Cost-sensitive Multi-class Classification from Probability Estimates

Deirdre B. O'Brien DEIRDRE@GOOGLE.COM
Google Inc., Mountain View, CA

Maya R. Gupta GUPTA@EE.WASHINGTON.EDU
University Of Washington, Seattle, WA

Robert M. Gray RMGRAY@STANFORD.EDU
Stanford University, Stanford, CA

Abstract

For two-class classification, it is common to classify by setting a threshold on class probability estimates, where the threshold is determined by ROC curve analysis. An analog for multi-class classification is learning a new class partitioning of the multiclass probability simplex to minimize empirical misclassification costs. We analyze the interplay between systematic errors in the class probability estimates and cost matrices for multiclass classification. We explore the effect on the class partitioning of five different transformations of the cost matrix. Experiments on benchmark datasets with naive Bayes and quadratic discriminant analysis show the effectiveness of learning a new partition matrix compared to previously proposed methods.

1. Introduction

Many classifiers first estimate class probabilities $\hat{p}_k(x)$ for each class $k \in \{1, \ldots, K\}$, then classify a test sample x as the class $\hat{y}(x)$ that minimizes the expected misclassification costs:

$$\hat{y}(x) = \arg \min_{i=1,\ldots,K} \sum_{j=1}^{K} c_{i|j} \hat{p}_j(x) \doteq g(\hat{p}(x); c), \qquad (1)$$

where $c_{i|j}$ is the i^{th} row, j^{th} column element of the cost matrix c, and the cost of classifying as class i when the true class is j. We define the function g to use as short-hand for this minimization.

Such probability-based classifiers can be interpreted as mapping each test sample to a point on the $\hat{p}$-simplex, where each corner of the simplex has $\hat{p}_j = 1$ for some j, and $\hat{p}_i = 0$ for all $i \neq j$; and that the cost matrix c induces a partitioning of the $\hat{p}$-simplex into regions assigned to each of the K classes. However, the probability estimation can suffer from systematic errors, e.g. oversmoothing the estimate towards class prior probabilities. The main contribution of this paper is an analytic and experimental investigation of how changing the partitioning of the $\hat{p}$-simplex can reduce the effect of such systematic errors on classification loss, analogous to ROC analysis for two-class classification.

First, we discuss systematic probability estimation errors and show how these errors can cause classification errors. Then in Section 3 we review methods to reduce the effect of such errors. In Section 4 we establish properties that describe how changing c affects the class-partitioning of the $\hat{p}$-simplex. In Section 5, we propose learning a partitioning of the $\hat{p}$-simplex that seeks to minimize the empirical misclassification costs for the given c, and we provide experimental evidence of the effectiveness of our approach in Section 6.

2. Systematic Error in Multi-class Probability Estimation

Friedman uses the term *oversmoothing* for cases where the probability estimates are systematically smoothed towards the class prior probabilities, and *undersmoothing* for cases where the class probability estimates produced are too confident, such as 1-NN (Friedman, 1997). Other systematic errors in the probability estimates can occur; Niculescu-Mizil and Caruana have documented the systematic errors introduced by various methods of probability estimation for two-class classification (Niculescu-Mizil & Caruana, 2005). Here we provide illustrative examples of over- and undersmoothing in multiclass tasks.

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

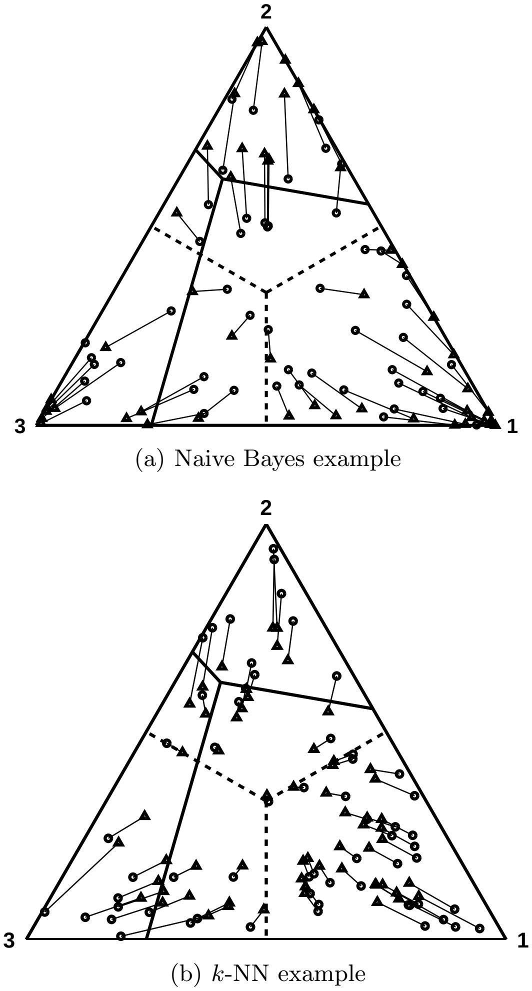
2.1. A Naive Bayes' Example

Consider a three-class problem with discrete features. We estimate class probabilities for test samples using naive Bayes (Hastie et al., 2001). The three classes are equally likely, and the feature vector consists of two identical copies of the same feature. Thus the naive Bayes' assumption of independent features is clearly violated. Figure 1(a) shows pairings of a true class probability (marked with an attached circle) and the associated naive Bayes' estimated probability (marked with a triangle). The incorrect feature-independence assumption *undersmooths* the probability estimates, pushing them towards the edges of the simplex.

When an estimated probability and the corresponding true probability fall in the same class partition, the undersmoothing does not cause any classification error. When the line attaching a triangle to a circle crosses a class partition line, a classification error occurs. One sees that undersmoothing does not cause errors given the 0/1 cost matrix (dashed lines), but causes many errors given an asymmetric cost matrix (solid lines).

2.2. A k-NN Example

Let N be the number of training samples. Then for the k-NN classifier as the number of nearest neighbors $k \to N$, the probability estimates are smoothed towards the class prior probabilities. Figure 1(b) illustrates an extreme example: $k = 2000$, $N = 3000$, and the samples are drawn iid and with equal probability from one of three class-conditional normal distributions. The oversmoothing does not cause errors given the 0/1 cost matrix (dashed lines), but causes many errors given an asymmetric cost matrix (solid lines).



(a) Naive Bayes example

(b) k-NN example

Figure 1. Circles mark the true probabilities, triangles mark the estimated probabilities, and each line connects a true probability to the corresponding estimate. The dashed lines mark the class partitioning of the $\hat{p}$-simplex induced by the 0-1 cost matrix, and the solid lines mark the class partitioning induced by an asymmetric cost matrix.

3. Related Work

Approaches to deal with the systematic errors in probability estimation can be analyzed in terms of the classification rule given in (1). Such approaches generally either change the partitioning of the $\hat{p}$-simplex, or change the probability estimates.

3.1. Related Work in Two-class Classification

For the two-class case, the $\hat{p}$-simplex is a line segment from $\hat{p}_1(x) = 0$ to $\hat{p}_1(x) = 1$, and a scalar threshold t partitions the two class regions. The optimal threshold $t^\star$ derived with respect to (1) is,

$$t^\star = \frac{c_{1|2} - c_{2|2}}{c_{1|2} + c_{2|1} - c_{1|1} - c_{2|2}}. \tag{2}$$

Classification errors can be reduced by changing the class-partitioning by specifying a threshold t that reduces the effect of systematic errors of the class probability estimates. The most common approach uses the receiver operating characteristic (ROC) curves (Egan, 1975; Hanley & McNeil, 1982). An ROC curve plots estimates of the probabilities $P_{\hat{Y}|Y}(2|2)$ versus $P_{\hat{Y}|Y}(2|1)$ for thresholds $t, 0 \le t \le 1$, where the estimates are derived from training or validation data. For a given cost matrix the desired point on the ROC curve is chosen and the associated threshold t is used for the classifier (Noe, 1983; Provost & Fawcett, 2001).

Other methods fix the threshold at the theoretical optimal $t^\star$ given by (2), and seek to improve classification by improving the probability estimates. Friedman considered adding a scalar a to the probability estimates for class 1 (Friedman, 1997); it is easy to show that this method is equivalent to using a threshold $\tilde{t} = t^\star - a$.

Zadrozny and Elkan use monotonic functions of the probability estimate $\hat{p}_1$ to give a *calibrated* estimate and show great improvements in cost-sensitive classification when the calibrated probability estimates are used in place of the original estimates (Zadrozny & Elkan, 2001; Zadrozny & Elkan, 2002). One of their approaches builds on Platt's earlier work to transform support vector machine (SVM) scores into probability estimates using a sigmoid function (Platt, 2000). The same approach can be applied to probability estimates rather than SVM scores. Zadrozny and Elkan propose two other approaches to perform the calibration: *binning* and *pair-adjacent violators*.

Binning takes the probability estimates obtained using cross-validation, orders these values and then groups them into B bins so that there are an equal number of samples in each bin (Zadrozny & Elkan, 2001). The upper and lower boundaries of each bin are determined, and for any test sample with a probability estimate falling in bin b, the updated probability estimate for class 1 is given by the fraction of validation samples in bin b that belong to class 1.

Pair-adjacent violators (PAV) monotonically transform the probability estimates using isotonic regression (Ayer et al., 1955). It has been shown that applying threshold $t^\star$ from (2) to the calibrated probability estimates obtained using PAV is equivalent to using a threshold chosen by ROC analysis on the original probabilities (O'Brien, 2006).

3.2. Related Work in Multi-class Classification

Zadrozny and Elkan extended their two-class solutions to multi-class problems by breaking the classification task into a number of binary classification tasks and using error correcting output codes (ECOC) to obtain multi-class probability estimates (Zadrozny & Elkan, 2002).

Other methods seek to extend the ROC thresholding approach to K-class classification. Instead of choosing a scalar threshold t, a partition of the $(K-1)$-dimensional $\hat{p}$-simplex must be specified. Mossman proposed a method for three class tasks using a very restrictive partitioning of the simplex (Mossman, 1999):

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

Lachiche and Flach proposed an alternative to the minimum expected misclassification cost assignment of (1) (Lachiche & Flach, 2003):

$$\hat{y}(x) = \arg\max_i w_i^* \hat{p}_i, \tag{4}$$

where the w_i^* are chosen by minimizing costs on the training set:

$$w^* = \arg\min_w \sum_{n=1}^{N} c_{(\arg\max_i w_i \hat{p}_i(x_n))|y_n}, \tag{5}$$

and x_n, y_n are the nth training sample and its associated class label. We refer to (4) and (5) as the *LF method*. Mossman's method and the LF method can both be viewed as learning a new partitioning for the $\hat{p}$-simplex. In Section 5, we show how these partitions can be achieved by using different cost matrices in equation (1).

MetaCost is a wrapper method that can be used with any classification algorithm and reduces the variance of the probability estimates by bootstrapping (Domingos, 1999). MetaCost reduces the variance of probability estimates, but is not designed to overcome systematic probability estimation errors.

4. The Effect of the Cost Matrix on the Class-Decision Boundaries

In this section we establish how different changes in the cost matrix affect the class partitioning of the $\hat{p}$-simplex enacted by (1). In Section 5 we use these properties to propose a new method to reduce the effect of systematic errors in probability estimation.

For a particular cost matrix c, and any two classes i, k that are adjacent in the partition of the $\hat{p}$-simplex, the partition-boundary between them is described by the hyperplane,

$$\sum_{j=1}^{K} c_{i|j} \hat{p}_j = \sum_{j=1}^{K} c_{k|j} \hat{p}_j. \tag{6}$$

We restrict attention to cost matrices where the cost of correct assignment is always less than the cost of incorrect assignment, that is, $c_{j|j} < c_{i|j}$, $\forall i \neq j$.

Property 1: *For any $\hat{p}$, the assigned class is the same for cost matrices c and αc for any scalar α.*

Proof: The minimization in (1) is unaffected by replacing $c_{i|j}$ by $\alpha c_{i|j}$, $\forall i, j$.

Property 2: *If the cost matrix c is full rank, then there is a point ζ^e where all two class boundaries as described by (6) intersect, and ζ^e may occur outside* *the probability simplex. We term ζ^e the "equal risk point".*

Proof: If c is full rank then the solution is $\zeta^e = c^{-1}\mathbf{1}/\|c^{-1}\mathbf{1}\|_1$ that will solve (6) for all classes. However, it can happen that $\|c^{-1}\mathbf{1}\|_1 = 0$, in which case the equal risk point can be said to be at infinity.

If c is not full rank there may still be an unique equal risk point. In general, ζ^e must solve $\left[\begin{array}{c|c} c & \mathbf{-1} \\ \hline \mathbf{1}^T & 0 \end{array}\right]\left[\begin{array}{c} \zeta^e \\ \gamma \end{array}\right] = \left[\begin{array}{c} \mathbf{0} \\ 1 \end{array}\right]$ for some $\gamma \in \mathbb{R}$. If the above matrix is full rank then there is a unique point solution for ζ^e, otherwise the above system is underdetermined and there is a hyperplane of solutions, or the above system can be overdetermined and there is no solution.

Property 3: *Adding a constant to all costs for a particular true class (equivalently, adding a constant to each term in any column of the cost matrix) does not affect the assignment.*

Proof: The minimization in (1) is unaffected by changing $c_{i|j}$ to $c_{i|j} + \alpha_j$, $\forall i$.

Property 4: *The class-partitioning produced by any cost matrix c can equivalently be produced by some cost matrix $\tilde{c}$ where $\tilde{c}_{i|i} = 0$ and $\tilde{c}_{i|j} > 0$ for $i \neq j$.*

Proof: Follows directly from Property 3.

Property 5: (See Fig. 2b) *Adding a constant α to the cost of assignment to class i irrespective of the true class (that is, adding α to each term in row i of a cost matrix), produces a new class-partitioning with partition boundaries parallel to those of the original.*

Proof: This change only affects the two-class boundary equations specified by (6) for class i and each class k:

$$\sum_{j=1}^{K}(c_{i|j} + \alpha)\hat{p}_j = \sum_{j=1}^{K} c_{k|j}\hat{p}_j = \sum_{j=1}^{K} c_{i|j}\hat{p}_j + \alpha.$$

Thus the new boundary between class i and k is parallel to the original boundary between class i and k. To maintain $c_{j|j} < c_{i|j}$, $\forall i \neq j$ requires $\max_{k\neq i}(c_{k|k} - c_{i|k}) < \alpha < \min_{k\neq i}(c_{k|i} - c_{i|i})$.

Property 6: (See Fig. 2c) *Scaling all costs where the true class is ℓ by a positive constant α (that is, multiplying column ℓ of c by α) moves an equal risk point along the line joining it to the corner of the simplex where $\hat{p}_\ell = 1$. The intersections of the class boundaries with the $\hat{p}_\ell = 0$ plane are unchanged.*

Proof: To prove that the new equal risk point $\tilde{\zeta}^e$ is on the line joining the original equal risk point ζ^e and the corner of the simplex where $\hat{p}_\ell = 1$, we show that there exists a constant β such that

$$\tilde{\zeta}^e_j = \beta\zeta^e_j + (1-\beta)\mathcal{I}_{(j=\ell)}, \tag{7}$$

for all j, and where $\mathcal{I}$ is the indicator function.

From (6), multiplying column ℓ by α, and requiring equal risk for classes ℓ and k:

$$\sum_{j=1,j\neq\ell}^{K}(c_{i|j} - c_{k|j})\tilde{\zeta}^e_j + (c_{i|\ell} - c_{k|\ell})(\alpha\tilde{\zeta}^e_\ell) = 0. \tag{8}$$

First note that if $\zeta^e_\ell = 0$, then ζ^e remains an equal risk point for the transformed cost matrix. Otherwise, for any ζ^e, we write $\zeta^e_j = s_j\zeta^e_\ell$. Comparing (6) and (8), there exists $\tilde{\zeta}$ such that $\tilde{\zeta}^e_j = \alpha(s_j\tilde{\zeta}^e_\ell)$, $\forall j \neq \ell$. Thus,

$$\tilde{\zeta}^e_j = \left(\frac{\alpha\tilde{\zeta}^e_\ell}{\zeta^e_\ell}\right)\zeta^e_j. \tag{9}$$

Let $\beta = \alpha\tilde{\zeta}^e_\ell/\zeta^e_\ell$, then (9) establishes (7) $\forall j \neq \ell$. Also,

$$\begin{aligned} \sum_{j\neq\ell}\tilde{\zeta}^e_j &= \beta\sum_{j\neq\ell}\zeta^e_j && (10)\\ \Rightarrow 1 - \tilde{\zeta}^e_\ell &= \beta(1 - \zeta^e_\ell), && (11) \end{aligned}$$

where (11) follows from (10) since components of $\tilde{\zeta}^e$ and ζ^e both sum to 1. This establishes (7) for ℓ.

Lastly, by setting $\hat{p}_\ell = 0$ in equation (6), it is evident that changes in $c_{i|\ell}$ and $c_{j|\ell}$ will not effect the intersections of the class boundaries with the $\hat{p}_\ell = 0$ plane.

Property 7: (See Fig. 2d) *Scaling all costs where the assigned class is i by a positive constant α (that is, multiplying all elements in row i by α) moves an equal risk point ζ^e along the hyper-plane where all classes but class i have equal expected misclassification costs.*

Proof: Let $\tilde{c}_{j|k} = c_{j|k}$ for all $j \neq i$, and let $\tilde{c}_{i|k} = \alpha c_{i|k}$. Then the equal risk point $\tilde{\zeta}^e$ produced by $\tilde{c}$ solves the same set of class boundary equations (6) specifying ζ^e, except

$$\sum_{j=1}^{K}(c_{i|j} - c_{K|j})\zeta^e_j = 0 \quad \Rightarrow \quad \sum_{j=1}^{K}(\alpha c_{i|j} - c_{K|j})\tilde{\zeta}^e_j = 0.$$

Because the constraints specifying that the other classes have equal misclassification costs still apply, the new equal risk point $\tilde{\zeta}^e$ must occur along the hyperplane specified by that subset of the constraints.

5. Learning the Partition Matrix

We propose changing the class partitions so that the class-partitioning corrects for the systematic error in the probability estimates. Changing the partition of the multi-class $\hat{p}$-simplex is analogous to the two-class practice of changing the threshold on $\hat{p}_1(x)$.

We split the cost matrix's role into two separate entities: a partition matrix (which partitions the $\hat{p}$-simplex with linear boundaries), and the misclassification cost matrix that specifies how misclassification errors are to be scored. From now on we will use the term *partition matrix* for the former role, and the term *cost matrix* only for the latter role.

Given a training set $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, where x_n has class probability estimate vector $\hat{p}(x_n)$, we propose to use a partition matrix $a^\star$ that solves

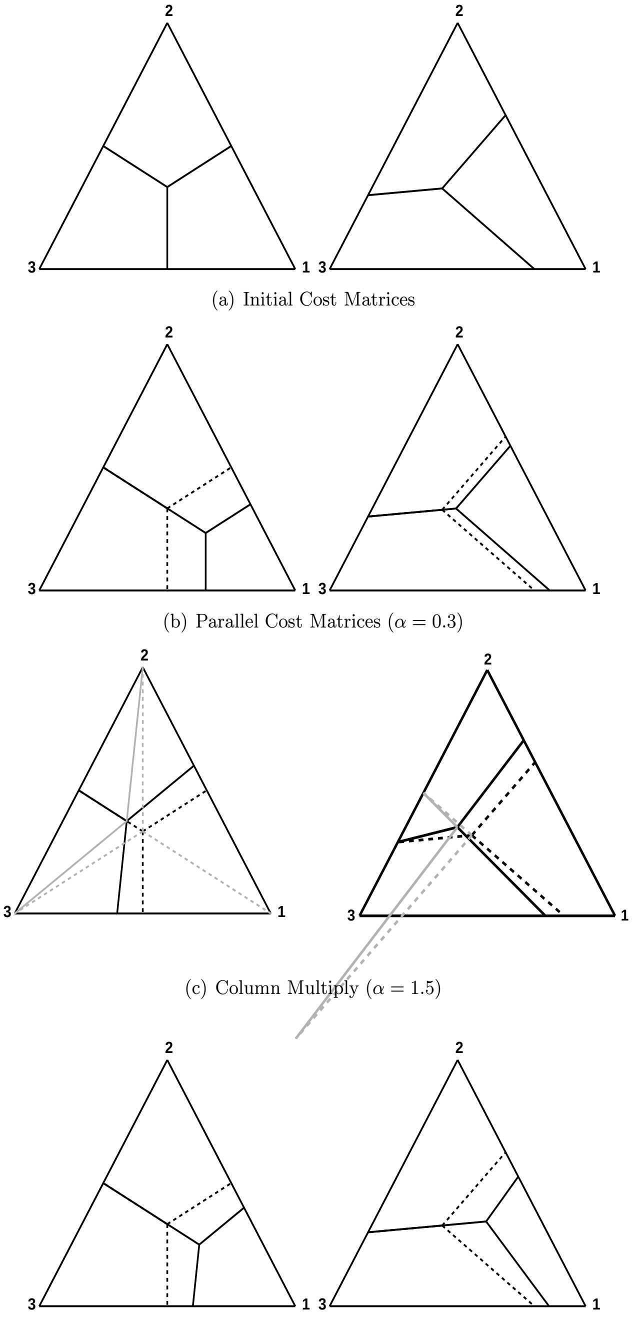
$$a^\star = \arg\min_a \sum_n c_{g(\hat{p}(x_n);a)|y_n}, \qquad (12)$$

where the function g in (12) is defined in (1). To avoid the issue of overfitting in learning a partition matrix, we restrict the partition to have linear boundaries that are parallel to the original decision boundaries produced by the cost matrix (see Fig. 2b). We have also considered learning partition matrices with different constraints, including partition matrices with all $K^2 - K$ free parameters, partition matrices that are column-multiply modifications of the original cost matrix (see Fig. 2c), and row-multiply modifications (see Fig. 2d). We found these different constraints resulted in similar performance, with the parallel partitioning working consistently well (O'Brien, 2006).

By Property 4 stated in Section 4, without loss of generality we consider only partition matrices a where $a_{i|i} = 0$ and $a_{i|j} > 0$ for $i \neq j$. From Property 5, adding α_i to row i of the cost matrix will yield a partition matrix with partition boundaries parallel to the partition boundaries induced by c. To maintain the requirement that $a_{i|i} = 0$ we subtract the same constant from column i without affecting the partition boundaries (Property 3). Thus given the cost matrix c, the partition matrix is a where $a_{i|j} = c_{i|j} + \alpha_i - \alpha_j$. This approach requires learning the parameters α_i for $i = 1, \ldots, K$.

Related methods can also be viewed as applications of (12) but with different restrictions on a. Mossman's method for three classes (Mossman, 1999) implicitly requires a to be of the form

$$\begin{bmatrix} 0 & 1 & L - \frac{\delta_2}{1-\delta_2} \\ \frac{1+\delta_2}{1-\delta_2} & 0 & L \\ \frac{\delta_1}{1-\delta_1}L & \frac{\delta_1}{1-\delta_1}L & 0 \end{bmatrix}, \qquad (13)$$



(a) Initial Cost Matrices

(b) Parallel Cost Matrices ($\alpha = 0.3$)

(c) Column Multiply ($\alpha = 1.5$)

(d) Row Multiply ($\alpha = 1.75$)

Figure 2. This figure illustrates Properties 5, 6, and 7 described in Section 4 for a three-class classification. The partition produced by the cost matrix is marked by solid lines: a 0/1 cost matrix for the left figures, and an asymmetric cost matrix for the right figures. The corners of the simplex are marked such that $\hat{p}_j = 1$ at corner j, and if a test sample has estimated class probability $\hat{p}$ that falls in the region including corner j, then the estimated class label is j. For each of the figures, manipulations are applied to the class 1 elements of the cost matrix. The dashed lines show the initial cost matrix partitions, the gray lines help to illustrate the properties.

where the δ terms are those used in (3) and $L \gg 1$. The LF method (Lachiche & Flach, 2003) is equivalent to choosing a partition matrix a such that $a_{i|j} = w_j z_{i|j}$ where $z_{i|j}$ is the 0-1 cost matrix and w_j is the weight used in equations (4) and (5). Thus the LF method is equal to a column-multiply of a 0-1 cost matrix (see Property 5 stated in Section 4, and Fig 2c), whereas our proposed method enacts a parallel shift based on the actual cost matrix c.

5.1. Optimizing the Partition Matrix

We learn the new partition matrix by minimizing an empirical loss calculated on a validation set of labeled samples using a greedy approach. The new partition matrix a is initialized to the cost matrix c. Then each free parameter is updated in turn. Let a denote the current partition matrix, and the new partition matrix $\tilde{a}$ will have $\tilde{a}_{i|j} = a_{i|j} + \alpha_i$ and $\tilde{a}_{j|i} = a_{j|i} - \alpha_i$ for all $j \neq i$. Suppose $\alpha_i = -\infty$ and interpret $\tilde{a}$ as a cost matrix – then there would be an infinitely negative cost to assigning a sample as class i, and thus every training sample would be assigned to class i. Suppose one increased α_i from negative infinity. For different values of α_i it would become more cost-effective to classify each of the training samples as a different class, call this classification choice $g^i(\hat{p}(x_n))$, where $g^i(\hat{p}(x_n)) = \arg\min_{k \neq i} \sum_{j=1}^{K} a_{k|j} \hat{p}_j(x)$. Let α_{in} denote the changepoint value – for $\alpha_i < \alpha_{in}$ training sample n would be assigned to class i and for $\alpha_i > \alpha_{in}$ training sample n would be assigned to class $g^i(\hat{p}(x_n))$.

We find these N changepoints α_{in} for $n = 1, \ldots, N$, by solving the N equations,

$$\sum_{j=1}^{K} \left(a_{g^i(\hat{p}(x_n))|j}\right) \hat{p}_j(x_n) = \sum_{j=1}^{K} \left(a_{i|j} + \alpha_{in}\right) \hat{p}_j(x_n).$$

Re-order the training data by their changepoints, so that $\{x_k, y_k\}$ denotes the training point with the kth largest changepoint. Then select N^* where,

$$N^* = \arg \min_{N_0=1,2,\ldots,N} \sum_{n<N_0} c_{g^i(\hat{p}(x_n))|y_n} + \sum_{n \geq N_0} c_{i|y_n}. \tag{14}$$

Note that α_{iN^*} to α_{iN*+1} defines the range of α_i that would yield the empirical cost given in (14); we set the parameter α_i to be the geometric mean of α_{iN^*} and α_{iN^*+1}. Since we require that $\tilde{a}_{j|j} < \tilde{a}_{k|j}$, for all $k \neq j$, it must be that $a_{j|j} < a_{i|j} + \alpha_i$ and $a_{i|i} + \alpha_i < a_{k|i}$, for all $j, k \neq i$, and so α_i is clipped to satisfy these conditions. In addition, if $\alpha_{iN^*} < 0 < \alpha_{iN^*+1}$, then α_i is set equal to 0, or equivalently $\tilde{a}$ is set equal to a.

Each class's partition matrix parameter is adjusted in this manner once, and the parameters are updated in order of class size from most populous to least. Preliminary experiments provided evidence that multiple passes through the parameters did not improve the final classification performance, and that performance was fairly robust to the parameter ordering.

6. Experiments

Experiments with UCI benchmark datasets compare the proposed parallel-partition matrix method to MetaCost (MC) (Domingos, 1999) and to the LF method (Lachiche & Flach, 2003).

For two-class problems the proposed partition matrix methods are equivalent to ROC analysis and therefore only multi-class problems are considered here.

There are two basic variants of MetaCost: the first variant is that the probability estimates are based on training samples not including the sample, while the other variant is that all-inclusive estimates are made. The results reported here are the better of the two variants for each dataset.

Randomized ten-fold cross-validation was done 100 times for each method and each dataset. In the cross-validation, 1/10 of the data was set aside as test data. For MetaCost, 100 resamples were generated using the remaining nine folds. For the proposed methods and the LF method the remaining nine folds were subject to a nine-fold cross-validation so that 8/10 of the data (eight folds) were used to estimate the probabilities for each of the nine folds. Then the partition matrix a and LF parameters were estimated using the nine folds' probability estimates. Finally, the learned cost-sensitive classifier was applied to the withheld 1/10 of the test data.

Experiments were done with two different probability estimation methods. For datasets with discrete features, multinomial naive Bayes was used with Laplacian correction for estimating probabilities. Any continuous features used with naive Bayes were quantized to 21 values. For datasets with continuous features, regularized quadratic discriminant analysis (QDA) (Friedman, 1989) was used. Each class's estimated covariance matrix was regularized as,

$$\hat{\Sigma} = (1 - \gamma - \lambda)\hat{\Sigma}_{ML} + \lambda\bar{\Sigma} + \lambda \frac{\text{trace}\left(\hat{\Sigma}_{ML}\right)}{d} I,$$

where $\hat{\Sigma}_{ML}$ is the maximum likelihood estimate of the full covariance matrix, $\bar{\Sigma}$ is the pooled maximum likelihood estimate, d is the dimensionality of the feature vector, and γ and λ were increased from zero until the condition number of $\hat{\Sigma}$ was less than 10^6.

Experiments were run with two different cost scenarios. In practical situations it is often the rare events that are of greatest interest, and therefore the cost of misclassifying samples from rare classes is higher. To simulate this situation, we set $c_{i|i} = 0$ and,

$$c_{i|j} = \frac{N_i}{N_i + N_j},$$

where N_i is the number of training samples labeled class i. For the second set of experiments, we set $c_{i|i} = 0$, and each element $c_{j|k}$ for $j \neq k$ was drawn randomly and uniformly from $[1, 10]$.

6.1. Discussion of Results

Results are presented in Tables 1 and 2 in terms of the mean increase in performance over the baseline of using the partitioning induced by the cost matrix. The datasets in the results tables are ordered by increasing geometric-mean class size for the training set. The results show that MetaCost did not consistently improve over the baseline. The LF method performed better and usually improved performance, but caused large increases in error in two cases: image segmentation with QDA, and dermatology with naive Bayes. In contrast, learning a new parallel partitioning showed a mean improvement of 10.8% for the rarity-based cost matrix and a mean improvement of 8.9% with the random cost matrix.

The LF method and proposed partition-learning method are designed to correct for systematic errors in the probability estimates. Such systematic errors can be interpreted as a bias that can cause the classifier to be wrong in the same way on average over many training sets. Thus, we expected to see a greater increase in performance for the LF method and proposed partition-learning method for larger datasets (further down in the tables) because performance given a large training sets is more likely to suffer from problems of bias than estimation variance. With smaller datasets, estimation variance is generally a larger concern, and the bias reduction offered by the LF and proposed method may not be very helpful. In addition, for small datasets it is harder to learn the systematic error from only a few training samples, and there is an increased risk of overfitting the learned parameters.

The datasets *iris* and *dermatology* had very low misclassification loss for the original probability estimates. For these datasets there was not much improvement possible, and we hypothesize that the methods that learned parameters were likely to overfit to small improvements in the training data.

		Mean % Improved		
Alg.	**Dataset**	**MC**	**LF**	**Par**
NB	Bridges 2 (type)	-1	-10	**5**
NB	Bridges 2 (material)	**16**	0	-8
NB	Audiology	-34	**4**	-3
NB	Horse (site)	-7	17	**18**
NB	Bridges 2 (rel-l)	-6	**-3**	-5
NB	Image segmentation	-53	**5**	0
QDA	Image segmentation	-5	-122	**9**
NB	Horse (code)	1	13	**16**
NB	Glass	-4	**18**	16
QDA	Glass	-3	27	**33**
NB	Flag (religion)	2	**7**	6
NB	Horse (type)	1	25	**27**
QDA	Iris	**-1**	-18	-4
NB	Ecoli	-13	**-5**	-7
QDA	Ecoli	-6	-9	**-1**
NB	Dermatology	**-12**	-100	-18
QDA	Wine	1	**64**	56
NB	Horse (subtype)	2	**22**	21
NB	Flare2 (common)	2	**18**	17
NB	Car	-16	10	**18**
NB	Nursery	-16	-8	**31**
	Mean	-7.2	-2.1	**10.8**
	Std. Dev.	14.4	40.4	17.3

Table 1. Performance for the rarity-based cost matrix with $c_{i|i} = 0$ and $c_{i|j} = N_i/(N_i + N_j)$. Largest average improvement for each dataset is in bold.

We used a greedy optimization approach to learn the new partition matrix for our method, but in some cases this leads only to a locally optimal solution. The LF method also uses a greedy search. However, Deng et al.'s results (Deng et al., 2006) show that improving the optimization of the LF objective can lead to an improvement in results. Similarly, we hypothesize that finding a globally optimal solution would also lead to an improvement in costs.

7. Discussion

We analyzed how changes in the cost matrix affect the partitioning of the $\hat{p}$-simplex due to the cost matrix. Based on this analysis, we explored correcting for systematic probability estimation errors by learning a partitioning of the $\hat{p}$-simplex that minimizes empirical misclassification costs on the training set. To reduce overfitting, we only considered partitionings parallel to the original partitioning induced by the cost matrix. Experiments with two standard classifiers showed that this post-processing worked best when the number of training samples per class is relatively large, and when the estimation error with the original cost matrix is large.

		Mean % Improved		
Alg.	**Dataset**	**MC**	**LF**	**Par**
NB	Bridges 2 (type)	-5	-12	**-3**
NB	Bridges 2 (material)	**12**	-10	2
NB	Audiology	-63	-46	**-9**
NB	Horse(site)	**2**	1	-1
NB	Bridges 2 (rel-l)	**-2**	-5	-5
NB	Image segmentation	-36	**8**	-3
QDA	Image segmentation	-11	-123	**35**
NB	Horse (code)	-2	**5**	1
NB	Glass	**5**	-4	1
QDA	Glass	-4	**7**	1
NB	Flag (religion)	-2	**13**	**13**
NB	Horse(type)	-2	**5**	1
QDA	Iris	**-1**	-7	-3
NB	Ecoli	-11	-17	**-10**
QDA	Ecoli	-3	-2	**3**
NB	Dermatology	-42	-132	**15**
QDA	Wine	-5	68	**73**
NB	Horse(subtype)	-19	**18**	**18**
NB	Flare2 (common)	-5	**35**	22
NB	Car	-22	25	**32**
NB	Nursery	-55	**10**	3
	Mean	-12.9	-7.8	**8.9**
	Std. Dev.	19.9	45.4	19.2

Table 2. Performance for random cost matrix. Largest average improvement for each dataset is in bold.

8. Acknowledgements

The authors thank Jerome Friedman and Richard Olshen for helpful discussions during the course of this research. This research was supported in part by the United States National Science Foundation Grant CCR-0073050.

References

Ayer, M., Brunk, H., Ewing, G., Reid, W., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, *4*, 641–647.

Deng, K., Bourke, C., Scott, S., & Vinodchandran, N. (2006). New algorithms for optimizing multi-class classifiers with ROC surfaces. *Proc. of the ICML 2006 Workshop on ROC Analysis in Machine Learning.* Pittsburgh, USA.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. *Proc. of 5th International Conference on Knowledge Discovery and Data Mining* (pp. 155–164). San Diego, CA.

Egan, J. (1975). *Signal detection theory and ROC-analysis.* New York: Academic Press.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*, 165–175.

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, *1*, 55–77.

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning.* New York: Springer-Verlag.

Lachiche, N., & Flach, P. (2003). Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. *Proc. of 20th International Conference on Machine Learning* (pp. 416–423). Washington DC.

Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, *19*, 78–98.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proc. of 22nd International Conference on Machine Learning.*

Noe, D. (1983). Selecting a diagnostic study's cutoff value by using its receiver operating characteristic curve. *Clinical Chemistry*, *29*, 571–2.

O'Brien, D. B. (2006). *Cost-sensitive performance of probability-estimation based classifiers: analysis and practice.* Doctoral dissertation, Stanford University.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers* (pp. 61–74).

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42*, 203 – 231.

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naïve Bayesian classifiers. *Proc. of 18th International Conference on Machine Learning* (pp. 609–616). Morgan Kaufmann Publishers, Inc.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proc. of 8th International Conference on Knowledge Discovery and Data Mining* (pp. 694–699). ACM Press.