USING UPPER BOUNDS ON ATTAINABLE DISCRIMINATION TO SELECT DISCRETE VALUED FEATURES

D. R. Lovell, C. R. Dance, M. Niranjan, R. W. Prager and K. J. Dalton‡

Cambridge University Engineering Department,
Trumpington St., Cambridge CB2 1PZ, UK
Tel: +44 1223 332754, Fax: +44 1223 332 662,
Email: {drl,crd,niranjan,rwp}@eng.cam.ac.uk

‡Cambridge University Department of Obstetrics and Gynaecology,
Rosie Maternity Hospital, Robinson Way, Cambridge, CB2 2SW, UK.

Abstract

Selection of features that will permit accurate pattern classification is, in general, a difficult task. However, if a particular data set is represented by discrete valued features, it becomes possible to determine empirically the contribution that each feature makes to the discrimination between classes. We describe how to calculate the maximum discrimination possible in a two alternative forced choice decision problem, when discrete valued features are used to represent a given data set. (In this paper, we measure discrimination in terms of the area under the receiver operating characteristic (ROC) curve.) Since this bound corresponds to the upper limit of classification achievable by any classifier (with that given data representation), we can use it to assess whether recognition errors are due to a lack of separability in the data or shortcomings in the classification technique. In comparison to the training and testing of artificial neural networks, the empirical bound on discrimination can be efficiently calculated, allowing an experimenter to decide whether subsequent development of neural network models is warranted.

We extend the discrimination bound method so that we can estimate both the maximum and average discrimination we can expect on unseen test data. These estimation techniques are the basis of a backwards elimination algorithm that can be used to rank features in order of their discriminative power. We use two problems to demonstrate this feature selection process: classification of the Mushroom Database, and a real-world, pregnancy related medical risk prediction task — assessment of risk of perinatal death.

1 INTRODUCTION

Accurate pattern classification requires patterns to be represented by features which, individually or in combination, discriminate between the classes of interest. The work presented in this paper was partly motivated by a need to determine which features, if any, would be useful in discriminating between mothers with low and high risk of adverse pregnancy outcome. In the databases we have access to, maternal information is encoded by hundreds of variables, some of which are clearly relevant to accurate classification, others which are clearly *irrelevant*, the remainder being of uncertain relevance. The techniques for feature selection that we present here should be useful to researchers who face similar problems in other domains.

Many existing feature selection techniques rely upon statistical measures of the between-class separation provided by different feature sets [1, 2]. One drawback with this is that, while certain features may provide a high degree of class separation, there is no guarantee they form a useful representation for making predictions about *new* data. In a medical context, for example, one could achieve perfect separation between cases with adverse or benign outcome by using each patient's social security number as a descriptive feature. Obviously, this feature will be of no use in predicting outcomes for new patients. The methods descibed in this paper use a *rotation error* estimate [3, p. 216] to measure the discrimination we can *expect* to attain when a given subset of features is used to represent new data.

Not only is it desirable to select discriminative features with which to train adaptive classifiers, it is also useful to establish a limit on the classification performance that can be achieved with those features. In the next section, we describe a situation where the empirical calculation of that limit is feasible and, in the rest of the paper, we show how that limit can be employed in the selection of discriminative features.

2 FRAMEWORK OF THE PROBLEM

This paper deals with *two-alternative forced choice* classification [4, 5]. Each pattern in our data set is associated with a vector of discrete valued features, $\boldsymbol{x} = (x_1, \ldots, x_n)$, and a class label, $y \in \{\mathcal{P}, \mathcal{Q}\}$. Cases with identical feature vectors (but not necessarily identical *labels*) are said to belong to the same *bin*. In more formal terms, if $(\boldsymbol{x}_i, y_i)$ is the i^{th} case in our data set, and $X_j \in X$, the set of all distinct feature vectors, then the j^{th} bin is defined by the set function

$$B_j \quad = \quad \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i = X_j\}.$$

We are interested in prediction systems, like artificial neural networks, that are trained to associate a continuous output value with each input vector of discrete features. In this context, classification is achieved by applying a threshold to that output, above which the feature vector is labelled as belonging to class $\mathcal{P}$, below which it is labelled class $\mathcal{Q}$.

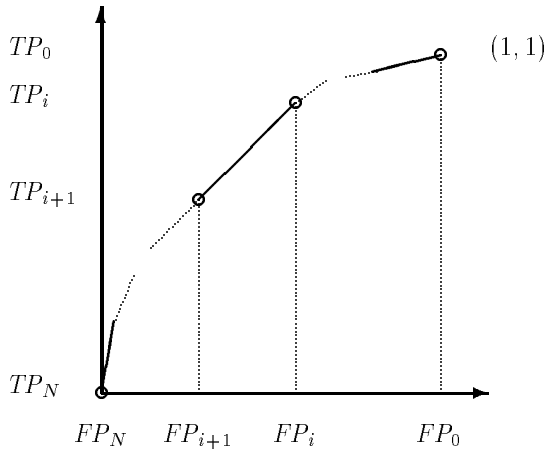


Figure 1: These sections of an ROC curve show true and false positive rates across the thresholds, $t_0, \ldots, t_N$, separating bins, $B_1, \ldots, B_N$, in an ordered list. The area under the curve is the sum of individual trapezoidal areas underneath the lines connecting adjacent points on the curve.

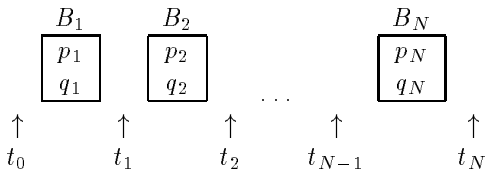
The output a prediction system associates with each bin imposes an order, or ranking, on the bins. Ideally, this ranking permits perfect discrimination between different classes of patterns. In practice, we have to deal with bins that contain patterns of both classes. What is the optimal way to assign outputs (*i.e.*, order bins) in this situation? To answer this, we need a means to assess how effectively a particular ranking of bins discriminates between classes.

3 DISCRIMINATION AND THE ROC

The ROC curve is a convenient way to measure — independent of threshold — how well the ranking imposed by the output of a prediction system separates two classes of data. The ROC curve is usually described in terms *true positive* and *false positive* classification rates. We shall regard patterns from class $\mathcal{P}$ and class $\mathcal{Q}$ as "positive" and "negative" cases, respectively.

Each point on an ROC curve (see Figure 1) plots the true *vs* false positive rate obtained when a particular classification threshold is applied to the continuous output of a prediction system. Let $L = (B_1, B_2, \ldots, B_N)$ define an ordered list of bins in which bin B_i contains p_i patterns from class $\mathcal{P}$ and

q_i patterns from class $\mathcal{Q}$. We can represent this ordering graphically:



where $t_0, \ldots, t_N$ indicate classification thresholds which separate each bin in the ordered list. If all patterns to the right of a threshold are labelled positive and all items to the left, negative, the ROC of this list plots TP_i *vs* FP_i for all thresholds $t_i \in \{t_0, \ldots, t_N\}$ where

$$TP_i = \sum_{j>i}^{N} \frac{p_j}{P} \tag{1}$$

$$FP_i = \sum_{j>i}^{N} \frac{q_j}{Q}, \tag{2}$$

and P and Q denote the number of patterns in classes $\mathcal{P}$ and $\mathcal{Q}$, respectively.

The area under the ROC curve is referred to as *accuracy* and Bamber [6] describes its relationship to the Mann-Whitney U statistic [7], a well known measure of discrimination. If X and Y are two sets of continuous observations, the U statistic is defined to be the total number of (x, y) pairs, for all $x \in X$ and $y \in Y$, in which $x < y$. Bamber shows that for the two-alternative forced choice classification of X and Y

$$\text{accuracy} = U/(|X|\,|Y|), \tag{3}$$

where $|X|$ and $|Y|$ are the number of observations in X and Y respectively.

For discrete observations, the U statistic is defined as

$$U = \sum_{\forall x \in X} \sum_{\forall y \in Y} u(x, y), \tag{4}$$

where

$$u(x, y) = \begin{cases} 1 & \text{if } x > y, \\ \frac{1}{2} & \text{if } x = y, \\ 0 & \text{if } x < y. \end{cases} \tag{5}$$

In the context of this paper, the U statistic is the total number of (class $\mathcal{P}$, class $\mathcal{Q}$) pairs in which the class $\mathcal{P}$ pattern is ranked higher than the class $\mathcal{Q}$ pattern. Clearly, the more discriminative the prediction system, the more pairs will be correctly ranked and the greater the U statistic of that system.

We have established the connection between a prediction system's ability to discriminate and the area under its ROC curve. Now we relate the

area under an ROC curve to the way a prediction system ranks bins. From Equations 1 and 2 we have that

$$TP_i = TP_{i+1} + \frac{p_{i+1}}{P}, \text{ and } FP_i = FP_{i+1} + \frac{q_{i+1}}{Q},$$

so, using the trapezoidal rule for integration (see Figure 1), the area under the ROC curve is

$$\begin{aligned} \text{Area} &= \sum_{i=0}^{N-1} \tfrac{1}{2}(TP_{i+1} + TP_i)(FP_i - FP_{i+1}) \\ &= \frac{1}{PQ}\left(1 \cdot \sum_{i=1}^{N}\sum_{j>i} p_j q_i + \tfrac{1}{2} \cdot \sum_{i=1}^{N} p_i q_i + 0 \cdot \sum_{i=1}^{N}\sum_{j<i} p_j q_i\right) \\ &= \frac{1}{PQ}\sum_{i=1}^{N}\sum_{j=1}^{N} u(B_i, B_j), \end{aligned} \tag{6}$$

where

$$u(B_i, B_j) = \begin{cases} p_j q_i & \text{if } i < j, \\ \frac{1}{2} p_i q_i & \text{if } i = j, \\ 0 & \text{if } i > j. \end{cases} \tag{7}$$

By analogy with Equation (3), the double summation term in Equation (6) represents the Mann-Whitney U statistic for a list of bins.

4 MAXIMIZING THE AREA UNDER THE ROC CURVE

Equation (3) shows that the ranking of bins that maximizes the area under the ROC curve also maximizes the Mann-Whitney U statistic. From Equation (6), the U statistic for a list of bins $L = (B_1, \ldots, B_N)$ can be written

$$\begin{aligned} U(L) &= \sum_{i=1}^{N}\sum_{j>i}^{N} u(B_i, B_j) + \text{constant} \\ &= \sum_{i=1}^{N}\sum_{j>i}^{N} p_j q_i + \text{constant} \\ &= \sum_{i=1}^{N}\sum_{j>i}^{N} q_i q_j \, a_j/(1 - a_j) + \text{constant}. \end{aligned} \tag{8}$$

where $a_j = p_j/(p_j + q_j) = \widehat{\Pr_j}(\mathcal{P})$, the estimated probability of a pattern in bin j belonging to class $\mathcal{P}$. The double summation in Equation (8) contains all $q_i q_j$ terms, for $i < j$, and hence depends only on the $a_j/(1 - a_j)$ ratios.

Because $a_j/(1-a_j)$ is a monotonically increasing function of a_j, the double summation will be maximized when $a_i \leq a_j$ for all $i < j$. Thus, the area under the ROC curve is maximized when

$$\widehat{\Pr_1}(\mathcal{P}) \leq \widehat{\Pr_2}(\mathcal{P}) \leq \cdots \leq \widehat{\Pr_N}(\mathcal{P}),$$

that is, when the bins are ranked in ascending order of likelihood of class $\mathcal{P}$ membership [4].

This result was stated in a slightly different context by Hanley and McNeil [8], however, one issue that they did not tackle was the question of the accuracy that can be *expected* with unseen test data. Breiman *et al.* consider this question in relation to misclassification in their CART model [9, Ch. 3] but do not explore the situation where discrete valued data permits estimation of bounds on accuracy. In the next section, we describe methods to estimate such bounds.

5 ESTIMATING EXPECTED ACCURACY

One classic method of estimating the expected performance of a system is to reserve, or "hold out" data to evaluate the performance of a trained system. If we employ this principle to estimate the maximum accuracy expected with a given set of discrete features, we use two sets of data drawn from the population we wish to classify. We then

1. Find the ordering of bins, L^1, to maximize the accuracy attainable on the first data set.

2. Apply that ordering to the corresponding bins in second data set to give L^2, and then measure the accuracy obtained with L^2.

One problem with this scheme is how to order bins in the second data set that do not appear in the first data set. Suppose B_i^1 and B_i^2 are corresponding bins in our two data sets. If $p_i^1 = q_i^1 = 0$ but $p_i^2 + q_i^2 \neq 0$, we say that bin B_i^2 is *unassigned*. There are two meaningful ways to insert an unassigned bin B_i^2 into the ordering of bins from the second data set:

1. Insert B_i^2 randomly into L^2. The accuracy obtained with L^2 will be an estimate of the *average* accuracy we can expect with our discrete features.

2. Insert B_i^2 so as to maximize $U(L^2)$, the Mann-Whitney U statistic for L^2. The accuracy obtained with L^2 will be an estimate of the *maximum* accuracy we can expect with our discrete features.

When more than one bin is unassigned, the second method raises the question: if we insert each unassigned bin into the position which maximizes $U(L^2)$ *for that bin only*, do we obtain the maximum $U(L^2)$ once *all* bins have been inserted? It can be shown [10] that the answer to this question is "yes".

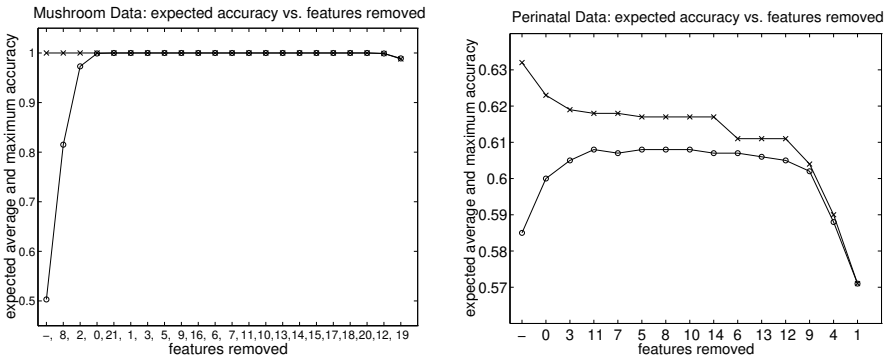


Figure 2: These graphs show how successive deletion of features affects the maximum ($\times$) and average ($\circ$) accuracy expected with the Mushroom Database and the perinatal mortality prediction problem. The saturated model (*i.e.*, the model incorporating all features) is indicated by "–".

These estimates of expected accuracy may be used to assess the contribution of individual discrete features to the attainable accuracy. We propose this *backwards deletion* strategy:

$$W_{n+1} \quad = \quad W_n - \underset{x_i \in W_n}{\text{argmax}} \ \text{accuracy}(W_n - \{x_i\}), \tag{9}$$

where the working set W_n initially contains all discrete features used to describe the data. The accuracy(W_n) function returns an estimate of the average accuracy we can expect using the features in W_n to represent a given data set. Features are removed from the working set in order of increasing importance to attainable accuracy.

The problem of selecting an optimal subset of features is NP hard and the backwards selection algorithm described here is not the only practical method for searching the space of possible models [3]. While forwards selection of features would require significantly less computation, commencing the search with the saturated model allows us to readily identify and retain variables whose *interaction* provides good discrimination between classes. In the following subsections we report how this strategy works in practice.

5.1 Feature selection in The Mushroom Database

The Mushroom Database [11] contains 8124 patterns describing hypothetical samples of the Agaricus and Lepiota Family of mushrooms. Each pattern consists of 22 discrete valued features and is labelled either "poisonous" or "edible". We took every even numbered pattern and applied the backwards deletion algorithm (Equation (9)) to the resulting data set, D. The left hand graph of Figure 2 shows the expected maximum and average accuracies

obtained. These estimates were formed by randomly partitioning D into data sets D^1, D^2 then applying the methods described in Section 5. Each point on the graph shows the average of 30 such estimates.

Since all features except 12, 19 and 4 can be removed before the expected maximum accuracy falls below 1.0, we can surmise that only these three features are relevant in discriminating between poisonous and edible mushrooms. Examination of the entire Mushroom Database confirms this conjecture.

The low expected average accuracy obtained for the saturated model is because each bin contains only one example when the data set D is described by all features. Consequently, when D is partitioned into D^1 and D^2, both partitions are disjoint so all bins in L^2 are unassigned and are ranked randomly. Hence, the expected accuracy is at the level of chance: 0.5.

5.2 Reviewing features used to predict perinatal mortality

The motivation behind the techniques described in this paper stemmed from the Quality Assurance in Maternity Care (QAMC) project which aims to develop neural network systems for the accurate prediction of risk of adverse pregnancy outcome. In this subsection, we consider a particular adverse outcome — perinatal death — that was used in a preliminary experiment to review an existing risk prediction scheme.

The Antenatal Prediction Score (ANPS) proposed by Chamberlain *et al.* in [12] is an additive score to assess risk of perinatal death based on information available at the first booking visit. It was natural to ask whether the performance of such a simple scheme could be bettered by a more sophisticated nonlinear model. However, before committing time and effort to developing new prediction systems, we wanted to know how much improvement could be obtained.

The basic ANPS uses 10 discrete features to score risk and Chamberlain *et al.* suggest a further 10 features to be incorporated into the score. However, since some of those features were unavailable in the perinatal databases we had access to, we used a total of 15 features to represent each of 771571 records from the Scottish Maternity Register. Because we were interested in the greatest discrimination attainable with these features, we applied the backwards deletion algorithm (Equation (9)) to the *entire* data set.

The resulting estimated accuracies shown in the right hand graph of Figure 2 indicate that effective discrimination in this real-world task is much harder than in the artificial Mushroom problem. The plot of average expected accuracy demonstrates the characteristic trade-off between fine (high variance) and coarse (high bias) grained solutions. For us, the most important aspect of the graph is the overall low discrimination afforded by the features. This is not surprising, given the nature of the prediction task, however, it does indicate that more discriminative features are required (as opposed to better modelling techniques) if risk is to be predicted with useful levels of accuracy. Thus, time that would have been wasted on building more

complex prediction systems can be devoted to uncovering better predictor variables.

6 CONCLUSIONS

The methods presented in this paper were developed to to evaluate how useful a set of discrete features (and all their interaction effects) are in discriminating between two classes. These techniques exploit the fact that empirical bounds on accuracy can be calculated when data is represented using discrete valued features. Once such bounds have been established for a particular task, an investigator can assess whether imperfect classification performance is attributable to shortcomings in the classification system or a lack of separability in the data.

Features can also be ranked in order of their discriminative power using the backwards elimination algorithm detailed in Section 5. Thus, for a given database, we can use the upper bounds on attainable discrimination to select a subset of discriminative features.

The methods we have described are particularly suited to problems where large amounts of data permit good estimates of class conditional probabilities (such as problem in Subsection 5.2). However, the results obtained on the Mushroom Database task show that accuracy estimates are still useful in "screening" the features used to describe smaller data sets.

One shortcoming of the techniques presented is their inability to extend to the discrimination of continuous features. An obvious solution to this would be to discretize any such features, then apply the accuracy estimation methods. However, this approach clearly depends on the coarseness of the discretization and introduces a familiar trade-off between model bias and variance. Such issues are beyond the scope of this paper and we must leave open the question of how to deal with continuous features.

In summary, this paper presents techniques that we have found useful in the preliminary screening of discrete valued features. We hope that researchers in other domains will find these methods of benefit.

Acknowledgements

We are pleased to acknowledge the assistance of Dr Susan Cole of the Scottish Home and Health Department for collecting and providing us with access to Scottish Maternity Register data. This work was funded by the Quality Assurance in Maternity Care project, EC Biomed Grant Ref. BIOMED 1 CT93 1113. The first author gratefully acknowledges The British Council for providing travel funding for him to take up the position of research associate with the QAMC project.

References

[1] J. Kittler, "Mathematical methods of feature selection in pattern recognition," *International Journal of Man-Machine Studies*, vol. 7, pp. 609–637, 1975.

[2] D. E. Boekee and J. C. A. Van der Lubbe, "Some aspects of error bounds in feature selection," *Pattern Recognition*, vol. 11, pp. 353–360, 1979.

[3] W. Siedlecki and J. Sklansky, "On automatic feature selection," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 2, no. 2, pp. 197–220, 1988.

[4] D. Green and J. Swets, *Signal Detection Theory and Psychophysics*. New York: John Wiley, 1966.

[5] J. A. Swets and R. M. Pickett, *Evaluation of diagnostic systems*. New York: Academic Press, 1982.

[6] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, vol. 12, pp. 387–415, 1975.

[7] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, pp. 50–60, 1947.

[8] J. Hanley and B. McNeil, "Maximum attainable discrimination and the utilization of radiologic examinations," *Journal of Chronic Disease*, vol. 35, pp. 601–611, 1982.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[10] D. R. Lovell, C. R. Dance, M. Niranjan, R. W. Prager, and K. J. Dalton, "Limits on the discrimination possible with discrete valued data, with application to medical risk prediction," Tech. Rep. CUED/F-INFENG/TR243, Cambridge University Engineering Department, January 1996.

[11] J. S. Schlimmer, "Mushroom database." `ftp://ics.uci.edu/pub/machine-learning/`, 1987.

[12] G. Chamberlain *et al.*, eds., *British births 1970: a survey under the joint auspices of the National Birthday Trust Fund and the Royal College of Obstetricians and Gynaecologists*, vol. 2: Obstetric care. London: Heinemann Medical, 1978.