The Error Coding Method and PICTs

Gareth James[*]
and
Trevor Hastie
Department of Statistics, Stanford University

March 29, 1998

Abstract

A new family of *plug-in* classification techniques has recently been developed in the statistics and machine learning literature. A plug-in classification technique (PICT) is a method that takes a standard classifier (such as LDA or TREES) and plugs it into an algorithm to produce a new classifier. The standard classifier is known as the *Base Classifier.* These methods often produce large improvements over using a single classifier. In this paper we investigate one of these methods and give some motivation for its success.

1 Introduction

A new family of classifiers has recently been developed in the statistics and machine learning communities. They involve taking a standard classifier (such as LDA or TREES) and *plugging* it into an algorithm to produce a new classifier. We refer to these *Plug in Classification Techniques* as PICTs and the plugged in classifier as the *Base Classifier.*

Several members of this family have recently received a great deal of discussion in the literature. Some examples include the *Bagging* algorithm (Breiman (1996b)), the *AdaBoost* algorithm (Freund and Schapire (1995)) and the *ECOC* algorithm (Dietterich and Bakiri (1995)). In this paper we investigate the latter procedure and give motivation for its success.

Dietterich and Bakiri (1995) suggested the following PICT, motivated by Error Correcting Coding Theory, for solving k class classification problems using binary classifiers.

- Produce a k by B (B large) binary coding matrix, i.e. a matrix of zeros and ones. We will denote this matrix by Z, its i,jth component by Z_{ij}, its ith row by $\mathbf{Z}_i$ and its jth column by $\mathbf{Z}^j$. The following is a possible coding matrix for a 10 class problem.

Class	$\mathbf{Z}^1$	$\mathbf{Z}^2$	$\mathbf{Z}^3$	$\mathbf{Z}^4$	$\mathbf{Z}^5$	$\mathbf{Z}^6$	$\dots$	$\mathbf{Z}^{15}$
0	1	1	0	0	0	0	$\dots$	1
1	0	0	1	1	1	1	$\dots$	0
2	1	0	0	1	0	0	$\dots$	1
3	0	0	1	1	0	1	$\dots$	1
4	1	1	1	0	1	0	$\dots$	0
5	0	1	0	0	1	1	$\dots$	0
6	1	0	1	1	1	0	$\dots$	1
7	0	0	0	1	1	1	$\dots$	0
8	1	1	0	1	0	1	$\dots$	1
9	0	1	1	1	0	0	$\dots$	0

[*] The authors were partially supported by grant DMS-9504495 from the National Science Foundation, and grant ROI-CA-72028-01 from the National Institutes of Health.

- Use the first column of the coding matrix ($\mathbf{Z}^1$) to create two *super* groups by assigning all classes with a one in the corresponding element of $\mathbf{Z}^1$ to super group one and all other classes to super group zero. So for example, with the above coding matrix, we would assign classes 0, 2, 4, 6 and 8 to super group one and the others to super group zero.
- Train your Base Classifier on the new two class problem.
- Repeat the process for each of the B columns ($\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^B$) to produce B trained classifiers.
- To a new test point apply each of the B classifiers. Each classifier will produce $\hat{p}_j$ which is the estimated probability the test point comes from the jth super group one. This will produce a vector of probability estimates, $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_B)^T$.
- To classify the point calculate $L_i = \sum_{j=1}^{B} |\hat{p}_j - Z_{ij}|$ for each of the k classes (i.e. for i from 1 to k). This is the L1 distance between $\hat{\mathbf{p}}$ and $\mathbf{Z}_i$ (the ith row of Z). Classify to the class with lowest L1 distance or equivalently $\arg\min_i L_i$.

We call this the ECOC PICT. Each row in the coding matrix corresponds to a unique (non-minimal) coding for the appropriate class. Dietterich's motivation was that this allowed *errors* in individual classifiers to be *corrected* so if a small number of classifiers gave a bad fit they did not unduly influence the final classification. Several Base Classifiers have been tested. The best results were obtained by using *trees*, so all the experiments in this paper use a standard CART classifier. Note however, that the theorems are general to any Base Classifier.

In the past it has been assumed that the improvements shown by this method were attributable to the error coding structure and much effort has been devoted to choosing an *optimal* coding matrix. In this paper we develop results which suggest that a randomized coding matrix should match (or exceed) the performance of a *designed* matrix.

2 The Coding Matrix

Empirical results (see Dietterich and Bakiri (1995)) suggest that the ECOC PICT can produce large improvements over a standard k class tree classifier. However, they do not shed any light on why this should be the case. The coding matrix, Z, is central to the PICT. In the past, the usual approach has been to choose Z so that the separation between rows ($\mathbf{Z}_i$) is as large as possible (in terms of Hamming distance) on the basis that this allows the largest number of *errors* to be corrected. In the next two sections we will examine tradeoffs between a *designed (deterministic)* and a *completely randomized* matrix.

Some of the results will make use of the following assumption:

$$E_{\mathcal{T}}[\hat{p}_j \mid Z, X] = \sum_{i=1}^{k} Z_{ij} q_i = \mathbf{Z}^{j\,T} \boldsymbol{q} \quad j = 1, \ldots, B \tag{1}$$

where $q_i = P(G = i \mid X)$ is the posterior probability that the test observation is from class i given that our predictor variable is X. Note that $\mathcal{T}$ is the training set so the expectation is taken over all possible training sets of a fixed size with the coding matrix held constant. This is an unbiasedness assumption. It states that on average our classifier will estimate the probability of being in super group one correctly. The assumption is probably not too bad given that trees are considered to have low bias.

2.1 Deterministic Coding Matrix

Let $\bar{D}_i = 1 - 2L_i/B$ for $i = 1, \ldots, k$. Notice that $\arg\min_i L_i = \arg\max_i \bar{D}_i$ so using $\bar{D}_i$ to classify is identical to the ECOC PICT. Theorem 3 in Section 2.2 explains why this is an intuitive transformation to use.

Obviously it is not possible for a PICT to outperform the Bayes Classifier. However, we would hope that, when we use the Bayes Classifier as our Base Classifier for each 2 class problem the PICT would achieve the Bayes Error Rate. We call this property Bayes Optimality.

Definition 1 *A PICT is said to be Bayes Optimal if, for any test set, it always classifies to the Bayes Class when the Bayes Classifier is our Base Classifier.*

Bayes Optimality implies a type of consistency. Under continuity assumptions, it implies that, if our Base Classifier converges to the Bayes Classifier, as for example, the training sample size increases, then so will the PICT.

For the ECOC PICT to be Bayes Optimal we need $\arg\max_i q_i = \arg\max_i \bar{D}_i$, when we use the Bayes Classifier as our Base Classifier. However, it can be shown that, if the Bayes Classifier is our Base Classifier, then

$$\bar{D}_i = 1 - \frac{2}{B}\sum_{l \neq i} q_l \sum_{j=1}^{B}(Z_{lj} - Z_{ij})^2 \quad i = 1, \dots, k$$

It is not clear from this expression why there should be any guarantee that $\arg\max_i \bar{D}_i = \arg\max_i q_i$. In fact Theorem 1 tells us that only in very restricted circumstances will the ECOC PICT be Bayes Optimal.

Theorem 1 *The Error Coding method is Bayes Optimal iff the Hamming distance between every pair of rows of the coding matrix is equal.*

The Hamming distance between two binary vectors is the number of points where they differ. For general B and k there is no known way to generate a matrix with this property so the ECOC PICT will not be Bayes Optimal.

2.2 Random Coding Matrix

We have seen in the previous section that there are potential problems with using a deterministic matrix. Now suppose we randomly generate a coding matrix by choosing a zero or one with equal probability for every coordinate. Let

$$\mu_i = E_Z(1 - 2|\hat{p}_1 - Z_{i1}| \mid \mathcal{T}) = E_Z(\bar{D}_i \mid \mathcal{T})$$

where the expectation is taken over all possible matrices of a fixed size for a fixed training set. Then μ_i is the conditional expectation of $\bar{D}_i$ and we can prove the following result.

Theorem 2 *For a random coding matrix, conditional on $\mathcal{T}$, $\arg\max_i \bar{D}_i \to \arg\max_i \mu_i$ a.s. as $B \to \infty$. Or in other words, the classification from the ECOC PICT approaches the classification from just using $\arg\max_i \mu_i$ a.s.*

The theorem is a consequence of the strong law. This leads to Corollary 1 which indicates we have eliminated the main concern with using a deterministic matrix.

Corollary 1 *When the coding matrix is randomly chosen the ECOC PICT is asymptotically Bayes Optimal i.e. $\arg\max_i \bar{D}_i \to \arg\max_i q_i$ a.s. as $B \to \infty$, provided the Bayes Classifier is used as the Base Classifier.*

Theorem 2 along with the following result provide motivation for the ECOC procedure.

Theorem 3 *Under Assumption 1 for a randomly generated coding matrix*

$$E_{\mathcal{T},Z}\bar{D}_i = E_{\mathcal{T}}\mu_i = q_i \quad i = 1, \dots, k$$

This tells us that $\bar{D}_i$ is an unbiased estimate of the conditional probability so classifying to the maximum is in a sense an unbiased estimate of the Bayes classification. Note that the expectation of $\bar{D}_i$ is taken over both Z and $\mathcal{T}$ while the expectation for μ_i is only over $\mathcal{T}$.

Now Theorem 2 tells us that for *large B* the ECOC PICT will be similar to classifying using $\arg\max_i \mu_i$. However what we mean by large depends on the rate of convergence. Theorem 4 tells us that this rate is in fact exponential.

Theorem 4 *If we randomly choose Z then, conditional on $\mathcal{T}$, for any fixed X*

$$Pr_Z(\arg\max_i \bar{D}_i \neq \arg\max_i \mu_i | \mathcal{T}) \leq (k-1)e^{-mB}$$

for some positive constant m.

Note that Theorem 4 does not depend on Assumption 1. This tells us that the error rate for the ECOC PICT is equal to the error rate using $\arg\max_i \mu_i$ plus a term which decreases exponentially in the limit. This result can be proved using Hoeffding's inequality (Hoeffding (1963)).

Of course Theorem 4 only gives an upper bound on the error rate and does not necessarily indicate the behavior for smaller values of B. Under certain conditions a Taylor expansion indicates that $Pr(\arg\max_i \bar{D}_i \neq \arg\max_i \mu_i) \approx 0.5 - m\sqrt{B}$ for small values of $m\sqrt{B}$. So we might expect that for smaller values of B the error rate decreases as some power of B but that as B increases the change looks more and more exponential.

To test this hypothesis we calculated the error rates for 6 different values of B $(15, 26, 40, 70, 100, 200)$ on the LETTER data set (available from the Irvine Repository of machine learning). For each value of B we generated 5 random matrices and 5 corresponding error rates. Figure 1 illustrates the results. Each point is the average over 20 random training sets. Here we have two curves. The lower curve is the best fit of $1/\sqrt{B}$ to the first four groups. It fits those groups well but under-predicts errors for the last two groups. The upper curve is the best fit of $1/B$ to the last four groups. It fits those groups well but over-predicts errors for the first two groups. This supports our hypothesis that the error rate is moving through the powers of B towards an exponential fit.

We can see from the figure that even for relatively low values of B the reduction in error rate has slowed substantially. This indicates that almost all the remaining misclassifications are a result of the error rate of $\arg\max_i \mu_i$ which we can not reduce by changing the coding matrix.

The coding matrix can be viewed as a method for sampling from the distribution of $1 - 2|\hat{p}_j - Z_{ij}|$. If we sample randomly we will estimate μ_i (its mean). It is well known that the optimal way to estimate such a parameter is by random sampling so it is not possible to improve on this by *designing* the coding matrix. Of course it may be possible to improve on $\arg\max_i \mu_i$ by using the training data to influence the sampling procedure and hence estimating a different quantity. However, a designed coding matrix does not use the training data.

3 Why does the ECOC PICT work?

The easiest way to motivate why the ECOC PICT works, in the case of tree classifiers, is to consider a very similar method which we call the Substitution PICT. We will show that under certain conditions the ECOC PICT is very similar to the Substitution PICT and then motivate the success of the latter.

3.1 The Substitution PICT

The Substitution PICT works in the following way :

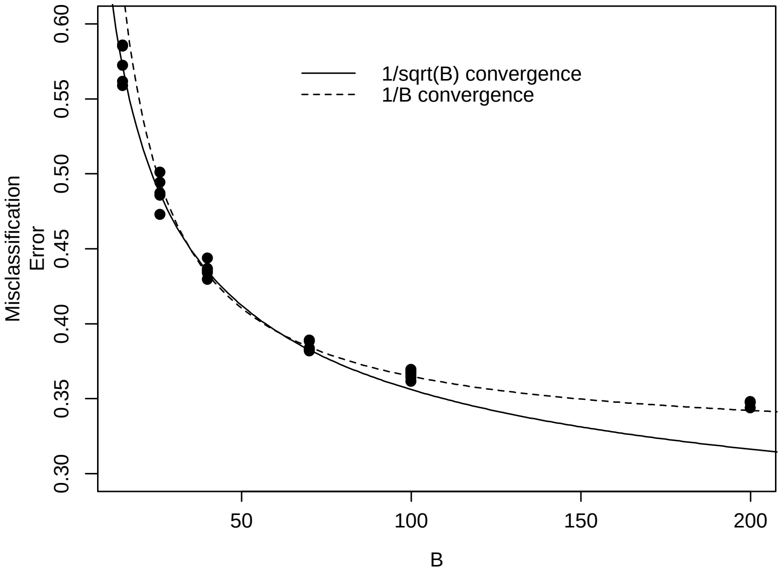

Figure 1: Best fit curves for rates $1/\sqrt{B}$ and $1/B$

Substitution Algorithm

- Produce a random binary coding matrix as with the ECOC PICT.
- Use the first column of the coding matrix ($\mathbf{Z}^1$) to create two *super* groups by assigning all classes with a one in the corresponding element of $\mathbf{Z}^1$ to super group one and all other classes to super group zero.
- Train your tree classifier on the new two class problem and repeat the process for each of the B columns. Each tree will form a partitioning of the predictor space.
- Now retain the partitioning of the predictor space that each tree has produced. Feed back into the trees the original k class training data. Use the training data to form probability estimates, just as one would do for any tree classifier. The only difference here is the rule that has been used to create the partitioning.
- To a new test point apply each of the B classifiers. The jth classifier will produce a k class probability estimate, p_{ij}, which is the estimated probability the test point comes from the ith class.
- To classify the point calculate

$$p_i^S = \frac{1}{B}\sum_{j=1}^{B} p_{ij} \tag{2}$$

and classify to $\arg\max_i p_i^S$

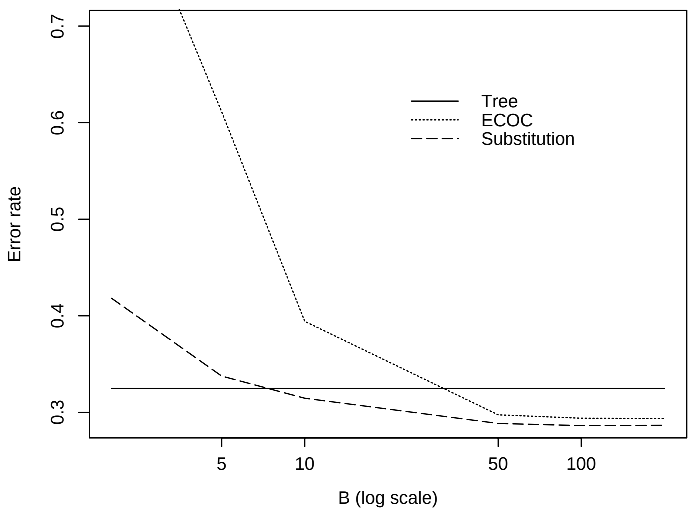

Figure 2: Error rates on the simulated data set for the tree method, Substitution PICT and ECOC PICT plotted against B (on log scale)

In summary, the Substitution PICT uses the coding matrix to form many different partitionings of the predictor space. Then, for each partitioning, it forms k class probability estimates by examining the proportions of each class, among the training data, that fall in the same region as the test point. The probability estimates are then combined by averaging over all the trees for each class. The final classification is to the maximum probability estimate.

Theorem 5 shows that under certain conditions the ECOC PICT can be thought of as an approximation to the Substitution PICT.

Theorem 5 *Suppose that p_{ij} is independent from $\mathbf{Z}^j$ (the jth column of Z), for all i and j. In other words the distribution of p_{ij} conditional on $\mathbf{Z}^j$ is identical to the unconditional distribution. Then*

$$E_Z[p_i^S \mid \mathcal{T}] = E_Z[\bar{D}_i \mid \mathcal{T}] = \mu_i$$

Therefore as B approaches infinity the ECOC PICT and Substitution PICT will converge for any given training set; i.e. they will give identical classification rules.

The theorem basicly states that under suitable conditions both p_i^S and $\bar{D}_i$ are unbiased estimates of μ_i and both will converge to μ_i almost surely. Both p_i^S and $\bar{D}_i$ are averages over random variables. In general the variance of p_{ij} is lower than that of $1 - 2|\hat{p}_i - Z_{ij}|$ so one would expect that the Substitution PICT will outperform the ECOC PICT for lower values of B. Figure 2 illustrates that this is indeed the case.

It is unlikely the assumption of independence is realistic. However, empirically it is well known that trees are unstable and a small change in the training data can cause a large change in the structure of the tree so it may be reasonable to suppose that the correlation between p_{ij} and $\mathbf{Z}^j$ is low.

To test this empirically we ran the ECOC and Substitution PICTs on a simulated data set. The data set was composed of 26 classes. Each class was distributed as a bivariate normal with identity covariance matrix and uniformly distributed means. Each training data set consisted of 10 observations from each class. Figure 3 shows a plot of the estimated probabilities for each of the 26 classes and 1040 test data points averaged over 10 training data sets. The probability estimates are calculated based on a matrix with 100 columns

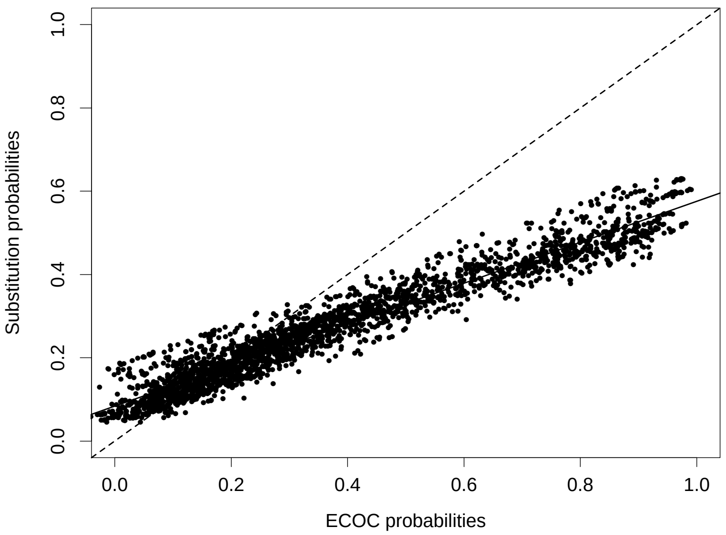

Figure 3: Probability estimates from both the ECOC and Substitution PICTs

(i.e. $B = 100$). Only points where the true posterior probability is greater than 0.01 have been plotted since classes with insignificant probabilities are unlikely to affect the classification. If the two methods were producing identical estimates we would expect the data points to lie on the dotted 45 degree line. Clearly this is not the case. The Substitution PICT is systematically shrinking the probability estimates. However there is a very clear linear relationship ($R^2 \approx 95\%$) and since we are only interested in the $\arg\max$ for each test point we might expect similar classifications. This is indeed the case. Fewer than 4% of points are correctly classified by one method but not the other.

3.2 Why does the Substitution PICT work?

The Substitution PICT is an example of a family of classification techniques known as *Majority Vote Classifiers*. A Majority Vote Classifier works by producing a large number of classifications and then classifying to the class that receives the greatest number of *votes* or classifications. (This voting may be weighted or unweighted). Some recent examples of such classifiers are Boosting (Freund and Schapire (1995)) and Bagging (Breiman (1996b)). Majority Vote Classifiers have shown a great deal of promise and a number of people have attempted to explain their success. The explanations generally fall into one of two categories which we call *Classical* (Kong and Dietterich (1995), Breiman (1996)) and *Modern* (Schapire and Freund et al. (1997), Breiman (1997)). The Classical theories rely on generalizations of bias and variance concepts as used in regression, while the Modern theories develop explanations that are more specific to classifiers. To date we do not believe that any one theory provides a comprehensive explanation. Here we present some ideas that fall in the Classical category. They are not intended as a rigorous theory but as an attempt to gain some insight.

The fact that p_i^S is an average of probability estimates suggests that a reduction in variability, without a complementary increase in bias, may be an explanation for the success of the Substitution PICT. This observation alone can not provide the answer, however, because it has been clearly demonstrated (see for example Friedman (1996)) that a reduction in variance of the probability estimates does not necessarily correspond to a reduction in the error rate. The quantity that we are interested in is not the individual probabilities but $\arg\max_j p_j$. Now $i = \arg\max_j p_j$ iff $p_i - p_j > 0 \quad \forall j \neq i$. So what we are really interested in are the random variables $p_i - p_j$. However, even the variances of these variables are not enough because variance is not independent of scale. For example by dividing all the probabilities by 2 we could reduce the variance by a factor of 4 but the probability that $p_i - p_j > 0$ would remain unchanged. A better quantity to consider is the coefficient of variation,

$$CV(p_i - p_j) = \sqrt{\frac{Var(p_i - p_j)}{(E(p_i - p_j))^2}}$$

If the probability estimates are normally distributed there is a direct correspondence between $CV(p_i - p_j)$ and the error rate i.e. the lower $CV(p_i - p_j)$ the lower the error rate. An assumption of normality may not be too bad, but in any case we would expect a similar relationship for any *reasonable* distribution. For example, if p_i^T is the probability estimate for the ith class from an ordinary k class tree classifier, we might suppose that the Substitution PICT will have a superior performance provided

$$CV_{\mathcal{T},Z}(p_i^S - p_j^S) < CV_{\mathcal{T}}(p_i^T - p_j^T) \tag{3}$$

Note that CV is calculated over both $\mathcal{T}$ and Z for the Substitution probabilities but only over $\mathcal{T}$ for the tree probabilities. To examine when (3) might hold we use the following semi-parametric model for the probability estimates,

$$\begin{aligned} p_i^S &= \alpha_S f(q_i) + \sigma_S \epsilon_i^S \quad E_{\mathcal{T},Z}\epsilon_i^S = 0 \\ p_i^T &= \alpha_T f(q_i) + \sigma_T \epsilon_i^T \quad E_{\mathcal{T}}\epsilon_i^T = 0 \end{aligned}$$

where f is an arbitrary increasing function, α_S and α_T are positive constants and $\boldsymbol{\epsilon}^S = (\epsilon_1^S, \ldots, \epsilon_k^S)$ and $\boldsymbol{\epsilon}^T = (\epsilon_1^T, \ldots, \epsilon_k^T)$ have arbitrary but identical distributions. Recall that $q_i = P(G = i \mid X)$. This model makes few assumptions about the specific form of the probability estimates but does assume that the ratio Ep_i^S/Ep_i^T is constant and that the error terms ($\epsilon^{\mathbf{S}}$ and $\epsilon^{\mathbf{T}}$) have the same distribution.

Under this modeling assumption it can be shown that (3) holds iff

$$\frac{\sigma_S}{\alpha_S} < \frac{\sigma_T}{\alpha_T} \tag{4}$$

(4) states that the standardized variance of the Substitution PICT is less than that for the tree classifier. Note that (4) is also equivalent to the signal to noise ratio of the k class tree classifier being less than that of the Substitution PICT.

The question remains, under what conditions will (4) hold? The probability estimates from the Substitution PICT are formed from an average of B correlated random variables (p_{ij}) so we know that σ_S (which depends on B) will decrease to a positive limit as B increases. Intuitively this suggests that (4) will hold provided

1. B is large enough (so we are close to the limit),
2. $$\gamma = \frac{Var_{\mathcal{T}}(p_i^T/\alpha_T)}{Var_{\mathcal{T},Z}(p_{i1}/\alpha_S)}$$

   is large enough (so the standardized variance of p_{ij} is not too large relative to that of p_i^T),
3. and $\rho = Corr_{\mathcal{T},Z}(p_{i1}, p_{i2})$ is low enough (so that a large enough reduction can be achieved by averaging).

Note that γ is the ratio of the squared noise to signal ratio (NSR) of the k class tree classifier to that of a *single tree* from the Substitution PICT. In fact we can formalize this intuition in the following theorem.

Theorem 6 *Under the previously stated semi-parametric model assumptions, (3) and (4) will hold iff*

$$\rho < \gamma \qquad (\rho \textit{ is small relative to } \gamma) \tag{5}$$

and

$$B \geq \frac{1-\rho}{\gamma-\rho} \qquad (B \textit{ is large enough}) \tag{6}$$

Further more, if either $k = 2$ (there are only 2 classes) or the error terms are normally distributed, then (5) and (6) are sufficient to guarantee a reduction in the error rate.

Now there is reason to believe that in general ρ will be small. This is a result of the empirical variability of tree classifiers. A small change in the training set can cause a large change in the structure of the tree and also the final probability estimates. So by changing the super group coding we might expect a probability estimate that is fairly unrelated to previous estimates and hence a low correlation.

To test the accuracy of this theory we examined the results from the simulation performed in Section 3.1. We wished to estimate γ and ρ. For this data it was clear that f could be well approximated by a linear function so our estimates for α_S and α_T were obtained using least squares. The following table summarizes our estimates for the variance and standardizing (α) terms from the simulated data set.

Classifier	$Var(p_i)$	α	$Var(p_i/\alpha)$
Substitution PICT	0.0515	0.3558	0.4068
Tree Method	0.0626	0.8225	0.0925

The table indicates that, when we account for the shrinkage in the Substitution PICT probability estimates ($\alpha_S = 0.3558$ vs $\alpha_T = 0.8225$), the NSR for a *single tree* from the Substitution PICT is over 4 times that of an ordinary k class tree (0.4068 vs 0.0925). In other words the estimate for γ is $\hat{\gamma} = 0.227$ so the signal to noise ratio of a single tree in the Substitution PICT is only about a quarter of that from an ordinary tree classifier. However, the estimate for ρ was very low at 0.125.

It is clear that ρ is less than γ so provided B is large enough we expect to see an improvement by using the Substitution PICT. From Theorem 6 we can estimate the required size of B as

$$B \geq \frac{1-\hat{\rho}}{\hat{\gamma}-\hat{\rho}} \approx 9$$

We see from Figure 2 that the Substitution error rate drops below that of the tree classifier at almost exactly this point, providing some validation for the theory.

4 Conclusion

The ECOC PICT was originally envisioned as an adaption of error coding ideas to classification problems. Our results indicate that the error coding matrix is simply a method for randomly sampling from a fixed distribution. This idea is very similar to the Bootstrap where one randomly samples B times from the empirical distribution for a fixed data set. There one is trying to estimate the variability of some parameter. The estimate will have two sources of error, randomness caused by sampling from the empirical distribution and the randomness from the data set itself. In our case we have the same two sources of error, error caused by sampling from $1 - 2|\hat{p}_j - Z_{ij}|$ to estimate μ_i and errors caused by using μ_i to classify. In both cases the first sort of error will decline rapidly and it is the second type which is of primary interest. It is possible to motivate the reduction in error rate from using $\arg\max_i \mu_i$ in terms of a decrease in variability, provided B is large enough and our correlation (ρ) is small enough.

References

Breiman, L. (1996) Bias, Variance, and Arcing Classifiers, Technical Report 460, Statistics Department, University of California, Berkeley

Breiman, L. (1996b) Bagging Predictors, Machine Learning 26, No. 2, pp. 123-140

Breiman, L. (1997) Arcing the Edge, Unpublished

Dietterich, T.G. and Bakiri G. (1995) Solving Multiclass Learning Problems via Error-Correcting Output Codes, Journal of Artificial Intelligence Research 2 (1995) 263-286

Dietterich, T. G. and Kong, E. B. (1995) Error-Correcting Output Coding Corrects Bias and Variance, Proceedings of the 12th International Conference on Machine Learning pp. 313-321 Morgan Kaufmann

Freund, Y. and Schapire, R. (1995), A decision-theoretic generalization of on-line learning and an application to boosting, to appear Journal of Computer and System Sciences

Friedman, J.H. (1996) On Bias, Variance, 0/1-loss, and the Curse of Dimensionality, Dept. of Statistics, Stanford University, Technical Report

Hoeffding, W. (1963) Probability Inequalities for Sums of Bounded Random Variables. "Journal of the American Statistical Association", March, 1963

Schapire, R., Freund, Y., Bartlett, P., and Lee, W. (1997) Boosting the Margin,
(available at http://www.research.att.com/ yoav)