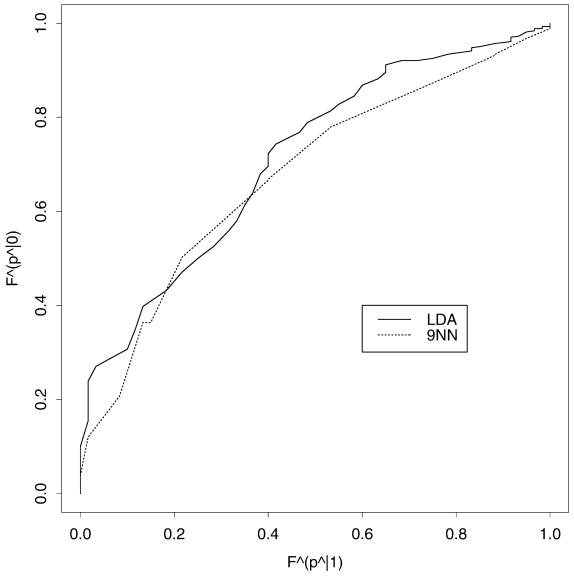

Fig. 8. ROC curves for banking data.

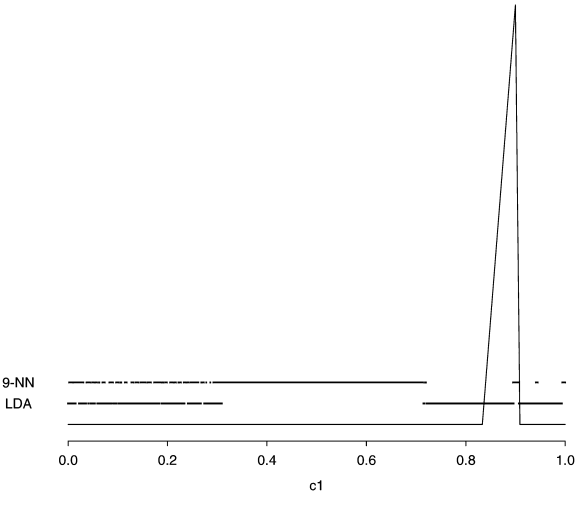

Fig. 9. The cost distribution and relative superiority of the two classifiers applied to the banking data.

and that for the nearest-neighbour method is 0.6745, suggesting that linear discriminant analysis has the edge. Experts in the bank give misclassifying a bad risk as 5–10 times as serious as the reverse. With a most likely value of 9 times as serious, Fig. 9 shows our plot. The LC index is -0.64, showing that linear discriminant analysis is also to be preferred using our criterion. However, the conclusion is now not as clear-cut as it might be. In circumstances such as this, as a referee pointed out, it may well be that relatively small alterations to the form of the belief function D could change one's confidence (or even conclusion) about which is the better classifier. The absolute size of the LC index serves as an indication of how confident one should be in the conclusion about which classifier is to be preferred.

6. Conclusion

The ROC curve and associated AUC statistic, equivalent to the Gini index and Wilcoxon test statistic, are often used as criteria by which to compare classification rules. However, they have major disadvantages. It is extremely awkward to use a pair of ROC curves to see for which cost pairs one classifier is superior to another, requiring a sophisticated projection operation. Only in the case that one classifier dominates another will the AUC be universally valid in a comparison of classifiers. In general, the AUC aggregates over all possible values of the ratio of the costs of the two kinds of misclassification. This is unrealistic, since almost always *something* will be known about the relative costs, even if they are not known precisely. We have proposed an alternative, the LC index, which makes use of what can be said about the relative sizes of the costs. In some situations this knowledge will include the fact that the interval of feasible values for the cost ratio is such that, within this interval, one classifier dominates (even though it may not dominate over the entire range of cost ratios). When this occurs there will be no doubt about which classifier is superior – and the LC index will take value $+1$ or -1. In other cases the feasible interval will include segments where each of the classifiers is superior, and in such cases the LC index will indicate which classifier is most likely to be better – and which should therefore be chosen. The size of the LC index will also indicate how much confidence one should have in the conclusions. A value near zero, though indicating preference for one classifier or the other, is an indication that departures from the assumptions necessarily made in constructing the index may lead to a reversal of the conclusion.

The LC index is a comparative and not an absolute index of performance. An absolute index is impossible because of the meaninglessness of combining losses corresponding to different cost pairs. That is, it does not make sense to argue that the cost pair $(c_0, c_1) = (0.1, 0.9)$

is better or worse than the cost pair $(c_0, c_1) = (0.2, 0.8)$. For the same reason, the LC index should not be interpreted as an expected loss. Although we have used a 'probability' distribution over likely values of the costs, the losses at these different values cannot be compared and hence cannot be combined to yield an expected value. The best that can be done is as we have done here and determine whether classifier A, say, is more likely to be favoured than classifier B.

Although the LC index is an improvement on the ROC plot and AUC, further improvements are possible. In particular, our model assumes that the costs of misclassifying individuals from any one class are all the same. This is unlikely to be the case in practice and problems in which misclassification costs vary from object to object, even for objects in the same class, is the focus of our current work.

Other workers have also recognised the importance of effective criteria for comparing classification rules. For example, Bradley [7] discusses the use of the area under the ROC curve as a performance measure in some detail. He presents an extensive series of applications to real data, in which he compares five different forms of classification rule. He notes the danger of extrapolating the results to different kinds of problems, and also avoids 'expert bias' by deliberately not tuning the algorithms to the specific problems. Analysis of variance and multiple range tests are used to summarise the performance differences between the methods. Provost and Fawcett [8] have also considered the problem of unknown costs. They defined the convex hull of a set of ROC curves in terms of the projection described in Section 1, so that the curve(s) producing minimum cost could be identified for any values of the costs. Wieand et al. [9] and McClish [10] also used a probability distribution or interval through which to weight differences between rules. However, they were interested in a rather different problem and defined their distribution over $f(\hat{p}\,|\,1)$ rather than the cost ratio.

This paper has focused on measures of performance and has not discussed estimating the standard deviation of those measures, and yet such indicators of precision are clearly important if confidence is to be placed in comparative statements about classification rules. This is the subject of current work. It is complicated by the fact that there are two sources of variation in the distribution of $\hat{p} = \hat{p}(0\,|\,x)$. One is that arising due to the fact that the test set is sampled. The other is that arising from the fact that the estimates are based on a design set which is itself sampled. For classifiers built using a given data set only the first source is relevant, but for studies aimed at general recommendations about which methods to use, both are relevant.

The fact that, in any real study, *something* will generally be known about the relative severity of the different kinds of costs indicates the limited value of many of the theoretical comparisons which have been conducted outside any practical context. Such studies tend to use AUC (or, perhaps worse, error rate) because they have no context from which to produce sensible (ranges of) cost ratios. But this very fact makes their conclusions of limited value. It does not make sense to ask 'which is the better/best rule' in general. It only makes sense to ask such a question in a particular problem context. The LC index has weaknesses, but these are perhaps not as severe as those of the popularly used alternatives.

7. Summary

Receiver operating characteristic, or ROC curves are popular ways of summarising the performance of two class classification rules, and are widely used in pattern recognition, epidemiology, and signal detection theory. The area under such curves (AUC) is often used as a measure of quality of the associated classification rule. This measure is equivalent to the *Gini* coefficient, and also to the Wilcoxon two sample test statistic. However, both ROC curves and the AUC measure have major disadvantages. In particular, deducing the relative performance of two classification rules for a given ratio of misclassification costs requires that an awkward projection operation be applied to the ROC curve, while the AUC integrates performance over all possible values of this ratio. This latter is generally inappropriate since, in any real problem, *something* will be known about the relative misclassification costs – even if it is only that one type of misclassification is more serious than the other. To overcome these problems we present an alternative plot which permits comparison between two classifiers, directly showing the values of the costs for which each is superior. Associated with this, we define a comparative measure of classifier performance, defined in terms of a feasible cost interval and the most likely value of the cost ratio. This measure, the LC index, takes values between $+1$ and -1, according to which of two classification rules is to be preferred. We illustrate with some real examples, showing that the conclusions reached using the LC index can be opposite to those reached using the common AUC measure.

Acknowledgements

The work of Niall Adams on this project was supported by grant GR/K55219 from the EPSRC, under its 'Neural networks – the key questions' initiative. Equipment was provided by grant CSM/2006/050 from the DRA. The paper was written while the second author was visiting the Isaac Newton Institute for the Mathematical Sciences in Cambridge as part of the Neural Networks and Machine Learning Programme. We are

grateful to Andrew Webb for helpful discussions and comments, to Gordon Blunt of Barclaycard for providing both the data and domain expertise for Section 5.3, and to Sam Korman of Abbey National for providing both the data and domain expertise for Section 5.4.

References

[1] D.J. Hand, Construction and Assessment of Classification Rules. Wiley, Chichester, 1997.
[2] J.P. Egan, Signal Detection Theory and ROC Analysis. Academic Press, New York, 1975.
[3] C.J. Merz, P.M. Murphy, UCI depository of machine learning database, 1996. [http://www.ic.uci.edu/mlearn/MLRepository.html].
[4] D. Michie, D.J. Spiegelhalter, C.C. Taylor, Machine Learning, Neural and Statistical Classification, Ellis Horwood, New York, 1994.
[5] D.J. Hand, W.E. Henley, Statistical classification methods in consumer credit scoring: a review, J. Roy. Statist. Soc. Ser. A 160 (1997) 523–541.
[6] D.J. Hand, S. Jacka (Eds.), Statistics in Finance, Edward Arnold, London, 1998.
[7] A.P. Bradley, The use of the area under the ROC curve in evaluation of machine learning algorithms, Pattern Recognition 30 (1997) 1145–1159.
[8] F. Provost, T. Fawcett, Analysis and visualisation of classifier performance: comparison under imprecise class and cost distributions, in: D. Heckerman, H. Mannila, D.P.R. Uthurusamy (Eds.), Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining, (1997) pp. 43–48.
[9] S. Wieand, M.H. Gail, B.R. James, K.L. James. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data, Biometrika 76 (1989) 585–592.
[10] D.K. McClish, Analyzing a portion of the ROC curve, Med. Decision Making 9 (1989) 190–195.

About the Author—NIALL ADAMS completed his Ph.D. in Statistical Computing at Liverpool John Moores University, UK. He joined the Open University Department of Statistics as a research fellow in October 1995. His primary research interests involve classification methodology, in particular the impact of misallocation costs.

About the Author—DAVID HAND has been Professor of Statistics at the Open University, UK since 1988. His methodological research interests include multivariate statistics, especially classification problems. His most recent book on the topic is *Construction and Assessment of Classification Rules* (Wiley, 1997). His applications interests include finance, medicine, and psychology.