Realisable Classifiers: Improving Operating Performance on Variable Cost Problems.

M.J.J. Scott, M. Niranjan, R.W. Prager,
Cambridge University Department of Engineering
Trumpington Street, Cambridge CB2 1PZ
mjjs@eng.cam.ac.uk

Abstract

A novel method is described for obtaining superior classification performance over a variable range of classification costs. By analysis of a set of existing classifiers using a receiver operating characteristic (ROC) curve, a set of new *realisable classifiers* may be obtained by a random combination of two of the existing classifiers. These classifiers lie on the convex hull that contains the original ROC points for the existing classifiers. This hull is the maximum realisable ROC ($MRROC$).

A theorem for this method is derived and proved from an observation about ROC data, and experimental results verify that a superior classification system may be constructed using only the existing classifiers and the information of the original ROC data. This new system is shown to produce the $MRROC$, and as such provides a powerful technique for improving classification systems in problem domains within which classification costs may not be known *a priori*. Empirical results are presented for artificial data, and for two real world data sets: an image segmentation task and the diagnosis of abnormal thyroid condition.

1 Introduction

A large fraction of decision support systems, particularly those used in medical diagnostics (e.g. diagnosis of cancer with digital mamography), are two-class pattern classification systems. Once a set of features and the functional form of the classifier have been chosen, the classifier is designed to optimise some cost function. When the costs of the different types of errors can be specified exactly, the optimum classifier may be designed to minimise the expected risk [5]. The particular feature set and the functional form chosen then define how well the performance of the classifier approaches the Bayes' performance.

In many real world applications, however, that cost of different types of errors is often not known at the time of designing the classifier. One also finds applications where the costs might change over time. Further, some costs cannot be specified quantitatively. In such situations we resort to specifying the classifier in the form of an adjustable threshold and a receiver operating characteristic (ROC) curve obtained by setting the threshold to various possible values. An example of such an ROC curve is shown in Figure 1. In the example, the classifier must classify a patient's condition as either adverse or benign.

BMVC 1998 doi:10.5244/C.12.31

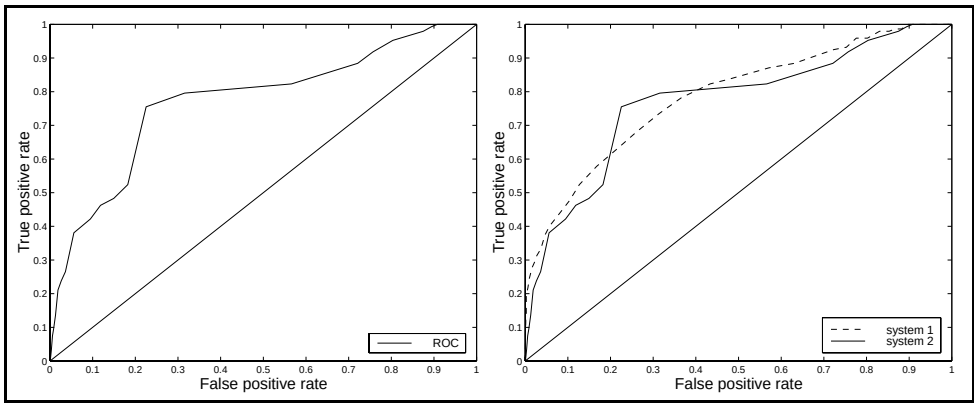

Figure 1: Left: an ROC curve for a medical diagnostic test for abnormal thyroid condition. The true positive rate corresponds to the probability that a sick patient will be diagnosed as sick, the false positive to the probability that a healthy patient will be diagnosed as ill. Right: the ROC curves of two classification systems cross.

The data in the example was obtained from the UCI Machine Learning repository [12], and represents the results of a number of diagnostic tests for abnormal thyroid conditions. A linear classifier was used, producing a continuous output, and a threshold is placed upon the output to determine the final classification. Two rates can be calculated for any series of classifications: the true positive and false negative rates. When an adverse case is *correctly* classified as adverse, a true positive has occurred, and a false negative when a benign case is *incorrectly* classified as adverse. By varying the level of the threshold, different degrees of true positive and false positive rates can be achieved, producing the ROC.

The ROC curve has been shown to be a useful mechanism in comparing the performance of different classifiers [7, 10, 9, 11, 18]. The area under the ROC ($AUROC$) curve is also known as the Wilcoxon statistic [7, 10]. Lovell *et al*, for example, use this statistic as a criterion for feature selection in a large obstetrics problem involving 48 features and $700,000$ cases.

$AUROC$, however, is a gross simplification of the information conveyed by a ROC, as noted by Hand in [7]. The costs of different operating points need to be taken into consideration. Hand further suggests this to be an important factor when the ROC curves of the classification systems that are being considered cross. Such an example is illustrated in Figure 1.

This paper describes a novel approach for combining classifiers to achieve desirable operating points that do not fall on any of the ROC curves of the available classification systems. More specifically, we show that a convex hull may be formed, encapsulating the 'best' operating points of many classification systems. We provide a theorem to show that an operating point on the convex hull is realisable in practice. Empirical results are provided for three classification tasks: artificial, medical diagnosis, and image segmentation.

2 Realisable classifiers

2.1 An observation about points in ROC space

We think of an ROC curve as representing curve joining a set of points in an ROC space. A point (fp_c, tp_c) represents an existing classifier c, that classifier producing false positives with a probability $Pr(\text{falsepositive} = fp_c)$, and true positive with probability $Pr(\text{truepositive} = tp_c)$.

Take two classifiers, c_a and c_b, each with distinct false positive and true positive rates. These two classifiers are the end points of a straight line in ROC space, L_{ab}. The line L_{ab} defines a set of classifiers, i.e. point $(fp_{c_x}, tp_{c_x}) \in L_{ab}$ represents the classifier that would produce those true positive and false positive rates.

We observe in this paper, that, given only c_a and c_b, one may realise the output of classifier c_x by randomly choosing between the output of c_a and c_b. The probability of choosing the output of c_a over that of c_b is determined by the distance along L_{ab} between c_x and c_a[1].

Theorem 1 *The realisable classifier. Two existing classifiers, c_a and c_b, produce true positive and false positive rates (tp_a, fp_a) and (tp_b, fp_b) respectively for a series of m inputs $x_1..x_m$. In a 2 dimensional plot of false positive rate against true positive (ROC space), call the straight line linking (fp_a, tp_a) and (fp_b, tp_b) L_{ab}.*

Any point (fp_r, tp_r) on L_{ab} corresponds to the point that would be produced by a classifier r. Call the set of classifiers corresponding to n points on L_{ab}, $\mathbf{R} = \{r_1, .., r_n\}$.

Given c_a and c_b, the output of a realisable classifier, $r_i \in \mathbf{R}$, for any input x_j, is a random variable that assumes the output of one or other of c_a and c_b with probability

$$\begin{aligned} Pr(r_i(\cdot) = c_b(\cdot)) &= \frac{fp_{r_i} - fp_a}{fp_b - fp_a} \\ Pr(r_i(\cdot) = c_a(\cdot)) &= 1 - Pr(r_i(\cdot) = c_b(\cdot)), \end{aligned}$$

where fp_{r_i} is the false positive rate of r_i.

The proof of Theorem 1 is straightforward. To construct the output of a realisable classifier $r_i$with false positive rate fp_{r_i}, randomly select between the outputs of c_a and c_b with the given probability. The expected false positive rate produced by doing so is

$$\begin{aligned} E[fp] &= Pr(r_i(\cdot) = c_b(\cdot)) * fp_b + Pr(r_i(\cdot) = c_a(\cdot)) * fp_a \\ &= \frac{fp_{r_i} - fp_a}{fp_b - fp_a} * fp_b + (1 - \frac{fp_{r_i} - fp_a}{fp_b - fp_a}) * fp_a \\ &= \frac{fp_b(fp_{r_i} - fp_a) + fp_a(fp_{r_i} - fp_a) - fp_a(fp_b - fp_a)}{fp_b - fp_a} \\ &= \frac{(fp_b - fp_a)(fp_{r_i} - fp_a) + fp_a(fp_b - fp_a)}{fp_b - fp_a} \\ &= fp_{r_i} \quad Q.E.D. \end{aligned}$$

[1] This technique has parallels in classical statistics. When estimating the power of a hypothesis test, the sample space of which has discrete probabilities, randomised decision rules could be employed. This allowed the estimation of specific power values, even when an observed estimate was unavailable [6, 17], in the context of using k-fold cross validation techniques to produce accurate ROC curves when data is scarce.

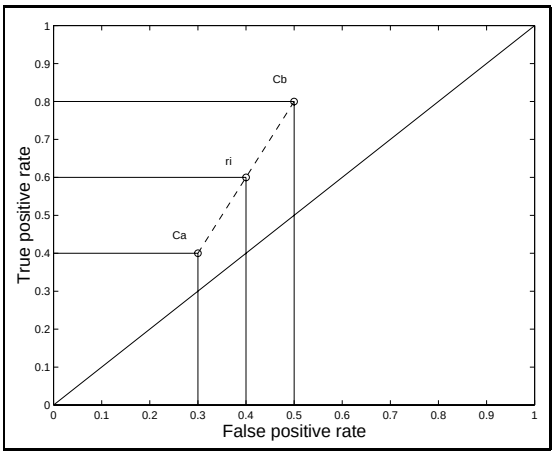

Figure 2: An example of a realisable classifier. The point r_i on the line joining c_a and c_b may be realised by the application of Theorem 1

And similarly for the true positive rate.

Figure 2 illustrates an example of a realisable classifier. The realisable classifier r_i, with false positive rate $fp_{r_i} = 0.4$, lies on the line between classifiers c_a and c_b, with false positive rates $fp_a = 0.3$, and $fp_b = 0.5$ respectively. To realise the output of r_i, calculate the probabilities for selecting the outputs of the existing classifiers using Theorem 1,

$$\begin{aligned} Pr(r_i(\cdot) = c_b(\cdot)) &= \frac{fp_{r_i} - fp_a}{fp_b - fp_a} \\ &= \frac{0.4 - 0.3}{0.5 - 0.3} \\ &= 0.5 \\ Pr(r_i(\cdot) = c_a(\cdot)) &= 1 - Pr(r_i(\cdot) = c_b(\cdot)) \\ &= 0.5. \end{aligned}$$

To obtain the classification output of r_i on a set of unseen cases, $\mathbf{x} = \{x_1, .., x_n\}$, the classifications of c_a and c_b would be calculated

$$\begin{aligned} c_a(\mathbf{x}) &\rightarrow \{(x_1 = \text{Adverse}), (x_2 = \text{Adverse}), (x_3 = \text{Benign}), .., (x_n = \text{Adverse})\} \\ c_b(\mathbf{x}) &\rightarrow \{(x_1 = \text{Benign}), (x_2 = \text{Adverse}), (x_3 = \text{Adverse}), .., (x_n = \text{Benign})\}. \end{aligned}$$

Using the probabilities calculated above, the output of r_i is then determined by randomly selecting one of the outputs, like so:

$$\begin{aligned} r_i(\mathbf{x}) &\rightarrow \{c_a(x_1), c_b(x_2), c_a(x_3), .., c_b(x_n)\} \\ r_i(\mathbf{x}) &\rightarrow \{(x_1 = \text{Adverse}), (x_2 = \text{Adverse}), (x_3 = \text{Benign}), .., (x_n = \text{Benign})\}. \end{aligned}$$

2.2 The maximum realisable ROC

We can now realise all classifiers that lie on straight line segments with end points formed by existing classifiers. What advantage can be gained by this? Take the example illustrated in Figure 3. The ROC is produced using a linear model on the 1 dimensional

classification problem shown. The steps in the ROC occur because the linear model cannot capture the multi modal nature of the data.

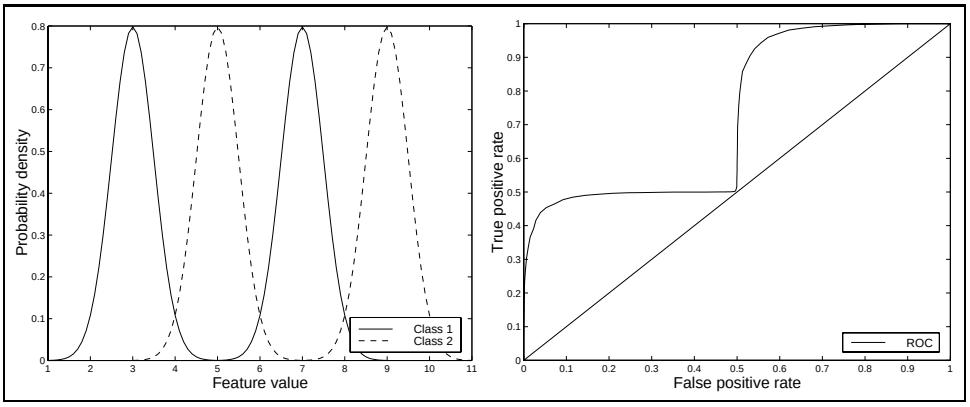

Figure 3: Left: an example of 2 class multi modal data. This data will present problems to linear classification systems. Right: the ROC curve produced by varying the threshold on the output of a linear model of the multi modal data shown. The ROC has a step like appearance because the linear model fails to capture the nature of the data.

We currently have a set of classifiers provided by the linear model, corresponding to the range of false positive rates from 0 to 1. The current ROC curve is produced by this set, and can be used to select the best available classifier for a particular false positive rate. It is possible, however, to obtain a new set of classifiers that will give better performance in terms of true positive rates than those provided by the linear model.

Calculate a convex hull [14] such that it contains all the points on the current ROC. The vertex points of the convex hull will be points corresponding to existing classifiers generated by the linear model. The facets of the hull are line segments with an existing classifier at each end point. We know from Theorem 1 that all the points on these lines represent realisable classifiers. It is immediately obvious that a realisable classifier r, with false positive rate fp_r, lying on a facet of this hull will have a greater true positive rate than the classifier with false positive rate fp_r found on the original ROC.

Given a classification algorithm such as the linear model, and the ROC curve produced by this, then the convex hull enclosing this ROC represents a set of realisable classifiers that will at all times be either equal or superior to those of the linear model, and that are generated by a subset of the original classifiers. The convex hull describes the maximum realisable ROC ($MRROC$) given the available existing classifiers.

3 Experimental results

3.1 Artificial data

Multi modal data was generated for the 1 dimensional, 2 class classification problem of Figure 3. A linear model was trained using 5000 training examples. By varying the threshold used on the output of the model when presented with 5000 test cases, the ROC curve of Figure 3 was obtained. The true positive rate was the rate of correct classifications

of class 1, the false positive rate was the rate of cases of class 2 being incorrectly classified as belonging to class 1..

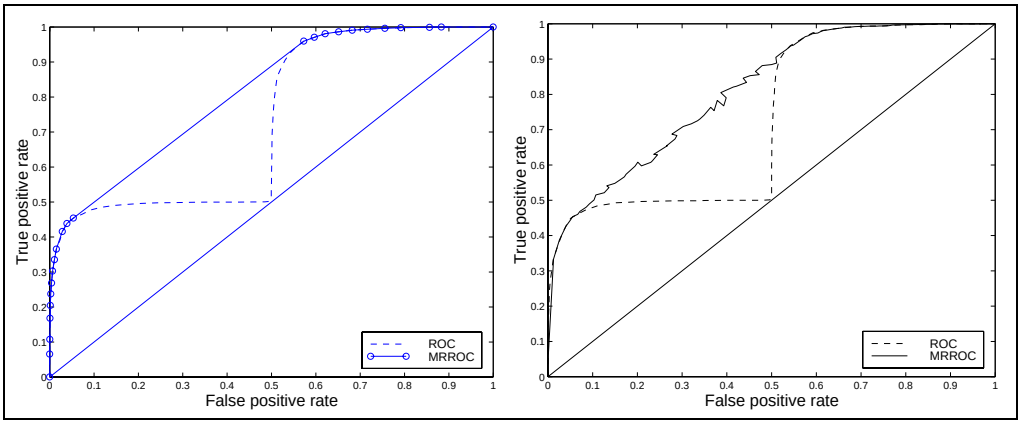

Figure 4: Left: the convex hull containing the ROC of a linear model is found. This hull is the $MRROC$ of the set of realisable classifiers produced from the set of existing linear classifiers. Right: the $MRROC$ plotted for an unseen data set. The $MRROC$ is consistent with that predicated.

Using the qhull software [1] the convex hull containing all the points in the ROC was obtained. Each vertex in the hull represented an existing classifier. Each of these existing classifiers was defined by the threshold used, on the output of the linear model, to yield a final classification for each test case. The hypothesis to being tested was that all the points on the facets of the convex hull, corresponding to classifiers that were not currently available, could be realised by application of Theorem 1 of this paper to the set of vertex classifiers. The $MRROC$ indicates the expected true positive rates, over the complete range of false positive rates, that one might hope to achieve using this approach.

Figure 4 plots the $MRROC$ over the ROC of the linear model on the test data.

To validate the hypothesis that the characteristic curve indicated by the $MRROC$ could actually be obtained, a third data set of 5000 validation cases was generated. This validation data was processed by the linear model. The thresholds corresponding to the existing classifiers in the convex hull were each applied to the outputs of the linear model, producing a number of sets of classifications. For any point on a facet of the hull, a classification for an individual validation case could be obtained by randomly selecting one of the classifications made by the two existing classifiers at the end points of the facet. As described above, this methodology leads to the realisation of the set of classifiers on the facets of the hull.

Figure 4 plots the characteristic curve given the set of realisable classifiers indicated with the $MRROC$ against the ROC of the linear model, on the validation data set. It can clearly be seen that the set of realisable classifiers produce an ROC consistent with the $MRROC$, and superior to the ROC of the linear model. The $MRROC$ appears slightly jagged. This is entirely consistent with the nature of the classifiers used to form it. The classifiers are random variables, whose central tendency will be to lie on the $MRROC$.

3.2 Thyroid data

A medical data set describing patients with abnormal thyroid conditions was obtained from the UCI machine learning repository [12]. The data was originally contained 7200 instances, with had 3 classes, *hyperthyroid*, *hypothyroid*, and *normal*, and 21 features. In this experiment, the classes were merged to form 2: *Adverse* and *Benign*. The data was randomly split into 3 data sets, Train, with 3800 instances, Test, with 1700 instances, and Unseen, with 1700 instances.

Two classification systems were made, System 1 and System 2, using a simple linear model trained with a single feature to describe the data. Figure 1 shows the ROC curves for both classification systems using the Test data to calculate the true and false positive rates (note that the curves cross).

Figure 5 shows the $MRROC$ predicted by the convex hull containing the Test ROC curves. The vertex points on the hull corresponded to existing classifiers. It was required to validate the hypothesis that all the points on the convex hull were realisable classifiers (by Theorem 1) and could be achieved in practice, resulting in the $MRROC$.

The ROC curves for both of the original systems, and for the set of realisable classifiers on the hull are plotted for the Unseen data in Figure 5. The $MRROC$ produced by application of Theorem 1 is consistent with that predicted, validating the hypothesis.

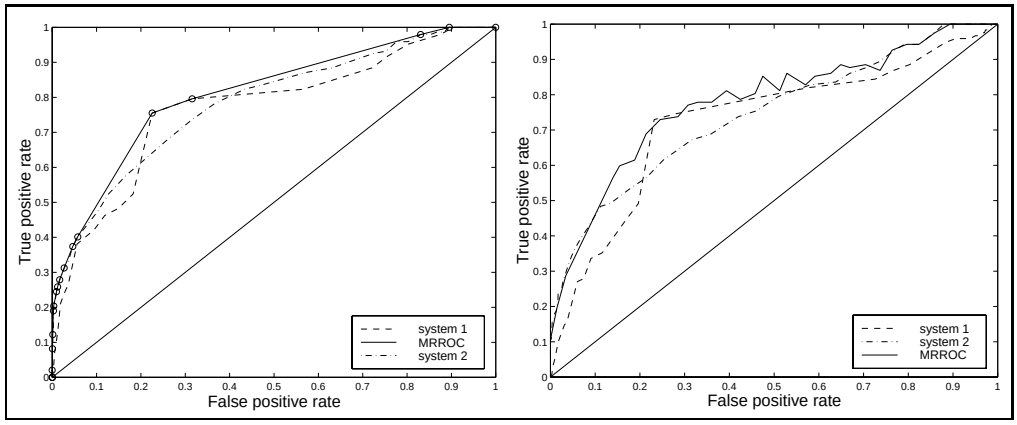

Figure 5: Left: the convex hull over both original ROC curves. This is the predicted $MRROC$. Right: the $MRROC$ plotted for the Unseen data set. The $MRROC$ is consistent with that predicated by the convex hull over the Test data ROC curves. The ROC curves for System 1 and System 2 using the Unseen data are plotted for comparison with the $MRROC$.

3.3 LandSat data

A LandSat image segmentation dataset, originally used in the Statlog project, was obtained from the UCI repository [13, 12]. The data consisted of multi-spectral values of pixels in 3×3 neighbourhoods in a satellite image. A classification was given with the central pixel of each neighbourhood. Originally there were 6 classes: *red soil*, *cotton crop*, *soil with vegetation stubble*, *grey soil*, *damp grey soil*, *very damp grey soil*. In this

experiment, the latter 3 classes have been combined into one class, *grey soil*, and the first 3 into *other*. The objective is to identify cases of *grey soil*, i.e. a correct classification of an example of this class is a *true positive*. As before, the data was split into 3 data sets: Train, with 3000 examples, Test with 1435 examples, and Unseen, with 2000 examples. The data had 36 dimensions, each with values in the range $0..255$.

A simple Bayes classifier was used [3], and the data was discretised using entropy based discretisation described in [4]. Feature subset selection was carried out using *sequential forwards float selection* [15, 8]. Error rate (zero-one loss), was used as a performance measure for subset selection. The error rate was estimating using the Train and Test sets. Improvements in performance were judged statistically significant using McNemars Test [2]. Selection was halted when no statistically significant improvement in classification accuracy could be achieved. The 5 feature subsets found during selection were saved.

Using the Train and Test data, the ROC curves for the classification systems corresponding to each feature subset were evaluated, Figure 6. It can be seen that when the costs vary from error rate, no single feature set produces a superior classification system. The $MRROC$ was predicted by fitting the convex hull over the 5 curves. The classifiers at each vertex were saved.

In Figure 7 the ROC curves for the 5 classification systems and the $MRROC$ on the Unseen data set are presented. The $MRROC$ obtained by application of Theorem 1 is consistent with that predicted, and is clearly superior to any of the individual systems. The $MRROC$ guarantees a maximisation of the Wilcoxon statistic [7], given the available classifiers.

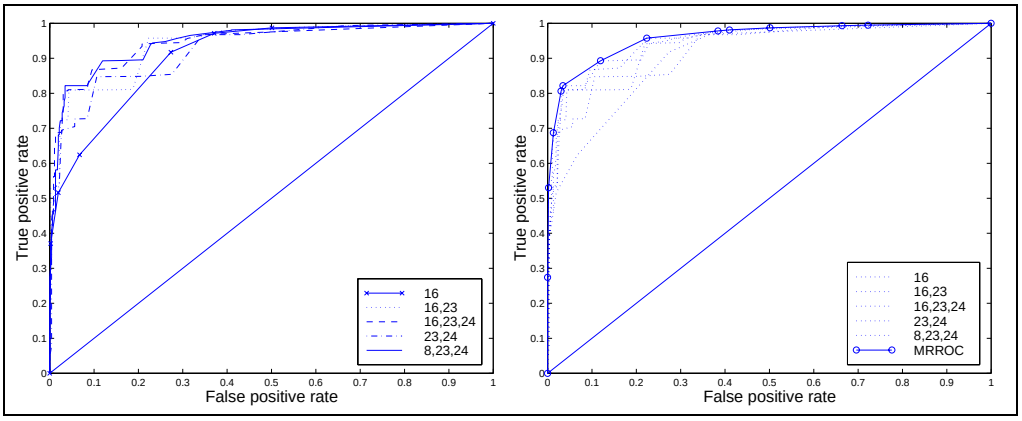

Figure 6: Left: the 5 feature subset ROCs on the Test data. Right: the $MRROC$ predicted by the Test ROCs

4 Conclusions

This technique for producing enhanced performance given a set of existing classifiers and the ROC formed by them may have profound implications for designers of classification systems in domains where classification costs may not be known *a priori*, or may

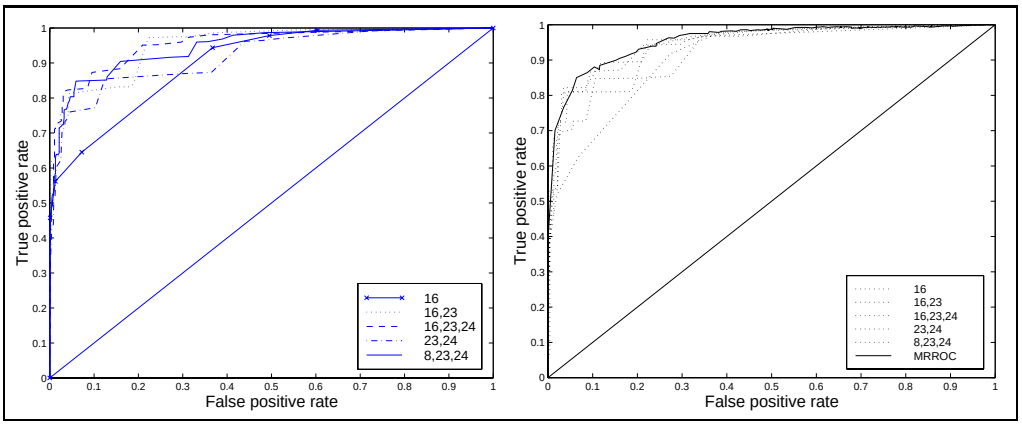

Figure 7: Left: the 5 feature subset ROCs in the Unseen data. Right: the $MRROC$ obtained on the Unseen data, by application of Theorem 1 to the existing classifiers found at the vertices of the Test $MRROC$.

change with time. It has been shown theoretically that an enhanced performance curve, the $MRROC$ can be achieved by application of the realisable classifier theorem, and empirical results provided, on both artificial and real world data, to validate this hypothesis. Given two ROC curves that cross, the $MRROC$ produced using both will be superior to either alone, and may realise operating points with true positive rates that were previously unavailable.

In the first experiment, with multi modal artificial data, the realisable classifiers lying on the facets of the convex hull represent operating points that are not attainable with the original linear classification system. These are not obtained at the expense of clarity or simplicity, nor do they require some degree of expert knowledge to be teased out of the system. The experiments using real world data indicate that this method is both applicable and feasible in such applications.

It is planned to apply this technique to a number of existing problems, such as those reported in [10, 9, 11]. Currently we are examining the application of this methodology to feature selection problems, having developed *Parcel* [16], a novel technique for selecting multiple feature sets across a range of costs.

Acknowledgements

The authors would like to thank the reviewers for their comments, and Dr. David Spiegelhalter for his suggestions about the *realisable classifier theorem*, and for indicating the parallels with classical statistical hypothesis testing.

References

[1] C.B. Barber, D.P. Dobkin, and H.T. Huhdanpaa. The quickhull algorithm for convex hulls,. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996.

[2] T.G. Diettrich. Statistical tests for comparing supervised classification learning algorithms. *Technical Report, Dept Computer Science, Oregon State University*, 1996.

[3] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Leanring*, 12(29):103–130, 1997.

[4] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretisation of continuous features. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 194–202. Morgan Kaufmann, 1995.

[5] R.H. Duda and P.E. Hart. *Pattern classification and scene analysis*. Wiley, 1973.

[6] T.S. Ferguson. *Mathematical Statistics, a Decision Theoretic Approach*. Academic Press, 1967.

[7] D.J. Hand. *Construction and Assessment of Classification Rules*. Wiley, 1997.

[8] A. Jain and D. Zongker. Feature selection: evaluation, application and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[9] D.R. Lovell, B. Rosario, M. Niranjan, R.W. Prager, K.J. Dalton, R. Derom, and J. Chalmers. Design, construction and evaluation of systems to predict risk in obstetrics. *Internationl Journal of Medical Informatics*, 46(3):159–173, 1997.

[10] D.R. Lovell, M.J.J. Scott, M. Niranjan, R.W. Prager, and K.J. Dalton. On the use of expected attainable discrimination for feature selection in large scale medical risk predication problems. Technical Report CUED/F-INFENG/TR.299, Department of Engineering, University of Cambridge, England, August 1997.

[11] D.G. Melvin. A comparison of statistical and connectionist techniques for liver transplant monitoring. Technical Report CUED/F-INFENG/TR.282, Department of Engineering, University of Cambridge, England, December 1996.

[12] C.J. Merz and P.M. Murphy. UCI repository of machine learning databases, 1998.

[13] D. Michie, D.J. Spiegelhalter, and C.C. Taylor. *Machine learning, neural and statistical classification*. Ellis Horwood, 1994.

[14] J. O'Rourke. *Computational Geometry in C*. Cambridge University Press, 1995.

[15] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[16] M.J.J. Scott, M. Niranjan, and R.W. Prager. Parcel: feature subset selection in variable cost domains (available from http://svr-www.eng.cam.ac.uk/reports/abstracts/scott_tr323.html). Technical Report CUED/F-INFENG/TR.323, Department of Engineering, University of Cambridge, England, May 1998.

[17] D. Silvey. *Statistical Inference*. Chapman and Hall, 1975.

[18] J.A. Swets and R.M. Pickett. *Evaluation of Diagnostic Systems*. Academic Press, 1982.