
Improving Classification Performance by Exploring the Role of Cost Matrices in Partitioning the Estimated Class Probability Space

Deirdre B. O’Brien

Robert M. Gray

Department of Electrical Engineering, Stanford University, Stanford, CA 94305-9510

DBOBRIEN@STANFORD.EDU

RMGRAY@STANFORD.EDU

Cost-sensitive classification, Bayes risk, empirical risk minimization, estimated class probabilities

Abstract

A cost matrix associates a cost with each error type; the costs are defined in the problem specification and are related to the real world meaning of the application at hand. In probability estimation based classifiers, cost matrices also determine how the estimated class probability space is divided into class regions. We explore this second role of cost matrices. There are two main contributions in this paper. First, we describe how specific variations in the cost matrix relate to changes in the partitioning of the estimated probability space. In addition, we provide a set of methods which improve this partition by seeking to minimize the empirical Bayes risk.

1. Introduction

Classification algorithms based on class probability estimation, such as naïve Bayes, provide an elegant means to incorporate misclassification costs. For classification problems with two classes the theory dictates that samples should be assigned to class 1 if the estimated probability for class 1 is greater than $(c_{1|2} - c_{2|2}) / (c_{1|2} - c_{2|2} + c_{2|1} - c_{1|1})$ and to class 2 otherwise, where $c_{i|j}$ is the real-life cost of assigning to class i a sample which belongs to class j . However, researchers often observe that the overall risk is reduced if a different threshold is used. The threshold is chosen to minimize empirical risk (Lachiche & Flach, 2003).

This paper provides an extension of this empirical risk minimizing thresholding method to multi-class prob-

lems. We begin in section 2 by studying the *local* and *global* risks in cost sensitive classification. Cost matrices specify penalties for particular mistakes and so allow for evaluation of classifier performance in terms of expected *global* risk. When used in probability estimation algorithms, cost matrices define the boundaries between classes in the probability estimate space based on the estimated *local* risk. In this role, we refer to the cost matrix as a *boundary* matrix. In section 3, we propose replacing the cost matrix in this role with an alternate *boundary* matrix, which is matched to the probability estimates. Some properties of the relationship between the cost matrix and the partition are detailed in section 4. These properties allow for a greater understanding of the methods used to select the alternative *boundary* matrices. The algorithmic details are presented in section 5. In section 6 we present experimental results showing the effectiveness of this approach.

2. Background

Many classification algorithms are implicitly probability estimators. Generative methods such as linear, quadratic or regularized discriminant analysis, Gaussian mixture methods, naïve Bayes etc. seek to estimate the conditional distribution of the features given the class. Combining these estimates with priors (either available or estimated from the training set) leads to estimated probabilities for class membership. Other methods such as logistic regression and probability estimation trees estimate the class probabilities directly. For x in the domain of the feature vectors \mathcal{X} , and k in the domain of the class labels \mathcal{K} (for simplicity we label the classes $\{1, 2, \dots, K\}$), the estimated class probabilities are given by $\{\hat{p}_1(x), \hat{p}_2(x), \dots, \hat{p}_K(x)\}$. We use y to denote the true class and $\hat{y}(x)$ the assigned class.

In cost sensitive classification we are given a cost matrix C ; the j^{th} element in the i^{th} row is $c_{i|j}$, the real-world cost of assigning a sample to class i when its true

This work was partially supported by Norsk Elektro Optikk and by NSF Grant No. NSF CCR-0309701.

Appearing in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005. Copyright 2005 by the author(s)/owner(s).

class is j . The estimated *local* risk of assigning a sample to class i is then, $\hat{r}_i(C, x) = \sum_{k \in \mathcal{K}} c_{i|k} \hat{p}_k(x)$. Since the goal of classification is to minimize the average or *global* risk given by

$$R_C = E_{X,Y}[c_{\hat{y}(X)|Y}] = E_X[r_{\hat{y}(X)}(C, X)], \quad (1)$$

the usual approach is to minimize the estimated *local* risk and so assign a sample with feature vector x to

$$\hat{y}(x) = \arg \min_i \hat{r}_i(C, x) = \arg \min_i \sum_{k \in \mathcal{K}} c_{i|k} \hat{p}_k(x). \quad (2)$$

We consider this method of cost-sensitive classification a *local risk* approach since the local risk for each class assignment at x is first estimated and then the sample is assigned to the class with minimum estimated risk.

If the probability estimation is accurate, this probability-based approach will indeed lead to minimum average risk classification. A number of previous methods in cost-sensitive classifier design have sought to improve these probability estimates (Zadrozny & Elkan, 2002; Domingos, 1999). However, accurate probability estimation is not essential in classification and for 0/1 costs many classification algorithms perform very well despite their poor performance as probability estimators (Friedman, 1997).

Examining (2), we see that the classification is performed by partitioning the space of estimated class probabilities. Each partition cell represents assignment to a particular class and the class boundaries are determined by the cost matrix. In this role, we refer to the cost matrix as a *boundary* matrix. The minimum risk classification requires only that for each x the class probability estimate vector lies in the same partition cell as the true class probability vector.

An alternate approach to classifier design, which we refer to as a *global risk* approach, involves finding the best boundaries in the sense of minimizing empirical risk. The empirical risk based on a validation set \mathcal{V} is \hat{R}_C ,

$$\hat{R}_C = \frac{1}{|\mathcal{V}|} \sum_{(x_i, y_i) \in \mathcal{V}} c_{\hat{y}(x_i)|y_i} = \sum_{k=1}^K \pi_k \sum_{j=1}^K c_{j|k} \hat{P}_{\hat{Y}|Y}(j|k),$$

where $\hat{P}_{\hat{Y}|Y}(j|k)$ can be estimated from counts on \mathcal{V} . Such classifiers usually require the class boundaries to be of a particular form and often rather than seeking to minimize \hat{R}_C directly, they optimize some other function of the samples which is correlated with \hat{R}_C . Examples of global risk classifiers are support vector machines and classification trees. In this design paradigm the function of the cost matrix is to penalize the incorrect assignments; this is the role considered when the cost matrix is given in the problem specification.

The goal of classification is to minimize R_C of equation (1). This is achieved by *global risk* methods if the approach taken to minimize \hat{R}_C is effective at minimizing R_C . It will also be achieved by *local risk* probabilistic models if the probability estimation step is such that the estimated probabilities lie in the same partition cell of the simplex as the true probabilities, i.e. if the minimum estimated risk assignment in (2) equals the minimum true risk assignment for all x .

2.1. Dealing with Inaccurate Probability Estimates in Tasks with Two Classes

For the two class problem the assignment in (2) is,

$$\hat{y}(x) = \begin{cases} 1 & \text{if } \hat{p}_1(x) > \frac{c_{1|2} - c_{2|2}}{c_{1|2} - c_{2|2} + c_{2|1} - c_{1|1}} \\ 2 & \text{if } \hat{p}_1(x) < \frac{c_{1|2} - c_{2|2}}{c_{1|2} - c_{2|2} + c_{2|1} - c_{1|1}} \end{cases}. \quad (3)$$

A common approach is to replace the right hand side of the inequalities in (3) with some other threshold t (Friedman, 1997). The threshold t is chosen to minimize the empirical risk, which for a two-class task can be written as

$$\hat{R}_C = \pi_1 c_{1|1} + \pi_2 c_{1|2} + \hat{P}_{\hat{Y}|Y}(2|1) \pi_1 (c_{2|1} - c_{1|1}) - \hat{P}_{\hat{Y}|Y}(2|2) \pi_2 (c_{1|2} - c_{2|2}).$$

Minimizing \hat{R}_C is equivalent to minimizing

$$\hat{P}_{\hat{Y}|Y}(2|1) \frac{\pi_1 (c_{2|1} - c_{1|1})}{\pi_2 (c_{1|2} - c_{2|2})} - \hat{P}_{\hat{Y}|Y}(2|2).$$

An ROC curve steps through thresholds from $t = 0$ to $t = 1$ and plots the corresponding values of $\hat{P}_{\hat{Y}|Y}(2|2)$ against $\hat{P}_{\hat{Y}|Y}(2|1)$. Given the prior and cost information, we can select the tangent to the ROC curve with slope $(\pi_1 (c_{2|1} - c_{1|1})) / (\pi_2 (c_{1|2} - c_{2|2}))$. The threshold used to generate this ROC point is that which should be used to replace the right hand side of the inequalities in (3) (Noe, 1983).

If the classifier design algorithm returns accurate probability estimates then the threshold returned by the ROC analysis will equal that given in (3), unfortunately this is often not the case. An example in (Lachiche & Flach, 2003) shows the significant difference in the classifier performance when the threshold is chosen to minimize empirical (global) risk versus when the threshold in (3) is used.

3. Matched Boundary Matrices

An alternative view of the ROC thresholding approach is to note that choosing a threshold t effectively changes the cost matrix used to partition the

probability estimates space. To extend the two-class thresholding method to multi-class problems, we use this idea to separate the two roles of the cost matrix. The role of the cost matrix in penalizing misclassifications is maintained, while in its role as a boundary matrix, the cost matrix is replaced by a more effective matrix.

Therefore, our approach is to find a “matched” boundary matrix B with elements b_{ij} that replaces C in (2). In the two class case, (3) becomes

$$\hat{y}^B(x) = \begin{cases} 1 & \text{if } \hat{p}_1(x) > \frac{b_{1|2} - b_{2|2}}{b_{1|2} - b_{2|2} + b_{2|1} - b_{1|1}} \\ 2 & \text{if } \hat{p}_1(x) < \frac{b_{1|2} - b_{2|2}}{b_{1|2} - b_{2|2} + b_{2|1} - b_{1|1}} \end{cases}.$$

In the general case, we are effectively changing the estimated local risk from $\hat{r}_i(C, x)$ to $\hat{r}_i(B, x)$. While $\hat{r}_i(B, x)$ is not meaningful in terms of the true misclassification costs, it is used in determining the boundaries associated with B . The class assignment becomes

$$\hat{y}^B(x) = \arg \min_i \hat{r}_i(B, x) = \arg \min_i \sum_{k \in \mathcal{K}} b_{i|k} \hat{p}_k(x). \quad (4)$$

Since the true costs determine the performance of the classifier, the true cost matrix is still used in the global empirical risk. Thus, for the classifier using the matched boundary matrix

$$\hat{R}_C = \frac{1}{|\mathcal{V}|} \sum_{(x_i, y_i) \in \mathcal{V}} c_{\hat{y}^B(x_i)|y_i}.$$

The cost matrix provides a division of the estimated probability simplex into classes, which would be optimal if $\hat{p}_k(x)$ and $p_k(x)$ lie in the same partition cell for all x . The matched boundary matrix seeks to provide an optimal partitioning of inaccurate probability estimates into classes and so minimize the empirical risk. B represents a “matched” boundary matrix in the sense that its elements are chosen to match the probability estimation and the cost matrix.

For a given probability estimation algorithm, changes to B in (4) cause changes in the $P_{\hat{Y}|Y}$ probabilities which describe the classifier’s position on the ROC curve. By varying the boundary matrix we can trace out the ROC curve for the given probability estimation algorithm and assignments based on local risk estimation. We seek B yielding $P_{\hat{Y}|Y}$ probabilities which minimize risk for the true cost matrix.

This could be viewed as an extension of a method described in (Mossman, 1999) where for three class tasks the $\hat{p}(x)$ simplex is divided as follows:

$$\hat{y}(x) = \begin{cases} 3 & \text{if } \hat{p}_3(x) > \delta_1 \\ 2 & \text{if } \hat{p}_3(x) \leq \delta_1 \text{ and } \hat{p}_2(x) - \hat{p}_1(x) > \delta_2 \\ 1 & \text{if } \hat{p}_3(x) \leq \delta_1 \text{ and } \hat{p}_2(x) - \hat{p}_1(x) \leq \delta_2. \end{cases}$$

Mossman’s approach is equivalent to setting

$$B = \begin{bmatrix} 0 & 1 & L - \frac{\delta_2}{1 - \delta_2} \\ \frac{1 + \delta_2}{1 - \delta_2} & 0 & L \\ \frac{\delta_1}{1 - \delta_1} L & \frac{\delta_1}{1 - \delta_1} L & 0 \end{bmatrix}, \quad L \gg 1.$$

In general, finding the matched boundary matrix is equivalent to finding linear separators of the classes in the estimated probability space. The method of matched boundary matrices differs from other global risk approaches in that the boundaries are found in the probability estimate space rather than in the original feature space.

4. Properties of Boundary Matrices

In this section we describe a number of features of cost matrices which relate to their role as boundary matrices. In section 5 these properties are used in designing algorithms to find effective matched boundary matrices. Some of these properties may be known to practitioners who work with cost matrices, but to the knowledge of the author they have not been presented in this context previously. Figure 1 illustrates these properties for a three class task.

The probability estimates are confined to the simplex where $\sum_{k \in \mathcal{K}} \hat{p}_k = 1$ and $\hat{p}_k \geq 0, \forall k \in \mathcal{K}$. However, when describing the boundary matrix partitions it is natural to consider the hyperplane where $\sum_{k \in \mathcal{K}} \hat{p}_k = 1$ without requiring $\hat{p}_k \geq 0, \forall k \in \mathcal{K}$. We refer to this hyperplane as the *extended simplex*. We term the boundary matrix A since we are considering boundary matrices in general rather than the true cost matrix C or the matched boundary matrix B .

We make a single assumption about classification tasks and the probability estimates: if the estimated probabilities are all zero except for one class with an estimated probability of one, then the sample will be assigned to that class for all boundary matrices. Even for very inaccurate probability estimates it seems reasonable to assume that such a confident guess for a particular class would be assigned as such. This assumption restricts our study to boundary matrices which satisfy,

$$a_{j|j} < a_{i|j}. \quad (5)$$

For boundary matrix A , we write the local risk for class i as $\rho_i(A, \cdot)$ where \cdot is a vector of class probabilities or probability estimates. Thus $r_i(A, x) = \rho_i(A, p(x))$ and $\hat{r}_i(A, x) = \rho_i(A, \hat{p}(x))$, emphasizing that the estimated local risk depends on x only through $\hat{p}(x)$.

The boundary between classes i and j in the estimated

probability space is given by,

$$\sum_{k \in \mathcal{K}} a_{i|k} \hat{p}_k \stackrel{i}{\leq} \sum_{k \in \mathcal{K}} a_{j|k} \hat{p}_k. \quad (6)$$

This boundary separates the extended simplex into half-planes based on whether $\rho_i(A, \hat{p}) < \rho_j(A, \hat{p})$. Because of (5), such a boundary must exist. Combining all two-class boundaries describes the classifier. A sufficient (but not necessary) condition for two boundary matrices to result in the same partition of the probability estimates space is that all two-class boundaries are equivalent. Thus, adding a constant to all elements of a boundary matrix or multiplying all elements by a positive constant does not effect the partitioning.

Property 1: *There is exactly one point in \hat{p} space which has equal estimated local risk for all classes (we term this point the equal local risk point).*

Proof: This point is the solution to K simultaneous equations. $K - 1$ equations are of the form:

$$\sum_{k \in \mathcal{K}} (a_{i|k} - a_{K|k}) \hat{p}_k = 0, i \in \{1, 2, \dots, K - 1\}, \quad (7)$$

the final equation requires that $\sum_{k=1}^K \hat{p}_k = 1$. Note that this point may not satisfy $\hat{p}_k \geq 0, \forall k$.

Property 2: *Adding a constant to each term in any column of a boundary matrix does not effect the partitioning.*

Proof: Adding the constant α to column m gives the following boundary between classes i and j

$$\sum_{k \in \mathcal{K}} a_{i|k} \hat{p}_k + \alpha \hat{p}_m \stackrel{i}{\leq} \sum_{k \in \mathcal{K}} a_{j|k} \hat{p}_k + \alpha \hat{p}_m.$$

This boundary is clearly equivalent to (6) for all \hat{p} . Thus the boundaries are unchanged and so the partition of the estimated probability space is unchanged.

For any boundary matrix satisfying (5), property 2 can be used to find a new boundary matrix in which all diagonal elements are zero and all off-diagonal elements are non-negative, and which has the same partitioning of the probability simplex as the original matrix.

Property 3: *Adding a constant to each term in any row of a boundary matrix provides a new boundary matrix whose class boundaries are parallel to those of the original.*

Proof: Adding the constant α to the m^{th} row affects only the estimated local risk of class m , which is increased by $\sum_{k \in \mathcal{K}} \alpha \hat{p}_k = \alpha$. Thus $\rho_m(A, \hat{p})$ is

increased by α for all \hat{p} . Since the local risk functions are all linear in the probability estimates, the new class boundaries will be parallel to the original boundaries. To maintain constraint (5) we require $\alpha \leq \min_{k \neq m} (b_{k|m} - b_{m|m})$.

Combining properties 2 and 3 gives a parallel transformation which preserve zero elements on the diagonal:

Property 3b: *Adding a constant to each term in the m^{th} row of a boundary matrix and subtracting the same constant from each term in the m^{th} column, provides a new boundary matrix whose class boundaries are parallel to those of the original.*

Property 4: *Multiplying all elements in column m of the boundary matrix by a positive constant moves the equal local risk point along the line joining the original equal local risk point to the corner of the simplex where $\hat{p}_m = 1$. The intersections of the class boundaries with the $\hat{p}_m = 0$ hyperplane are unchanged.*

Proof: All points along the line joining a point \hat{q} to the corner where $\hat{p}_m = 1$ can be written as a linear combination of \hat{q} and e_m , and so are of the form $\hat{q}_j^\alpha = \alpha \hat{q}_j$ for $j \neq m$ and $\hat{q}_m^\alpha = \alpha \hat{q}_m + (1 - \alpha)$.

Let \hat{q} be the equal local risk point for the original boundary matrix. Solving the $K - 1$ equations given in (7) allows us to write $\hat{q}_k = s_k \hat{q}_m$, where s_k depends on A and $s_m = 1$. The unity constraint requires

$$\sum_{k=1}^K s_k \hat{q}_m = 1 \Rightarrow \hat{q}_m = \frac{1}{\sum_{k=1}^K s_k}, \hat{q}_j = \frac{s_j}{\sum_{k=1}^K s_k}. \quad (8)$$

Let \hat{v} be the equal local risk point for the new boundary matrix when column m is multiplied by w , then for $i \in \{1, 2, \dots, K - 1\}$,

$$\sum_{k=1, k \neq m}^K (a_{i|k} - a_{K|k}) \hat{v}_k + w(a_{i|m} - a_{K|m}) \hat{v}_j = 0, \quad (9)$$

and $\sum_{k=1}^K \hat{v}_k = 1$. Let $\hat{u}_k = \hat{v}_k, k \neq m$ and $\hat{u}_m = w \hat{v}_m$, then (9) becomes,

$$\sum_{k=1}^K (a_{i|k} - a_{K|k}) \hat{u}_k = 0, i \in \{1, 2, \dots, K - 1\},$$

which can be solved to write $\hat{u}_k = s_k \hat{u}_m$ as before. Combined with the unity constraint on \hat{v} this yields

$$\sum_{k=1}^K s_k \hat{u}_m = \sum_{k=1}^K \hat{u}_k = \sum_{k=1}^K \hat{v}_k - \hat{v}_m + \hat{u}_m = 1 + \hat{u}_m \left(1 - \frac{1}{w}\right),$$

giving

$$\hat{u}_m = \frac{1}{\sum_{k=1}^K s_k + \left(\frac{1}{w} - 1\right)}.$$

Thus for $k \neq m$

$$\hat{v}_k = \frac{s_k}{\sum_{j=1}^K s_j + (\frac{1}{w} - 1)} = \hat{q}_k \times \frac{\sum_{j=1}^K s_j}{\sum_{j=1}^K s_j + (\frac{1}{w} - 1)},$$

or $\hat{v}_k = \alpha \hat{q}_k$ for $k \neq m$. Therefore, \hat{v} lies on the line joining \hat{q} to e_m . $\alpha = 1/(1 + \hat{q}_m(1/w - 1))$ and so depends on the m^{th} component of the original equal local risk point. Note that by (8), if $\sum_{k=1}^K s_k > 1$ then $0 < \hat{q}_m < 1$. In this case if $w < 1$, then $\alpha < 1$ and the new equal local risk point is closer to the $\hat{p}_m = 1$ corner of the simplex than \hat{q} ; conversely if $w > 1$ then $\alpha > 1$ and the new equal local risk point is further from the $\hat{p}_m = 1$ corner of the simplex than \hat{q} . By considering the three ranges $\hat{q}_m < 0, 0 \leq \hat{q}_m \leq 1$ and $\hat{q}_m > 1$ we can show that \hat{v} will lie within the simplex if and only if \hat{q} does.

The i, j class boundary for the new matrix is

$$\sum_{k=1, k \neq m}^K a_{i|k} \hat{p}_k + \alpha a_{i|m} \hat{p}_m \stackrel{i}{\leq} \sum_{k=1, k \neq m}^K a_{j|k} \hat{p}_k + \alpha a_{j|m} \hat{p}_m,$$

which at $\hat{p}_m = 0$ is equal to that of the original class boundary at $\hat{p}_m = 0$.

Property 5 *Multiplying all elements in row m by a positive constant w , moves the equal local risk point along the line where all classes but class m have equal estimated local risk. The intersection of the new i, j boundary with the old i, j boundary does not depend on w .*

Proof: Referring back to property 1, we see the only equation in the K simultaneous equations affected by this multiplication is $\sum_{j=1}^K (a_{m|j} - a_{K|j}) \hat{p}_j = 0$. Removing this constraint returns a line in the space which satisfies all the other constraints, therefore changing w moves the equal local risk point along this line. This is in fact the case for any change in the boundary matrix that affects only row m .

Boundaries involving class m become

$$\sum_{k \in \mathcal{K}} a_{i|k} \hat{p}_k \stackrel{i}{\leq} w \sum_{k \in \mathcal{K}} a_{m|k} \hat{p}_k. \quad (10)$$

The extended simplex is a $K - 1$ dimensional hyperplane, so this boundary is a $K - 2$ dimensional hyperplane. Consider the $K - 3$ dimensional hyperplane

$$H_{im} = \left\{ \hat{p} \left| \sum_{k \in \mathcal{K}} a_{i|k} \hat{p}_k = 0, \sum_{k \in \mathcal{K}} a_{m|k} \hat{p}_k = 0, \sum_{k \in \mathcal{K}} \hat{p}_k = 1 \right. \right\}.$$

All points in H_{im} satisfy (10) for any value of w . Thus, we can think of the class boundary between i and m as hinged at the hyperplane H_{im} and the equal local risk point hinged along the line described above.

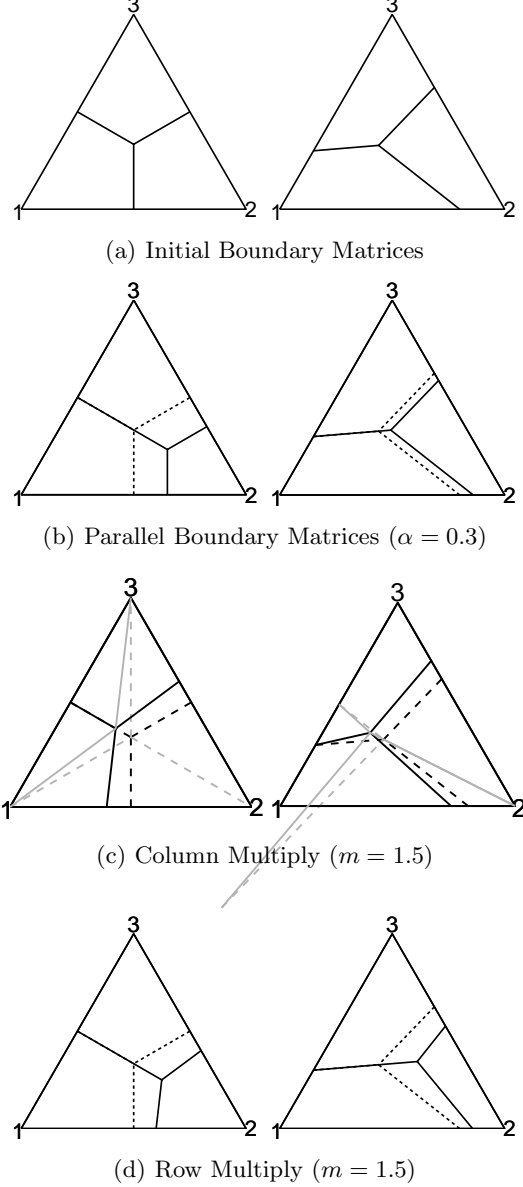


Figure 1. This figure illustrates the properties described in section 4 for a three class task. Each triangle represents the \hat{p} simplex. The corners are labeled 1,2 and 3 corresponding to $(\hat{p}_1, \hat{p}_2, \hat{p}_3) = (1, 0, 0), (0, 1, 0)$ and $(0, 0, 1)$ respectively. For any probability \hat{p} , its position in the simplex is such that the perpendicular distance to the edge opposite corner i , is \hat{p}_i . In each case the partition is shown by solid lines; the class i assignment region is the partition cell including corner i . All manipulations are applied to the class 2 elements of the boundary matrix. The initial boundary matrix for the figures on the left is the 0/1 cost matrix, and for those on the right it is $\begin{bmatrix} 0 & 1 & 7 \\ 4 & 0 & 3 \\ 3 & 5 & 0 \end{bmatrix}$. The dashed

lines show the initial boundary matrix partitions, the gray lines show boundary extensions or other lines necessary to illustrate the properties.

5. Selecting Matched Boundary Matrices

In section 3, we motivated the use of matched boundary matrices in probability estimate partitioning. The properties in section 4 provide ideas for finding an appropriate matched boundary matrix. We can restrict our attention to matched boundary matrices that have zeros along the diagonal and non-negative off-diagonal terms. We consider a general case which allows any such boundary matrix. We also consider restrictive approaches which force the matched boundary matrix to be a transformation of the original cost matrix. These restrictions are motivated by the wide variation in partitions, achievable by simple manipulations of the cost matrix as described in section 4 and illustrated in figure 1. We consider a number of such transformations:

Parallel: B is obtained from C , by adding α_m to row m and subtracting α_m from column m , for $m \in \{1, 2, \dots, K-1\}$. The optimization task is to find the best such α_m 's in the sense that classification using (4) minimizes the empirical risk.

Row multiply: B is obtained from C , by multiplying row m by w_m for $m \in \{1, 2, \dots, K-1\}$. The optimization task is to find the best such w_m 's in the sense that classification using (4) minimizes the empirical risk.

Column multiply: The approach is as for the row multiply case, but here w_m multiplies column m .

It is easy to show that the probability weighting approach described in (Lachiche & Flach, 2003) is equivalent to setting the boundary matrix equal to a column multiplied 0/1 cost matrix. Our optimization methods use an approach similar to theirs.

5.1. Optimization

We describe a greedy method of finding the required parameters. While such a method will find only a local rather than global optimum, it provides a good indication of the effectiveness of the overall approach. Future work will investigate more sophisticated schemes.

In each case we begin by initializing the matched boundary matrix B to the real cost matrix C . We then update each parameter one at a time in a greedy fashion. The update step is described below for each method. We investigated looping through the parameters until no further improvement was seen on the validation set. However, the changes made to the matched boundary matrix after the first iteration often had little effect on the performance. Therefore, only one pass

through the parameters was performed. The parameters were updated in order of the class size to which they related; those associated with the largest class were updated first and so on.

The following result is used in many of the methods described: Let $\hat{y}_m(\hat{p}) = \arg \min_{i \in \mathcal{K}, i \neq m} \sum_{k \in \mathcal{K}} b_{i|k} \hat{p}_k$, and let the matched boundary matrix be altered such that only two-class boundaries of the form (6) which include m as one class are changed. Then following the alteration, samples with probability estimates \hat{p} will be assigned to either class m or class $\hat{y}_m(\hat{p})$.

Parallel: (|| in table 1) For a given matched boundary matrix B and a particular class m , we seek to find a constant α_m that defines a new matched boundary matrix \tilde{B} with $\tilde{b}_{i|j} = b_{i|j}$, $m \notin \{i, j\}$, $\tilde{b}_{m|j} = b_{m|j} + \alpha_m$, $\tilde{b}_{i|m} = b_{i|m} - \alpha_m$, $\tilde{b}_{m|m} = 0$. (Having found \tilde{B} we will then set $B = \tilde{B}$.) Note that only boundaries involving class m are altered by this update. The new boundaries are

$$\sum_{k \in \mathcal{K}} b_{i|k} \hat{p}_k - \alpha_m \hat{p}_m \stackrel{i}{\leq} \sum_{k \in \mathcal{K}} b_{m|k} \hat{p}_k + \alpha_m \sum_{\substack{k \in \mathcal{K} \\ k \neq m}} \hat{p}_k.$$

Thus a sample with probability estimates \hat{p} will be assigned to classes $\hat{y}_m(\hat{p})$ or m according to

$$\sum_{k \in \mathcal{K}} b_{\hat{y}_m(\hat{p})|k} \hat{p}_k - \sum_{k \in \mathcal{K}} b_{m|k} \hat{p}_k \stackrel{\hat{y}_m(\hat{p})}{\leq} \alpha_m. \quad (11)$$

For each sample in the validation set, we find the cost (using the true specified cost matrix, C) of assigning it to class m and the cost of assigning it to class $\hat{y}_m(\hat{p}(x))$. We then select α_m which minimizes the empirical risk when the validation set is classified according to (11). There will in fact be a range of values which minimize the empirical risk; we select α_m equal to the geometric mean of the end points of this range.

Row multiply: (\times R in table 1) For a given matched boundary matrix B and a particular class m , we seek to find a constant w_m that defines a new matched boundary matrix \tilde{B} with $\tilde{b}_{i|j} = b_{i|j}$, $i \neq m$, $\tilde{b}_{m|j} = w_m b_{m|j}$. Again only boundaries involving class m are altered by this update. The new boundaries are at

$$\sum_{k \in \mathcal{K}} b_{i|k} \hat{p}_k \stackrel{i}{\leq} w_m \sum_{k \in \mathcal{K}} b_{m|k} \hat{p}_k,$$

so a sample with probability estimates \hat{p} will be assigned to classes $\hat{y}_m(\hat{p})$ or m according to

$$\frac{\sum_{k \in \mathcal{K}} b_{\hat{y}_m(\hat{p})|k} \hat{p}_k}{\sum_{k \in \mathcal{K}} b_{m|k} \hat{p}_k} \stackrel{\hat{y}_m(\hat{p})}{\leq} w_m. \quad (12)$$

As in the previous case we select w_m so that the empirical risk is minimized.

Column multiply: ($\times C$ in table 1) Finding the column multiplier is not as straightforward as the other optimizations described, since altering this multiplier effects all class boundaries. For a given matched boundary matrix B and a particular class m , we seek to find a constant w_m that defines a new matched boundary matrix \tilde{B} with $\tilde{b}_{i|j} = b_{i|j}$, $j \neq m$, $\tilde{b}_{i|m} = w_m b_{i|m}$. We considered both a method which found an optimal value of w_m and a simpler method described here. Both methods performed similarly when evaluated on unseen test sets and so we include only the simpler suboptimal approach here.

For any \hat{p} we consider assignment only to m or to $\bar{m}(\hat{p})$, where $\bar{m}(\hat{p}) = \arg \min_{i \in \mathcal{K}, i \neq m} \sum_{\substack{k \in \mathcal{K} \\ k \neq m}} b_{i|k} \hat{p}_k$. The assignment rule used is:

$$\sum_{\substack{k \in \mathcal{K} \\ k \neq m}} b_{\bar{m}(\hat{p})|k} \hat{p}_k + w_m b_{\bar{m}(\hat{p})|m} \hat{p}_m \stackrel{\bar{m}(\hat{p})}{\leq} \sum_{\substack{k \in \mathcal{K} \\ k \neq m}} b_{m|k} \hat{p}_k$$

$$\Leftrightarrow w_m \stackrel{\bar{m}(\hat{p})}{\leq} \frac{\sum_{\substack{k \in \mathcal{K} \\ k \neq m}} (b_{m|k} - b_{\bar{m}(\hat{p})|k}) \hat{p}_k}{b_{\bar{m}(\hat{p})|m} \hat{p}_m}.$$

The value of w_m was then found in a similar way to the method used to find α_m in the parallel approach.

General Update: (Gen in table 1) The final approach is to allow each non-diagonal element of the cost matrix to be updated. As before we start with B equal to the real cost matrix C and use a greedy approach to set each element consecutively.

For the (m, n) update, we seek to find a constant $\alpha_{m|n}$ that defines a new matched boundary matrix given by \tilde{B} with $\tilde{b}_{i|j} = b_{i|j}$, $(i, j) \neq (m, n)$, $\tilde{b}_{m|n} = \alpha_{m|n}$. Only boundaries involving class m are altered by this update. The new boundaries are at

$$\sum_{k \in \mathcal{K}} b_{i|k} \hat{p}_k \stackrel{i}{\leq} \sum_{\substack{k \in \mathcal{K} \\ k \neq n}} b_{m|k} \hat{p}_k + \alpha_{m|n} \hat{p}_n.$$

Thus a sample with probability estimates \hat{p} will be assigned to classes $\hat{y}_{\bar{m}}(\hat{p})$ or m according to

$$\frac{\sum_{k \in \mathcal{K}} b_{\hat{y}_{\bar{m}}(\hat{p})|k} \hat{p}_k - \sum_{\substack{k \in \mathcal{K} \\ k \neq n}} b_{m|k} \hat{p}_k}{\hat{p}_n} \stackrel{m}{\leq} \alpha_{m|n}. \quad (13)$$

For each sample in the validation set, we find the cost (using the real cost matrix C) of assigning it to class m and the cost of assigning it to class $\hat{y}_{\bar{m}}(\hat{p}(x))$. We then select $\alpha_{m|n}$ to minimize the empirical risk when the validation set is classified according to (13).

6. Experiments and Results

The experiments were conducted using datasets available at the UCI Repository (Blake & Merz, 1998). For two class problems the methods described in section 5 are all equivalent to choosing a threshold on the probability estimate of one class. This approach has already been shown to be effective (Lachiche & Flach, 2003). Therefore, only datasets containing more than two classes were considered. In addition, classes with fewer than 10 samples in the dataset were not included.

Naïve Bayes was used to estimate the class probabilities. Continuous random variables were converted to discrete variables by partitioning them into bins. The middle bin was centered at the mean of the data and each bin was one standard deviation wide. Ten-fold crossvalidation was used. The complete dataset was divided into ten parts. During each step in the crossvalidation one part was set aside as test. The naïve Bayes parameters were estimated using eight of the remaining parts and the probability estimates were calculated on the ninth. Having done this for each of the nine parts the matched boundary matrix was optimized on these probability estimates. For each dataset the experiments were run 100 times with random crossvalidation partitions and the results were averaged.

Table 1 shows the results ordered by average number of samples per class (SPC). The left side of the table shows the test-set empirical probability of error (i.e. C equal to the 0/1 cost matrix). The right side shows the test-set empirical risk for a particular cost matrix. Here we include results using the method in (Lachiche & Flach, 2003) which is similar to a column multiply approach with a 0/1 initialization – this method is labeled $w\hat{p}$. The cost matrix was randomly generated with off diagonal elements between 1 and 10. For a task with K classes, the top left $K \times K$ elements were used as the cost matrix. The complete generated cost matrix was

$$C = \begin{bmatrix} 0 & 2 & 4 & 9 & 3 & 5 & 2 \\ 3 & 0 & 2 & 6 & 5 & 4 & 3 \\ 7 & 2 & 0 & 4 & 1 & 9 & 7 \\ 5 & 7 & 7 & 0 & 10 & 1 & 7 \\ 9 & 3 & 4 & 6 & 0 & 8 & 4 \\ 8 & 2 & 6 & 5 & 5 & 0 & 6 \\ 5 & 1 & 2 & 7 & 6 & 10 & 0 \end{bmatrix}.$$

The matched boundary matrix approaches were most effective for the non-0/1 cost matrix. The good performance of the baseline in 0/1 cost tasks is in agreement with observations for two class problems (Friedman, 1997). For each of the experiments the matched boundary matrix reduced the empirical risk on the training set significantly. However, for datasets with small SPC the matched boundary matrix often overfits the training set and the performance on the test set deteriorates compared with the baseline (using C in the

Table 1. Results. The bold entries show where the optimized classifier outperforms the baseline. The best performance is underlined. Entries in italics indicate a less than 95% statistically significant difference to the baseline performance.

Dataset	[SPC]	K	Prob. of Error (0/1 cost)					Empirical Risk (cost matrix C)					
			0/1	$\ $	$\times R$	$\times C$	Gen	C	$w\hat{p}$	$\ $	$\times R$	$\times C$	Gen
Bridges 2 (type)	[17.5]	6	0.41	<u>0.40</u>	<i>0.41</i>	0.40	<i>0.41</i>	<u>1.43</u>	1.60	1.46	1.49	1.46	1.50
Audiology	[29.6]	5	<u>0.14</u>	0.16	0.16	0.16	0.17	<u>0.67</u>	0.99	0.70	0.71	0.74	0.79
Image segmentation	[30.0]	7	0.17	0.17	0.18	0.17	<u>0.16</u>	0.89	0.83	0.91	<i>0.90</i>	0.86	<u>0.60</u>
Bridges 2 (rel-1)	[34.3]	3	<u>0.29</u>	0.29	0.30	0.29	0.29	<u>1.17</u>	1.23	1.22	1.23	1.22	1.25
Bridges 2 (material)	[35.3]	3	0.15	0.14	0.14	<u>0.14</u>	0.14	0.36	0.39	0.35	<i>0.36</i>	<i>0.36</i>	0.39
Flag (religion)	[35.6]	5	<u>0.51</u>	0.52	0.52	0.51	0.53	2.52	2.21	2.20	<u>2.19</u>	2.22	2.20
Horse-colic (code)	[38.4]	7	<u>0.56</u>	0.57	0.57	0.57	0.56	2.86	<u>2.64</u>	2.81	2.81	2.68	2.76
Horse-colic (site)	[39.3]	7	<u>0.47</u>	0.49	0.49	0.49	0.48	<u>2.43</u>	2.63	2.48	2.48	2.52	2.53
Glass	[41.0]	5	<u>0.34</u>	0.35	0.35	0.35	<i>0.34</i>	1.17	1.22	1.20	1.20	1.21	<u>1.13</u>
Horse-colic (type)	[60.0]	5	<u>0.46</u>	0.49	0.49	0.49	0.48	2.31	2.22	2.18	2.17	2.22	<u>2.17</u>
Dermatology	[61.0]	6	<u>0.02</u>	0.02	0.03	0.02	0.04	0.10	0.26	0.09	<u>0.09</u>	<i>0.10</i>	0.13
Ecoli	[65.4]	5	<u>0.15</u>	0.15	0.16	0.15	0.16	<u>0.48</u>	0.56	0.52	0.52	0.54	0.56
Horse-colic (subtype)	[99.7]	3	0.39	0.32	<u>0.32</u>	0.32	0.32	1.32	1.14	1.14	1.14	1.15	<u>1.14</u>
Flare2 (common)	[262.3]	4	0.22	0.16	0.16	<u>0.16</u>	0.16	0.79	0.52	0.61	<u>0.52</u>	0.52	0.52
Car	[432.0]	4	0.14	0.13	0.13	0.13	<u>0.11</u>	0.58	0.43	0.46	0.43	0.39	<u>0.36</u>
Nursery	[3239.5]	4	0.10	0.10	0.10	0.10	<u>0.08</u>	0.58	<u>0.52</u>	0.57	0.57	0.58	0.53

estimated local risk). As SPC increases, the performance of the proposed methods improve substantially over that of the baseline. For the larger datasets, using more update iterations improves the performance further, e.g. for the general update algorithm the average empirical risk on the Nursery dataset was just 0.43 after 12 iterations.

Section 4 and figure 1 show that the restrictive update methods (parallel, row multiply and column multiply) allow significant flexibility in the resulting partition. This is also evident in the results where these restrictive approaches often perform similarly to the general update method. In fact in some cases we found that the bulk of the performance improvement was achieved by updating just the first parameter and leaving all others as they were in the real cost matrix.

7. Conclusion and Future Work

The experiments show that matched boundary matrices provide a method of improving the cost sensitive performance of probability estimation based classifiers. For small datasets care must be taken not to overfit the training set. For large datasets the results demonstrate the method’s robustness to such overfitting. Future work will seek to develop more sophisticated methods for selecting matched boundary matrices aimed particularly at reducing overfitting.

In this paper we used a naïve Bayes classifier. We plan to investigate the use of matched boundary matrix methods with other classification algorithms. In par-

ticular we will consider the effect of using this approach in classifier design using MetaCost (Domingos, 1999).

References

- Blake, C., & Merz, C. (1998). UCI repository of machine learning databases.
- Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. *Fifth International Conferences on Knowledge Discovery and Data Mining* (pp. 155–164).
- Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1, 55–77.
- Lachiche, N., & Flach, P. (2003). Improving accuracy and cost of two-class and multi-class probabilistic classifiers using roc curves. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*.
- Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, 19, 78–98.
- Noe, D. (1983). Selecting a diagnostic study’s cutoff value by using its receiver operating characteristic curve. *Clinical Chemistry*, 29, 571–2.
- Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Eight International conference on knowledge discovery and data mining* (pp. 694–699). ACM Press.