Feature selection using Expected Attainable Discrimination

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

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.

‡Katholieke Universiteit Leuven, Center for Human Genetics,
Herestraat 49, B-3000 Leuven, Belgium

Abstract

We propose *expected attainable discrimination* (EAD) as a measure to select discrete valued features for reliable discrimination between two classes of data. EAD is an average of the area under the ROC curves obtained when a simple histogram probability density model is trained and tested on many random partitions of a data set. EAD can be incorporated into various stepwise search methods to determine promising subsets of features, particularly when misclassification costs are difficult or impossible to specify. Experimental application to the problem of risk prediction in pregnancy is described.

Keywords: Receiver operating characteristic (ROC); Area under the ROC curve; Feature selection; Risk prediction in pregnancy; Failure to progress.

1 Introduction

Feature selection is a key step towards solving any practical classification problem (Kittler, 1975). Even though a Bayes classifier can never be improved by omission of features (Van Campenhout, 1982), there are compelling practical and theoretical reasons why features should be carefully selected. In practice, we cannot build Bayes classifiers because the true class conditional probabilities that characterise our data are not known. Instead, we build parameterised classifiers and use training data to estimate suitable parameter values. Finite amounts of data impose a trade off between the complexity of a classifier and the accuracy with which its parameters can be estimated. Omission of one or more predictors simplifies a classifier and can lead to more accurate parameter estimates, thereby improving performance (Siedlecki and Sklansky, 1988; Duda and Hart, 1973, Section 3.8).

There are two main kinds of feature selection methods: those based on probabilistic separability measures applied to the data, and those based on the error rate of a classifier as a design criterion. In this paper, we explore the latter approach. As Siedlecki and Sklansky (1988, Section 3.4) point out, such an approach "...must be recognised as a process in which [the] classifier is optimised and, therefore, trained. Hence, the only error rate that can be computed for this classifier is an apparent error rate, which is known to be very biased". The feature selection criterion proposed in this paper is based around a classification model that is both fast to train and, through the use of hold-out data, able to provide a performance estimate less biased than apparent error rate.

In many situations, rather than minimise the error rate of a classifier, it is desirable to minimise the *risk*, *i.e.*, the false positive and false negative rates weighted by their respective misclassification costs. Sometimes (particularly in medical problems) these costs may be difficult or impossible to specify. However, the area under the *receiver operating characteristic* (ROC)

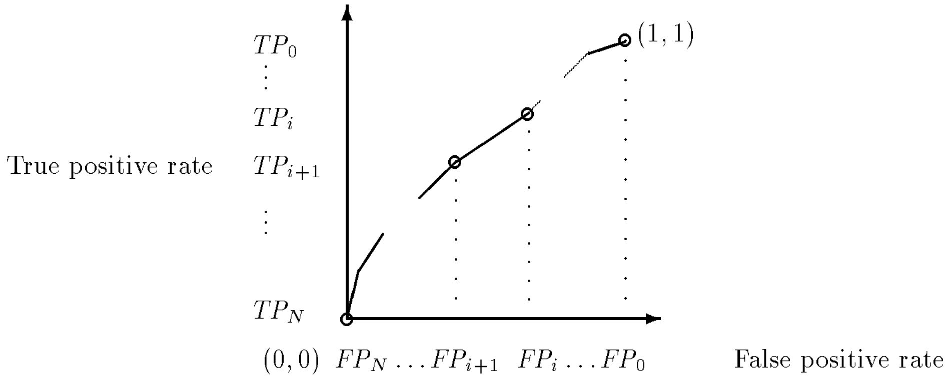

Figure 1: These sections of an ROC curve show true and false positive rates across the thresholds, $t_0, \ldots, t_N$, separating bins, $B_1, \ldots, B_N$, in an ordered list. Straightforward methods to calculate the area under this curve (and its standard error) are presented by Bamber (1975).

curve can provide a useful measure of a classifier's performance across all possible costs. This area (AUROC for short) represents the extent to which a classifier can discriminate between two classes of data in a given data set. In the following sections, we describe the problem setting, and show how the AUROC can be used to estimate the discrimination expected to be attainable with a given set of features. We then describe how this *expected attainable discrimination* can be used in a sub-optimal search for discriminative features.

2 Problem setting

We wish to choose features that will allow us to discriminate between patterns of class $\mathcal{P}$ and patterns of class $\mathcal{Q}$. Each training pattern is represented by a multidimensional vector of discrete valued features, $\boldsymbol{x} = (x_1, \ldots, x_n)$, and a class label, $y \in \{\mathcal{P}, \mathcal{Q}\}$. We implement a histogram probability density model by grouping patterns with identical feature vectors into *bins* such that bin B_i contains p_i patterns from class $\mathcal{P}$ and q_i patterns from class $\mathcal{Q}$. With each bin, we associate an estimate of the probability that data from that bin belong to class $\mathcal{P}$, that is, $\hat{P}(y = \mathcal{P}|\boldsymbol{x} \in B_i)$ for bin i. This imposes an ordering on the bins that we can represent as:

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

where bins are arranged from left to right in ascending order of $\hat{P}(y = \mathcal{P}|\boldsymbol{x} \in B_i)$, and classification thresholds $t_0, \ldots, t_N$ separate each bin in the list. Using this representation, we define

$$\text{true positive rate at } t_i\text{:} \quad TP_i = \sum_{j>i}^{N} \frac{p_j}{|\mathcal{P}|} \tag{1}$$

$$\text{false positive rate at } t_i\text{:} \quad FP_i = \sum_{j>i}^{N} \frac{q_j}{|\mathcal{Q}|}, \tag{2}$$

where $|\mathcal{P}|$ and $|\mathcal{Q}|$ denote the number of patterns in classes $\mathcal{P}$ and $\mathcal{Q}$, respectively. The ROC curve (Figure 1) plots TP_i versus FP_i across all thresholds t_i, and the AUROC measures how well a prediction system discriminates between two classes of outcome (Green and Swets, 1966).

The AUROC depends on how bins are ordered, and that is determined by the estimates of $\hat{P}(y = \mathcal{P}|\boldsymbol{x} \in B_i)$. In Section 3 we discuss the form of these estimates that maximises the AUROC. Many classifiers work by estimating $\hat{P}(y = \mathcal{P}|\boldsymbol{x} \in B_i)$ (or some monotonically increasing function thereof) and then applying a threshold to that estimate, above which $\boldsymbol{x}$ is deemed to belong to class $\mathcal{P}$. The optimum value of the threshold is determined by the misclassification costs and the class prior probabilities. The AUROC provides a summary of the classifier's behaviour across *all* thresholds and is, thus, an appropriate measure of performance when misclassification costs are not clearly specified.

Unlike most measures of probabilistic separation, the AUROC has a number of physical interpretations, depending on the experimental setting in which it was obtained (Green and Swets, 1966, Chapter 2). In this paper, the following interpretation is relevant: given two patterns, one randomly chosen from class $\mathcal{P}$, the other randomly chosen from class $\mathcal{Q}$, the probability that a classifier correctly decides which pattern is class $\mathcal{P}$ is equal to the area under the classifier's ROC curve. Throughout the remainder of this paper, we shall use the terms "discrimination" and "area under the ROC curve" synonymously.

3 Maximum attainable discrimination

It is possible to calculate the maximum extent to which one can discriminate between two classes within a given set of data. From Equations 1 and 2 we have that

$$\begin{aligned} TP_i &= TP_{i-1} - \frac{p_i}{|\mathcal{P}|} \\ FP_i &= FP_{i-1} - \frac{q_i}{|\mathcal{Q}|}. \end{aligned}$$

Since $|\mathcal{P}|$ and $|\mathcal{Q}|$ are constant in a given data set, the slope of the i^{th} segment of the ROC curve (between adjacent points $(FP_i, TP_i), (FP_{i-1}, TP_{i-1})$) is proportional to p_i/q_i, and equal to the estimated likelihood ratio in the i^{th} bin. Thus, if we place bins in ascending order of estimated likelihood ratio[1], the line segments that make up the ROC curve appear, from left to right, in decreasing order of slope, and the AUROC is maximized (see step (1) in Figure 2) (Green and Swets, 1966; Hanley and McNeil, 1982).

This *maximum attainable discrimination* — or *MAD*, as Hanley and McNeil (1982) refer to it — is easily calculated and applies to *any* system which associates outputs with discrete valued data in a consistent manner (*i.e.*, labels identical patterns identically). In the context of discrete valued features, MAD is directly related to the Bayes error rate defined by the distribution of the two classes in a given data set. As we shall see in Section 6, MAD helps determine whether a prediction system's performance is limited by an inherent lack of separability in the data, or shortcomings of the prediction model.

Note that MAD increases with the number of features used to represent a given data set. However, our ability to generalise from the characteristics of a finite amount of data tends to *decrease* once a certain representational complexity is exceeded.

4 Expected attainable discrimination

Our aim in feature selection is to choose from a set of available predictors, those which afford the greatest degree of discrimination between two classes of data. Implicit in this objective are the ideas that a classification system will eventually be built using the selected representation(s) of the data, and that that system will be used to classify *new* data.

A natural way to select features is simply to evaluate the performance of a classification model using a variety of representations of new data. Ignoring the combinatorial nature of the number of possible feature subsets for the time being, we propose a practical classification scheme and a way to estimate its performance on new data. The scheme uses the histogram probability density model described in Section 2 to estimate *expected attainable discrimination* (EAD). Given two sets of data, D^1 and D^2, drawn from the population we wish to classify (see Figure 2), we

[1] or, equivalently, $p_i/(p_i + q_i)$. This maximum likelihood estimate of the probability that data from bin B_i belongs to class $\mathcal{P}$ is what we would obtain using a simple histogram model (Bishop, 1995) of the probability density function underlying the data set.

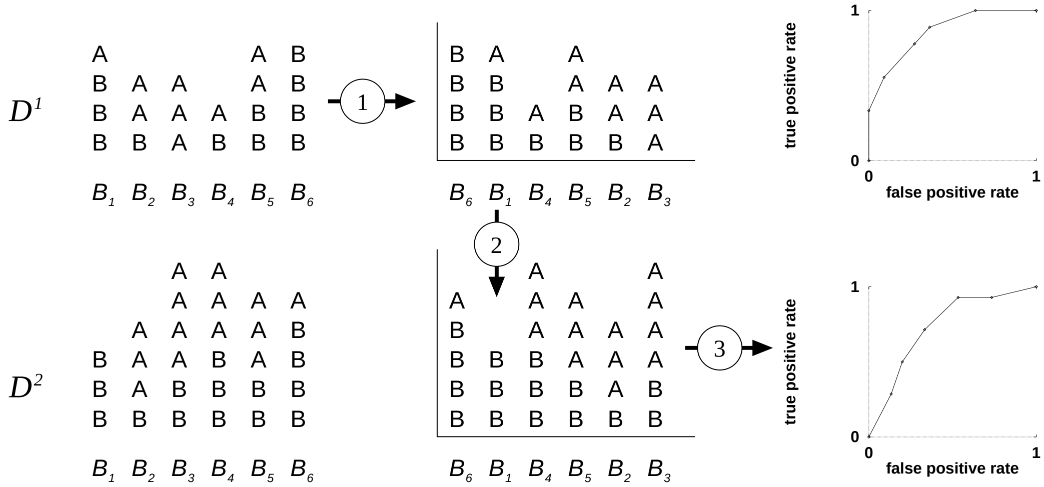

Figure 2: To estimate the discrimination we can *expect* to obtain with a given representation of a specific dataset, we randomly partition our dataset into two disjoint subsets, D^1 and D^2. The D^1 bins, B_1–B_6, are placed in MAD order (1). The D^1 MAD ordering is applied to the bins in the D^2 subset (2). The area under the D^2 ROC curve is an estimate of the expected discrimination (3).

1. obtain the MAD with D^1 by placing its bins in ascending order of $p_i/(p_i + q_i)$ to form the list, L^1;
2. apply that ordering to the corresponding bins in D^2 to give L^2, and measure the discrimination obtained with that list.

The area under the ROC curve for L^2 estimates EAD, and indicates how well the features representing the data allow us to predict the outcome of new data.

Any bins that appear in D^2 but not D^1 are inserted randomly into the L^2 ordering. There are other ways to deal with these "unassigned" bins (such as the Bayesian approach described by MacKay and Peto (1995)), however this method was used throughout the experiments described in this paper.

Data sets D^1 and D^2 should be independent to avoid highly biased estimates of the classifier's performance. Methods such as cross-validation and the bootstrap exist to generate training and test data sets from a single data set so as to mitigate such bias (see Weiss and Kulikowski (1991) for a summary). Given the large amount of data available to us in our experiments, it was convenient to randomly select two-thirds of our data to create set D^1 and use the remaining third to form D^2. An *rotation error* estimate (Siedlecki and Sklansky, 1988, p. 216) of EAD was formed by averaging EAD across ten such random partitions.

5 Using EAD to select discriminative features

EAD provides a measure of the degree of discrimination we can expect to achieve when a particular subset of features is used to represent a given data set. Thus we can use EAD in a search for the best subset of features among a number of feasible subsets. Since this search is NP hard, we must resort to sub-optimal methods (Kittler, 1978).

The experiments described in this paper use *sequential forward selection* with EAD (which we refer to as FEAD). (The more computationally intensive sequential backwards selection (BEAD) is reported by Lovell et al. (1996).) FEAD is an iterative search method defined by

$$W_{n+1} \quad = \quad W_n + \operatorname*{argmax}_{x_i \notin W_n} \mathrm{EAD}(W_n + \{x_i\}), \tag{3}$$

where the working set W_n contains the n discrete features used to describe the data at the current search step. Using the methods described in the previous section, $\mathrm{EAD}(W_n)$ returns an estimate

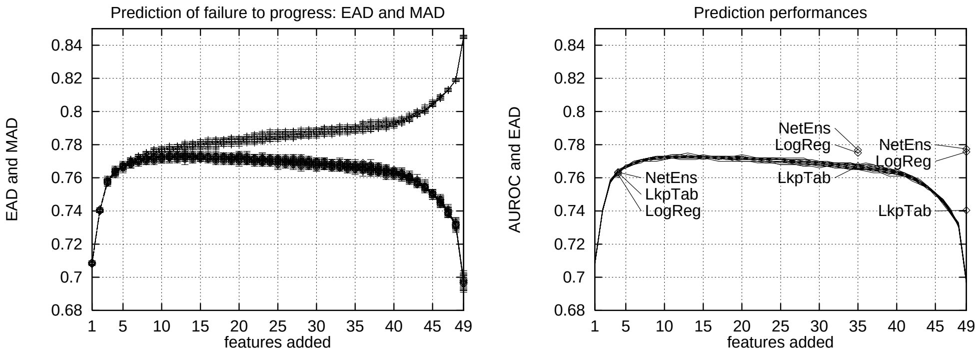

Figure 3: The left graph shows the MAD (upper curves) and EAD (lower curves) for prediction of failure to progress, obtained over 20 runs of FEAD. Error bars show one standard deviation above and below each point estimate. Features are added, one at a time, from left to right, starting with a single feature model. The order of addition is shown in Table 1.

The right graph plots test set prediction performances of different models against the number of features used in each model. Performance is summarised by the AUROC obtained on an independent test set of 176 812 cases. The EAD (lower curves) obtained over 20 runs of FEAD is shown for comparison.

of the discrimination we can expect when the features in W_n are used to represent a given data set. Thus, features will be added from the working set in order of decreasing importance to attainable discrimination. Since EAD(W_n) is an estimate based on several random partitions of the data, the order in which features are selected is not deterministic. For this reason, we recommend performing several runs of the algorithm to observe *trends* in the order of feature selection. In the following section we report how this strategy works in practice.

6 Selecting features to 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 outcome considered in the QAMC project is *Cæsarian section for failure to progress* and in this section we focus on 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 prior to labour. Before this could be achieved, we had to determine which features to use to make that prediction. Furthermore, we had to select features *without* knowledge of the costs of misclassification.

We used data from the Scottish Morbidity Record (known as the SMR2) on singleton births that occurred between 1980–88. Cases in which elective Cæsarian section or breech presentation occurred were removed from the dataset[2], leaving 540 905 records. Since the SMR2 uses the 9th Revision of the International Classification of Diseases (ICD-9) (World Health Organization, 1975) to encode information about maternal condition, that data contained many features relating to conditions prior to the onset of labour (over 300 distinct ICD-9 codes appeared). From this set of features, a subset of 49 features was selected using the Cramér coefficient (Siegel and Castellan, 1988) as a univariate measure of the strength of association between each feature and the outcome. While not strictly necessary, this pre selection relieved much of the computational burden on the FEAD algorithm. We acknowledge that this step could possibly remove variables important to the *multivariate* prediction of risk, but our aim here is to illustrate the use of EAD for feature subset selection, not to come up with a clinically unassailable prediction model.

[2] In almost all cases, these factors will be known before labour commences and indicate that delivery will be by Cæsarean section so that failure to progress in labour cannot occur.

Failure to progress												
Feature, ICD code and Range			**Order of addition**									
`Parity`		0-4	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
`MumAge`		0-6	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
`CSections`		0,1	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
`SUPRV HIGH-RISK PREG NEC`	V238	0,1	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
`THREATENED LABOR NEC`	6441	0,1	**6**	**6**	**5**	**5**	**5**	**5**	**5**	**5**	**6**	**8**
`Height`		0-2	**5**	**8**	**6**	**6**	**6**	**6**	**7**	**6**	**9**	**5**
`ANEMIA IN PREGNANCY`	6482	0,1	**7**	**5**	**10**	**9**	**7**	**9**	**6**	**10**	**5**	**7**
`POOR FETAL GROWTH`	6565	0,1	**9**	**11**	**8**	**12**	**10**	**12**	**8**	**7**	**8**	**6**
`PREM SEPARATION PLACENTA`	6412	0,1	13	**10**	18	**13**	**12**	**10**	12	**12**	12	23
`HIGH HEAD AT TERM`	6525	0,1	27	16	**12**	15	**13**	23	25	**9**	20	10
`OTH PLACENTAL CONDITIONS`	6567	0,1	22	14	**11**	28	**9**	**11**	22	29	**11**	18
`ABNORMAL VULVA IN PREG`	6548	0,1	19	15	19	22	**11**	20	**9**	32	21	14
`NeoDxs`		0,1	12	31	16	26	17	**13**	13	**8**	36	17
`HEMORR FROM PLACENT PREV`	6411	0,1	17	**9**	33	**8**	22	28	20	38	35	28
`TRANSVERSE/OBLIQUE LIE`	6523	0,1	28	17	30	18	26	24	19	**11**	14	11
`ANTEPARTUM HEMORR NEC`	6418	0,1	**10**	24	24	20	29	16	36	22	39	32
`INFECTIV DIS IN PREG NEC`	6478	0,1	31	27	28	16	15	25	17	25	17	20
`FETAL DISPROPORTION NOS`	6535	0,1	11	35	20	34	33	33	18	17	26	21
`INDICAT CARE LAB/DEL NEC`	6598	0,1	35	18	26	39	14	19	30	16	15	27
`MALPOSITION NEC`	6528	0,1	**8**	12	34	**7**	24	22	21	37	33	39
`ABO ISOIMMUNIZATION`	6562	0,1	32	22	13	19	34	26	29	18	22	15
`MALPOSITION NOS`	6529	0,1	14	13	17	27	18	38	23	20	37	35
`OBESITY`	2780	0,1	26	26	14	37	25	29	10	**13**	16	38
`EXCESSIVE FETAL GROWTH`	6566	0,1	24	38	22	21	20	42	11	15	25	13
`PREGNANCY COMPL NOS`	6469	0,1	37	29	25	31	21	27	27	24	23	31
`CEPHALIC VERSION NOS`	6521	0,1	29	34	23	17	35	**14**	16	27	18	16
`CERVIX INCOMPET IN PREG`	6545	0,1	18	20	**9**	14	19	21	38	21	27	22
`PREG W POOR REPRODUCT HX`	V235	0,1	30	19	35	24	28	30	15	23	30	26
`UTERINE TUMOR IN PREG`	6541	0,1	16	37	32	**11**	16	41	24	19	29	34
`VOMITING COMPL PREG NEC`	6438	0,1	25	21	31	32	31	15	14	14	19	33
`POLYHYDRAMNIOS`	6579	0,1	23	25	21	41	36	32	26	30	42	19
`PREG W HX OF INFERTILITY`	V230	0,1	20	30	37	23	37	**7**	34	33	40	29
`DIABETES MELLIT IN PREG`	6480	0,1	34	36	15	29	27	35	39	36	**10**	24
`TRANS HYPERTENSION PREG`	6423	0,1	40	32	39	36	**8**	34	35	42	13	12
`RHESUS ISOIMMUNIZATION`	6561	0,1	41	40	36	25	38	36	37	26	32	40
`PREG W POOR OBSTETRIC HX`	V234	0,1	36	23	29	35	39	17	28	34	24	37
`MILD HYPEREMESIS GRAVID`	6430	0,1	21	33	27	38	30	39	33	28	28	36
`BONE DISORDER IN PREG`	6487	0,1	33	39	41	40	23	40	32	31	41	25
`THREATENED ABORTION`	6400	0,1	38	42	40	43	42	37	42	40	31	41
`PROLONGED PREGNANCY`	6459	0,1	47	**7**	**7**	46	40	45	46	45	**7**	**9**
`ABN GLUC TOLERAN IN PREG`	6488	0,1	39	28	38	30	32	31	40	35	38	30
`THREATEN PREMATURE LABOR`	6440	0,1	15	41	43	42	44	18	41	39	43	42
`MILD/NOS PRE-ECLAMPSIA`	6424	0,1	42	45	42	44	46	43	44	43	45	45
`EDEMA IN PREGNANCY`	6461	0,1	43	43	44	33	41	44	31	41	34	44
`HYPERTENS COMPL PREG NOS`	6429	0,1	45	47	45	**10**	45	**8**	45	44	44	46
`ThrptAbortions`		0,1	44	44	46	47	47	46	43	46	46	43
`Support`		0,1	48	48	47	45	43	47	47	47	47	48
`SpontAbortions`		0,1	46	46	48	48	48	48	48	48	48	47
`SocClass`		0-6	49	49	49	49	49	49	49	49	49	49

Table 1: The discrete variables used to predict failure to progress, and the order in which they were added to the working set across 10 of the 20 separate runs of FEAD. Capitalised entries indicate ICD-9 code variables describing maternal conditions. Numbers in bold indicate the subset of features that attained maximum EAD in each run. For the prediction models described in Section 6, polytomous variables (like `MumAge` and `Parity`) were encoded using the *reference variable* method (Hosmer and Lemeshow, 1989).

Using the 49 candidate features, 20 separate runs of FEAD were performed. For each subset of features considered in the forward search, EAD was averaged across 10 random partitions of the data, as explained in Section 4.The left plot of Figure 3 shows the MAD and EAD obtained at each forward selection step of the 20 runs. Clearly, once a certain representational complexity is reached, using additional features to describe the data results in a *decrease* in EAD. Table 1 lists the order in which features were selected in 10 of the 20 runs performed, and shows in bold the subsets which attained maximum EAD. The maximum EAD values ranged between 0.773 and 0.775 across the 20 runs, and occurred when the working set contained 9–15 variables.

The FEAD algorithm adds features to the working set in decreasing order of discriminative power. From a clinical perspective, the ranking of features shown in Table 1 accords with conventional obstetric wisdom. This is notable since the algorithm operated on database information alone, without recourse to expert opinion or prior assumptions.

We can make a more quantitative assessment of EAD as a feature selection criterion by evaluating the prediction performance of models built using selected subsets of features. The right plot of Figure 3 shows the performance achieved by different prediction models on an independent test set of 176 812 cases (SMR2 data from 1989–91). Logistic regression (LogReg) (Hosmer and Lemeshow, 1989), ensembles of neural networks (NetEns and NetMix) (Bishop, 1995) and a lookup table smoothed with a Dirichlet prior (LkpTab) (MacKay and Peto, 1995) were trained[3] using a variety of feature subsets to represent the SMR2 data from 1980–88. We describe results obtained with three different feature subsets:

1. the full 49-variable set (*i.e.*, all features listed in Table 1);
2. a 35-variable subset made up of all features that were in *any* of the subsets that attained maximum EAD in *any* run of FEAD (*i.e.*, if we considered only 10 runs of FEAD, this approach would select predictors with a bold face entry anywhere in their row of Table 1);
3. a 4-variable subset that used only non-ICD code information (`Parity`, `MumAge`, `CSections` and `NeoDxs`). This subset was selected for clinical reasons.

FEAD provides two benchmarks to which we can compare the prediction performance of the different models. At each step of forward selection, the estimate of EAD tells us the performance we could expect from a simple histogram probability density model, while the MAD gives an upper bound on performance. Thus, when a model achieves close to MAD (as is the case with all models using the 4-variable representation), we know that prediction accuracy is limited by the amount of discriminative information in the data rather than shortcomings in the model itself.

Although the 4 non-ICD code features `MumAge`, `Parity`, `CSections` and `NeoDxs` were encoded in a 12 dimensional binary vector, the 1-of-N representation of `MumAge` and `Parity` meant that only 126 distinct feature vectors appeared in the training data. Given that the actual number of patient records outnumbered that by three orders of magnitude, it is not surprising that the NetEns, LkpTab and LogReg models achieved close to MAD in this setting.

As the representation of the SMR2 data becomes more complex, the ability to resolve individual test cases increases. This explains the non-decreasing MAD curve in Figure 3. However, our ability to make good predictions about the risk associated with each vector of patient characteristics does not necessarily increase. While the amount of training data remains the same as in the 4-feature model, the 35 variable representation increases the number of distinct feature vectors to 13 500.

When all 49 features are used to represent the SMR2 data, EAD and the performance of the LkpTab model fall dramatically. The smoothing effect of the Dirichlet prior keeps the LkpTab model performance above that of the simple histogram density model. The number of parameters in the LogReg and NetEns models are, respectively, linear and quadratic functions of the number of features. As such, their generalisation performance is not yet diminished by the high dimensional representation of the data.

Note that the performances of NetEns and LogReg (AUROC of 0.776 and 0.775, respectively) are only slightly greater than the largest values of EAD observed in each run of the FEAD algorithm. In this instance, the histogram probability density model underlying EAD has provided a good indication of the performance of more computationally intensive models.

[3] Full details of the training and testing of these models are given in (Lovell et al., 1997).

7 Discussion

The main strength of EAD as a feature selection criterion is that it provides a performance benchmark, rather than an abstract statistical measure of interclass separation. EAD directly estimates the performance we could expect from a histogram density model of the data. With statistical separation measures, it is necessary to build prediction models using the selected features before we can obtain an idea of the practical value of that data representation.

The complexity of all other models will lie between that of the histogram density model and the null model (*i.e.*, the model that predicts constant risk). Thus, EAD is a sensible measure to evaluate in exploring the discriminative power of feature subsets. The idea of providing a simple, fast benchmark to which other, more complicated prediction strategies could be compared has also been put forward by Holte (1993), though not in the context of feature selection.

Because it is based on AUROC as a measure of discrimination, EAD can be used in situations where misclassification costs are not specified. AUROC has been criticised by Jiang et al. (1996), who state that it "is not a meaningful summary of clinical diagnostic performance when high sensitivity must be maintained". We must stress that we use AUROC primarily to direct our search for discriminative subsets of features. Proper evaluation of models trained on the selected features requires additional measures of performance.

To avoid distributional assumptions in calculating MAD and EAD, we have constrained data to be represented by discrete valued features but, with a finite quantity of data, we cannot elude the curse of dimensionality. The more features we use to represent our data, the less accurate our estimates of the likelihood ratios in each bin.

Increasing the number of features used to represent a fixed amount of data inevitably leads to bins that contain only one class of data. Such bins give rise to infinite likelihood ratio estimates, but this does not cause difficulties when the area under the ROC curve is used as a measure of discrimination. Hills (1967) proposed a similar framework to the one we have described but used information divergence as a measure of discrimination. Hills dealt with the problem of infinite likelihood ratio estimates by "adding one observation" to each class in any bin that initially contained data from only one class. This approach is equivalent to forming a Bayesian estimate of the probability of adverse outcome using a uniform prior.

In this paper, we have demonstrated FEAD on a large database, but it has also been successfully applied to smaller amounts of data (~ 300000 and ~ 8000 records (Lovell et al., 1996)). Experimental results suggest that EAD is a good feature selection criterion when the number of different feature combinations *actually appearing in the data* is small enough for reasonably accurate likelihood estimates to be obtained.

Because EAD is estimated on the basis of several random partitions of a dataset, different searches that use EAD as a criterion will return slightly different results (see Table 1). This is a natural consequence of the uncertainty about the features that are important in solving a given problem. This uncertainty arises because we are trying to estimate model parameters with a finite amount of data. As far as we are aware, most other approaches to feature selection do not perturb the data used to estimate the "goodness" of feature subsets. By avoiding the issue of how sensitive a feature selection criterion is to different samples of the available data, these methods can present an over-confident statement of the discriminative power of variables. EAD straightforwardly accounts for uncertainty about the importance of different features.

8 Conclusion

The methods described in this paper were motivated by a need to determine which features, individually or in combination, would form a useful representation for making predictions about new data. Using the area under the ROC curve as a non-parametric measure of discrimination between two classes, we have shown how to calculate MAD: the *maximum attainable discrimination* when a given set of discrete valued features is used to represent a dataset. By using a "hold out" data set, it is straightforward to estimate EAD: the discrimination we can *expect* to obtain when these features are used to represent new data.

We have shown how EAD can be used as a feature selection criterion. Since it corresponds to the performance expected from the bias-free histogram probability density model of data, EAD provides a meaningful performance benchmark. However, in situations where the ratio of data to the number of distinct features is low, EAD gives pessimistic estimates of the prediction performance attainable by suitably biased models.

EAD can be used as a feature selection criterion in settings where Bayes risk cannot be evaluated because of variable, or unspecified misclassification costs. This was the case in the obstetric risk prediction problem we described and it remains to be seen whether EAD will be of benefit in other problem domains.

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 SMR2 data. We also thank Dr Jim Chalmers of the Information and Statistics Division of the National Health Service in Scotland. 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 post of research associate in the QAMC project.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387–415.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley.

Green, D. and Swets, J. (1966). *Signal Detection Theory and Psychophysics*. John Wiley, New York.

Hanley, J. and McNeil, B. (1982). Maximum attainable discrimination and the utilization of radiologic examinations. *Journal of Chronic Disease*, *35*, 601–611.

Hills, M. (1967). Discrimination and allocation with discrete data. *Applied Statistics*, *16*, 237–250.

Holte, R. C. (1993). Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*, *11*(1), 63–91.

Hosmer, D. and Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York; Chichester.

Jiang, Y. L., Metz, C. E., and Nishikawa, R. M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, *201*(3), 745–750.

Kittler, J. (1975). Mathematical methods of feature selection in pattern recognition. *International Journal of Man-Machine Studies*, *7*, 609–637.

Kittler, J. (1978). Features set search algorithms. In Chen, C. H. Ed., *Pattern Recognition and Signal Processing*, pp. 41–60. Sijthoff and Noordhoff, The Netherlands.

Lovell, D. R., Dance, C. R., Niranjan, M., Prager, R. W., and Dalton, K. J. (1996). Using upper bounds on discrimination to select discrete valued features. In Usui, S., Tohkura, Y., Katagiri, S., and Wilson, E. Eds., *Neural Networks for Signal Processing VI*, pp. 233–242 Keihanna, Japan.

Lovell, D. R., Scott, M. J. J., Niranjan, M., Prager, R. W., Dalton, K. J., and Derom, R. (1997). On the use of Expected Attainable Discrimination for feature selection in large scale medical risk prediction problems. Tech. rep. CUED/F-INFENG/TR299, Cambridge University Engineering Department.

MacKay, D. J. C. and Peto, L. (1995). A Hierarchical Dirichlet Language Model. *Natural Language Engineering*, *1*(3), 1–19.

Siedlecki, W. and Sklansky, J. (1988). On automatic feature selection. *International Journal of Pattern Recognition and Artificial Intelligence*, *2*(2), 197–220.

Siegel, S. and Castellan, N. J. (1988). *Nonparametric statistics for the behavioural sciences*. McGraw-Hill.

Van Campenhout, J. M. (1982). Topics in Measurement Selection. In Krishnaiah, P. R. and Kanal, L. N. Eds., *Handbook of Statistics, Vol. 2*, pp. 793–803. North-Holland.

Weiss, S. M. and Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural networks, machine learning, and expert systems*. Morgan Kaufmann, San Mateo.

World Health Organization (1975). *International Classification of Diseases, Ninth Revision*. Geneva: World Health Organization.