Ranking the effect of different features on the classification of discrete valued data

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

There is a fundamental upper limit on the extent that any classifier — neural network or otherwise — can discriminate between two classes of discrete valued data. When classes overlap in feature space, discrimination will be less than perfect.

In a two alternative forced-choice decision problem, a classifier's discrimination ability is often measured in terms of the area under its receiver operating characteristic (ROC) curve. In this paper, we show how to calculate the maximum possible area under the ROC curve obtained with a particular representation of a given set of discrete valued data. This result corresponds to the fundamental upper limit of discrimination.

We extend this result to show how to estimate both the maximum and the average discrimination we can expect to achieve on unseen test data, again, for a particular representation of a discrete valued data set.

These results have practical engineering application in the selection of discriminative features for neural network training. We show how the bound on discrimination can be used in a backwards elimination algorithm to rank discrete features in order of their discriminative power. The algorithm commences with the saturated model (*i.e.*, using all features and all interactions to represent the data) and removes the feature that has least impact on the upper limit of the model's discrimination. Thus, features are removed from the model in reverse order of their discriminative power.

This feature ranking technique is applied to a machine learning benchmark: the classification of the Mushroom Database. We also describe its application to a larger real-world medical risk prediction problem: prediction of risk of adverse pregnancy outcome from information available at initial hospital booking. The results obtained with the latter task suggest that improved risk prediction would require more discriminative information, rather than better predictive techniques. This demonstrates how we can determine whether a classifier's performance is being limited by an inherent inseparability of the data.

Keywords: medical risk prediction, receiver operating characteristic, failure to progress.

1 Introduction

Selecting discriminative features is the first and perhaps the most important step in implementing a classifier. Accurate pattern classification requires that data be represented by features which, individually or in combination, discriminate between the classes of interest. Even if we avoid making a "hard" decision about a pattern's class by estimating probability of class membership instead, there are many situations where we must eventually choose between two alternative actions, based on that estimate. Such situations arise frequently in medical risk prediction, the area which provides the principal motivation for the work described here.

The methods we present were developed as a practical means to determine which features, if any, would be useful in predicting a mother's 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 risk prediction, 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.

Not only is it desirable to select discriminative features with which to train prediction systems, it is also useful to establish a limit on the performance achievable 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 [1]. 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*. More formally, 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

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

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 $\mathcal{Q}$.

The output that 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 curve

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. More formally, 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:

$$\begin{array}{ccccccccc}
 & B_1 & & B_2 & & & & B_N & \\
 & \boxed{\begin{array}{c} p_1 \\ q_1 \end{array}} & & \boxed{\begin{array}{c} p_2 \\ q_2 \end{array}} & & \ldots & & \boxed{\begin{array}{c} p_N \\ q_N \end{array}} & \\
\uparrow & & \uparrow & & \uparrow & & \uparrow & & \uparrow \\
t_0 & & t_1 & & t_2 & & t_{N-1} & & t_N
\end{array}$$

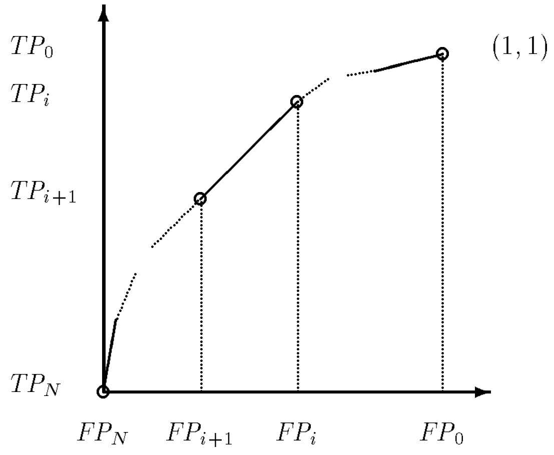


Figure 1: These sections of an ROC curve show true and false positive rates across the thresholds, $t_0, \ldots, t_N$, separating bins, $B_1, \ldots, B_N$, in an ordered list. The area under the curve is the sum of individual trapezoidal areas underneath the lines connecting adjacent points on the curve.

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 [2] describes its relationship to the Mann-Whitney U statistic [3], 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

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

where

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

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

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

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.

This result was stated in a slightly different context by Hanley and McNeil [4], 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 [5, 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 Estimation of 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 must 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 [6] that the answer to this question is "yes".

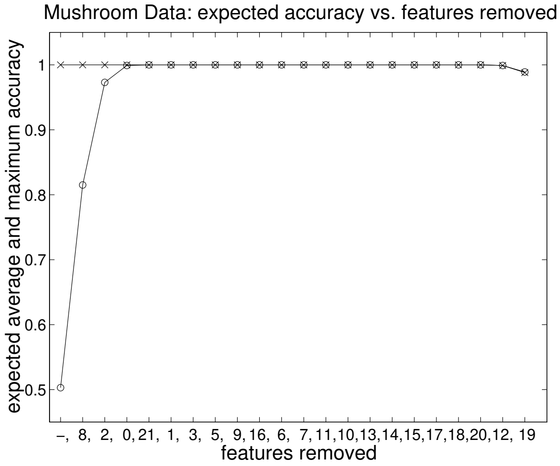
These estimates of expected accuracy may be used to assess the contribution of individual discrete features to the attainable accuracy. We propose this *backwards elimination* strategy:

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

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. In the following subsections we look at how this strategy works in practice.

5.1 Selecting useful features from The Mushroom Database

The Mushroom Database [7] 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 elimination algorithm (Equation (9)) to the resulting data set, D. 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.



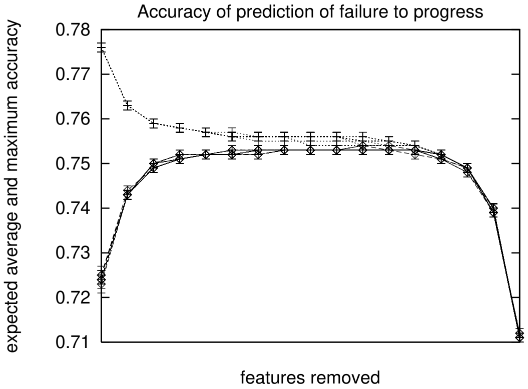


Figure 2: This graph shows how successive elimination of features affects the maximum ($\times$) and average ($\circ$) accuracy expected with the Mushroom Database. The saturated model (*i.e.*, the model incorporating all features) is indicated by "–". Note that feature 4 is retained throughout the backwards elimination process.

Figure 3: This graph shows the maximum (upper curves) and average (lower curves) expected accuracies of prediction of failure to progress, obtained over 10 runs of the backwards elimination algorithm. Table 1 shows the order in which features were removed from the saturated model. The error bars show one standard deviation above and below each point estimate.

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 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 What information helps us predict "failure to progress"?

The methods we have described were developed under the Quality Assurance in Maternity Care (QAMC) project as a practical means (1) to establish upper bounds on the accuracy of prediction of adverse pregnancy outcomes and (2) discover which features best determine risk of adverse outcome. One of the outcomes considered in the QAMC project is *Caesarian section for failure to progress* and in this section we discuss an investigation into the precursor to that outcome: failure to progress. Put simply, a failure to progress means that the natural course of labour has stalled before delivery, placing the wellbeing of both mother and child at risk. In short, our aim was to determine how accurately a prediction of failure to progress could be made on the basis of information available at a mother's initial visit to her doctor.

We used data from the Scottish Maternity Register (SMR) on 771571 singleton births that occurred between 1980–91. Our first task was to decide which variables within that data were likely to be relevant to the prediction of failure to progress. Since the SMR uses the 9th Revision of the International Classification of Diseases (ICD-9) [8] to encode information about maternal condition, we had to pre-select a manageable and meaningful subset of conditions from over 1900

Failure to progress											
Feature and Range		**Order of removal**									
Social class	0-5	1	1	1	1	1	1	1	1	1	1
Blood group	0-3	2	2	2	2	2	2	2	2	2	2
Previous spontaneous abortions	0,1	3	3	4	4	4	3	3	3	3	3
Previous theraputic abortions	0,1	4	11	3	3	3	4	5	4	4	5
Poor reproductive history*	0,1	6	4	6	6	5	7	9	6	7	4
Poor obstetric history*	0,1	8	6	5	5	8	6	6	9	5	9
History of abortions*	0,1	7	7	9	9	6	9	4	5	6	6
History of infertility*	0,1	11	8	7	11	11	5	11	8	9	7
Illegitimate pregnancy*	0,1	9	9	12	7	7	10	10	7	8	10
Fam. hist. endocr./metab. disorder*	0,1	5	12	11	10	9	8	8	10	10	8
Mother supported	0,1	12	5	8	8	10	12	12	12	12	12
Unwanted pregnancy*	0,1	10	10	10	12	12	11	7	11	11	11
Previous neonatal deaths	0,1	13	13	13	13	13	13	13	13	13	13
Height > 155cm	0,1	14	14	14	14	14	14	14	14	14	14
Previous Caesarian sections	0,1	15	15	15	15	15	15	15	15	15	15
Age	0-5	16	16	16	16	16	16	16	16	16	16
Parity	0-5	17	17	17	17	17	17	17	17	17	17

Table 1: The discrete variables used to predict failure to progress, and the order in which they were removed from the saturated model across 10 separate runs of the backwards elimination algorithm. An asterisk indicates a variable abstracted from the list of ICD-9 codes describing maternal conditions.

distinct codes appearing in the database. This was done by assessing the relative strength of association (using Cramer's V statistic) between failure to progress and each maternal condition, then applying an arbitrary cutoff to select only those conditions strongly associated with the outcome. Obviously, this univariate screening process is not the optimal approach to choosing promising variables for *multivariate* prediction, however, we are not aware of a more effective scheme that could be practically applied to such a large set of candidate variables.

Further screening was done by retaining only those ICD-9 codes signifying conditions certain to be present at a mother's initial booking visit. The resulting set of 17 discrete valued predictors was used to represent the SMR records[1] and Table 1 shows the order in which 10 separate runs of the backwards elimination algorithm removed these predictors from the saturated model. Figure 3 plots the maximum and average expected accuracies obtained at each step of backwards elimination.

Clearly perfect prediction of failure to progress is not possible with the information contained in the our 17 discrete valued predictors. Furthermore, Figure 3 suggests that only the last 5 features in Table 1 appear to be useful discriminants. Had we proceeded directly from initial screening of variables to the development of neural networks for outcome prediction — without investigating the maximum attainable discrimination — we would probably have spent much time and effort trying to achieve impossible levels of accuracy with unnecessarily large numbers of predictors. Instead, the experiment described here (along with several others dealing with different adverse outcomes) has suggested a promising new direction that is in accord with conventional obstetric wisdom: adopt different prediction methods for first pregnancies (parity 0) and later pregnancies (parity > 0).

[1]Records with missing/invalid values in predictor variables (around 12.3% of the database) were deleted.

6 Conclusions

We have shown how to estimate the average and the maximum accuracy attainable when discrete valued features are used to represent a given data set. We have also shown how such estimates can be employed in a backwards elimination algorithm to remove predictor variables from a saturated model in increasing order of their discriminative power. By using this approach to rank the effect that different features have on prediction accuracy, we can determine those variables that could be usefully employed in actual prediction systems. (Note that since the backwards elimination algorithm considers all possible interaction effects, this method of feature selection is particularly apposite as a precursor to neural network modelling.)

The methods we have described are well suited to problems where large amounts of data permit good estimates of class conditional probabilities (as was the case in the prediction of failure to progress). 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. In both experiments, the graphs of estimated average accuracy show the characteristic trade-off between high variance models (with too many parameters to generalise effectively) and high bias models (with too few parameters to adequately model the data).

It is important to remember that the backwards elimination algorithm is not deterministic. As Table 1 shows, the order in which features are eliminated is not always the same (though certain features did tend to be eliminated early, whilst others were retained to the last). This stochastic element arises because the algorithm employs *estimates* of attainable accuracy, estimates formed by randomly partitioning an entire data set into halves. Thus, while each run of the algorithm does rank features in ascending order of discriminative power, it is advisable *not* to attach too much significance to a single ranking; better to look at trends across a series of experiments.

In summary, this paper presents techniques we have found useful in preliminary screening of discrete valued features. We hope 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. A. Swets and R. M. Pickett, *Evaluation of diagnostic systems.* New York: Academic Press, 1982.

[2] 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.

[3] 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.

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

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

[6] 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.

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

[8] World Health Organization, *International Classification of Diseases, Ninth Revision.* Geneva: World Health Organization, 1975.