Error coding and PaCT's

Gareth James
and
Trevor Hastie
Dept. of Statistics, Stanford University

January 10, 1997

Abstract

A new class of *plug in* classification techniques have recently been developed in the statistics literature. A plug in classification technique (PaCT) is a method that takes a standard classifier (such as LDA or nearest neighbors) and plugs it into an algorithm to produce a new classifier. The standard classifier is known as the *Plug in Classifier* (PiC). 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

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

- Produce a k by n (n large) binary coding matrix, ie 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 $\mathbf{Z}^j$.
- Use the first column of the coding matrix ($\mathbf{Z}^1$) to create two *super* groups by assigning all groups with a one in the corresponding element of $\mathbf{Z}^1$ to super group one and all other groups to super group zero.
- Train your plug in classifier (PiC) on these two super groups.
- Repeat the process for each of the n columns $(\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^n)$ to produce n trained classifiers.
- For a new test point apply each of the n classifiers to it. Each classifier will produce a $\hat{p}_j$ which is the estimated probability the test point comes from the jth super group one. This will produce a vector of probability estimates, $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)^T$.
- To classify the point calculate $D_i = \sum_{j=1}^{n} |\hat{p}_j - Z_{ij}|$ for each of the k groups (ie 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 group with lowest L1 distance or equivalently $\arg_i \min D_i$

I call this the L1 PaCT. 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 PiC's have been tested. The best results were obtained by using *tree's*, so all the experiments in this paper are stated using a standard CART PiC. Note however, that the theorems are completely general to any PiC.

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 indicate that a randomized coding matrix should match (or exceed) the performance of a *designed* matrix.

2 Bias and Variance Effects

As in regression estimation problems, the accuracy of a classifier can be considered in terms of bias and variance. It has been postulated that the L1 PaCT is a method for converting a tree classifier (which is high variance - low bias) into a low variance - low bias classifier. To test this hypothesis one must first define what is meant by bias and variance of a classifier. There have recently been a number of definitions proposed (Dietterich and Bakiri (1995), Breiman (1996), Friedman (1996), Kohavi and Wolpert (1996), Tibshirani(1996)). James (1997) suggests the following decomposition.

$$ER = \underbrace{P(Y \neq SY)}_{\text{Bayes ER}} + \underbrace{[P(Y \neq SC) - P(Y \neq SY)]}_{\text{Systematic Effect (SE)}} + \underbrace{[P(Y \neq C) - P(Y \neq SC)]}_{\text{Variance Effect (VE)}}$$

where Y is the class label, SY is the Bayes classifier, C is our classifier, and SC is the mode of C (ie the most common group that C classifies to at each value of X). Note that C is a random variable — it depends on the training data. The intuition here is that the mode (SC) is the *closest* non random classifier to C (in terms of minimizing the 0-1 loss function). So $P(Y \neq SC) - P(Y \neq SY)$ is the increase in error of using the systematic part of C (SC) rather than the Bayes Classifier. Similarly $P(Y \neq C) - P(Y \neq SC)$ is the change in error introduced by the variability of C about SC. SE is the equivalent of Bias^2 in the regression setting and VE equates to $Var(\hat{Y})$.

To illustrate the effect of the L1 PaCT on these quantities we produced a simulated data set of 26 classes. Each class was distributed as a bivariate normal with identity covariance matrix and uniformly distributed means. The training data consisted of 10 observations from each group. The estimates are averaged over 20 different random training sets and each of the L1 PaCT's are also averaged over 5 random coding matrices. The test data consisted of 40 observations from each group. The Bayes error rate (which is the minimum possible error rate) for this test set was 23%. We tested three classifiers. The first was CART, the second was the L1 PaCT with $n = 26$ and the last was the L1 PaCT with $n = 100$. The following table illustrates the results for the three classifiers tested.

Classifier	Bayes Error	SE	VE	Reducible Error	Total Error
CART	0.231	0.020 (0.002)	0.073 (0.002)	0.093 (0.003)	0.324 (0.003)
L1 PaCT ($n = 26$)	0.231	0.015 (0.001)	0.062 (0.001)	0.077 (0.001)	0.308 (0.001)
L1 PaCT ($n = 100$)	0.231	0.017 (0.001)	0.042 (0.001)	0.059 (0.001)	0.290 (0.001)

The numbers in parentheses are approximate standard errors using a mixture of bootstrap and normal approximations. It is clear that the L1 PaCT is lowering the reducible error (Total error - Bayes Error). With $n = 100$ the reducible error has declined by 36% over using CART. Most of this reduction is attributable to the Variance Effect though there does seem to be some indication of a decrease in the Systematic Effect also. These figures tend to support the hypothesis that the method is reducing errors by decreasing the variance.

Note that there is a decrease in the error rate between the 26 and 100 column classifiers. This effect is very common and is apparent in real data sets also. In fact if we consider figure 1, which is a plot of the number of columns (n) of our coding matrix vs error rate for the LETTER data set (available from the Irvine Repository of machine learning), it is noticeable that there is a strong relationship. We will see later that the exact structure of this relationship can be theoretically explained.

3 Theory

The above results show that the L1 PaCT reduces the error rate by reducing the variance effect and to a lesser extent also the systematic effect. However, they do not shed any light on why this should be the case. To explore this question we need to develop the probability structure of the L1 PaCT. The coding matrix, Z, is central to the L1 PaCT. In the past the usual approach has been to choose one with as large a separation between rows ($\mathbf{Z}_i$) 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 the tradeoffs between a

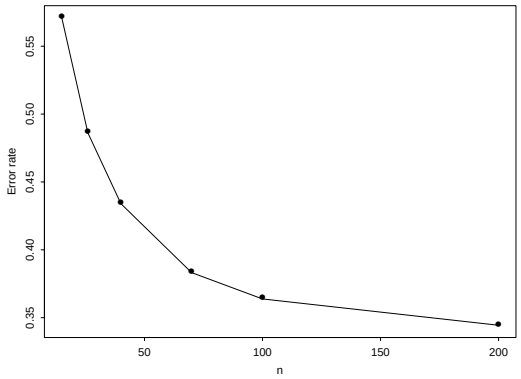

Figure 1: Error rates for differing values of n

designed (deterministic) and a *completely randomized* matrix.

Some of the results that follow will make use of the following assumption.

$$E[\hat{p}_j \mid Z, X] = \sum_{i=1}^{k} Z_{ij}\beta_i^x = \mathbf{Z}^{j^T}\boldsymbol{\beta}^x \quad j = 1 \dots n \tag{1}$$

where $\beta_i^x = P(G_i \mid X)$ is the posterior probability that the test observation is from group i given that our predictor variable is X. 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.

3.1 Deterministic Coding Matrix

Obviously the L1 PaCT can not outperform the Bayes Classifier. However we would hope that it would achieve the Bayes Error Rate when we use the Bayes Classifier as our PiC. If not we would be better off using the PiC directly. We have defined this property as Bayes Optimality.

Definition 1 *A PaCT is said to be Bayes Optimal if, for any test set, it always classifies to the bayes group when the Bayes Classifier is our PiC.*

For the L1 PaCT this means that $\arg_i \max \beta_i^x = \arg_i \min D_i^x$ for all values of X when we use the Bayes Classifier as our PiC. The following theorem and corollary will help us to determine under what circumstances the L1 PaCT is Bayes Optimal.

Theorem 1 *If assumption 1 holds and Z is deterministic then*

$$ED_i^x = \sum_{l \neq i} \beta_l^x \sum_{j=1}^{n} (Z_{lj} - Z_{ij})^2 \quad i = 1 \dots k$$

Corollary 1 *If we use the Bayes classifier as our plug in classifier (ie use the Bayes classifier to produce $\hat{p}_j$) then D_i^x is non random and*

$$D_i^x = \sum_{l \neq i} \beta_l^x \sum_{j=1}^{n} (Z_{lj} - Z_{ij})^2 \quad i = 1 \dots k$$

The corollary suggests that $\arg_i \max \beta_i^x$ may not equal $\arg_i \min D_i^x = \arg_i \min \sum_{l \neq i} \beta_l^x \sum_{j=1}^n (Z_{lj} - Z_{ij})^2$ for all X. In fact the following theorem tells us that only in very restricted circumstances will the L1 PaCT be Bayes Optimal.

Theorem 2 *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 n and k there is no known way to generate a matrix with this property so the L1 PaCT will not be Bayes Optimal.

3.2 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 $\bar{D}_i^x = D_i^x/n = \sum |\hat{p}_j - Z_{ij}|/n$. Then we have the following theorem which indicates that by randomizing we have eliminated one of the concerns with a deterministic matrix.

Theorem 3 *When the coding matrix is randomly chosen the L1 PaCT is asymptotically Bayes Optimal ie* $Pr(\arg_i \min \bar{D}_i^x = \arg_i \max \beta_i^x) \to 1$ *as* $n \to \infty$

This theorem is a consequence of the strong law and theorem 4.

Theorem 4 *Under assumption 1 for a randomly generated coding matrix*

$$E\bar{D}_i^x = \frac{1}{2}(1 - \beta_i^x) \quad i = 1 \ldots k$$

or in vector notation

$$E\bar{\mathbf{D}}^x = \frac{1}{2}(1 - \boldsymbol{\beta}^{\mathbf{x}})$$

This tells us that $\arg_i \min E\bar{D}_i^x = \arg_i \max \beta_i^x$ which gives us the first indication of why the L1 PaCT is successful. If $\bar{\mathbf{D}}^x$ had no randomness this would say that the L1 PaCT is equivalent to the Bayes Rule.

Of course in general $\bar{\mathbf{D}}^x$ will vary so there is no guarantee that $\arg_i \min \bar{D}_i^x = \arg_i \max \beta_i^x$. However if the variability of $\bar{\mathbf{D}}^x$ is low we might hope that $Pr(\arg_i \min \bar{D}_i^x = \arg_i \max \beta_i^x)$ is high. To evaluate this probability we need to consider the distribution of $\bar{\mathbf{D}}^x$.

Let $\mu_i^x = E[|\hat{p}_1 - Z_{i1}| \mid \text{Training set}]$ and $\boldsymbol{\mu}^x = (\mu_1^x, \mu_2^x, \ldots, \mu_k^x)^T$. Then $\boldsymbol{\mu}^x$ is the expected value of $\bar{\mathbf{D}}^x$ conditional on the training set. Theorem 5 gives the asymptotic distribution of $\bar{\mathbf{D}}^x$.

Theorem 5 *For any fixed training set*

$$\sqrt{n}(\bar{\mathbf{D}}^x - \boldsymbol{\mu}^x) \Rightarrow \mathbf{N}(\mathbf{0}, \Sigma)$$

If we remove the conditioning on the training set, $\boldsymbol{\mu}^x$ *is a random variable with* $E\boldsymbol{\mu}^x = \frac{1}{2}(1 - \boldsymbol{\beta}^x)$.

Theorem 5 is a simple application of the multi-variate central limit theorem. Notice that D_i^x is just an average of n random variables ($|\hat{p}_j - Z_{ij}|$). If we condition on a training set then each of these random variables are independent and identically distributed (with mean μ_i^x) because each one will only depend on $\mathbf{Z}^j$.

This leads to an important result.

Theorem 6 *If we randomly choose* Z *then for any fixed* X

$$|Pr(\arg_i \min \bar{D}_i^x = \arg_i \max \beta_i^x) - Pr(\arg_i \min \mu_i^x = \arg_i \max \beta_i^x)| \leq c \cdot e^{-mn}$$

for some constants c *and* m *or equivalently*

$$Pr(\arg_i \min \bar{D}_i^x \neq \arg_i \min \mu_i^x) \leq c \cdot e^{-mn}$$

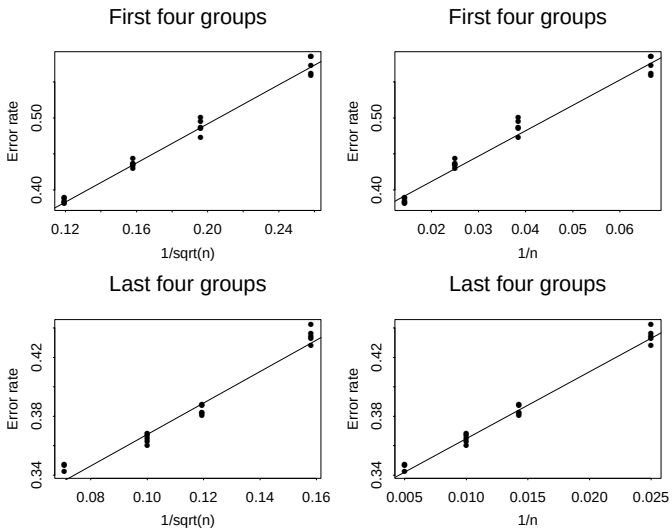

Figure 2: Comparison's of $1/\sqrt{n}$ and $1/n$ vs error rates

Note that theorem 6 does not depend on assumption 1. What this tells us is that the probability we correctly classify a point using $\bar{\mathbf{D}}^x$ is equal to the probability using $\boldsymbol{\mu}^x$ plus an error term which decreases exponentially in the limit. This exponential decay is a result of the central limit theorem. From the CLT we know that $\bar{\mathbf{D}}^x = \boldsymbol{\mu}^x + O_p(1/\sqrt{n})$ so $\bar{\mathbf{D}}^x$ only approaches $\boldsymbol{\mu}^x$ at a rate of $1/\sqrt{n}$. However because of the normality of $\bar{\mathbf{D}}^x$ it can be shown that

$$Pr(\arg_i \min \bar{D}_i^x \neq \arg_i \min \mu_i^x) \leq \frac{k}{m\sqrt{n}} \frac{1}{\sqrt{2\pi}} e^{-m^2 n/2} \quad \text{for some } m > 0$$

which gives an exponential decay. This only gives an upper bound on the error rate and does not necessarily indicate the behavior for smaller values of n. Under certain conditions a Taylor expansion indicates that $Pr(\arg_i \min \bar{D}_i^x \neq \arg_i \min \mu_i^x) \approx 0.5 - m\sqrt{n}$ for small values of $m\sqrt{n}$. So we might expect that for smaller values of n the error rate decreases as some power of n but that as n increases the change in error rate looks more and more exponential.

To test this hypothesis we calculated the error rates for 6 different values of n $(15, 26, 40, 70, 100, 200)$ on the LETTER data set. Figure 2 illustrates the results. The first row shows plots of the error rate vs $1/n$ or $1/\sqrt{n}$ for the first four values of n $(15, 26, 40, 70)$. Each value of n contains 5 points corresponding to 5 random matrices. Each point is the average over 20 random training sets. It is clear that the error rate looks far more like $1/\sqrt{n}$ than $1/n$. However if we consider the second row which contains plots of the last four groups $(40, 70, 100, 200)$ the trend is reversed. This supports our hypothesis that the error rate is moving through the powers of n towards an exponential fit. Figure 3 illustrates this effect in an alternate manor. Here we have two curves. The lower curve is the best fit of $1/\sqrt{n}$ 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 $1/n$ to the last four groups. It fits those groups well but over predicts errors for the first two groups.

We can see from figure 3 that even for relatively low values of n the reduction in error rate has slowed substantially. This indicates that almost all the remaining errors are as a result of $\arg_i \min \mu_i^x \neq \arg\max \beta_i^x$ which we can not eliminate by changing the coding matrix. Thus, the coding matrix is simply a method for randomly sampling from the distribution of $|\hat{p}_j - \mathbf{Z}^j|$ to estimate $\boldsymbol{\mu}^x$ (its mean). It is well known that the optimal way to estimate such a parameter is by random sampling so it should not be possible to improve on this by *designing* the coding matrix. It is clear that what we are really interested in is $\boldsymbol{\mu}^x$. We can, with relative ease, estimate $\boldsymbol{\mu}^x$ to a high level of accuracy. From theorem 5 we know that $E\boldsymbol{\mu}^x = \frac{1}{2}(1 - \boldsymbol{\beta}^x)$ so if the variability of $\boldsymbol{\mu}^x$ is low so will be the error rate. In general the variability of $\boldsymbol{\mu}^x$ will depend on the

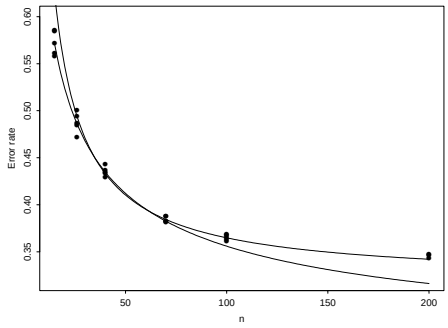

Figure 3: Best fit curves for rates $1/\sqrt{n}$ and $1/n$

variability of the PiC. Theorem 7 shows how these quantities are related.

Theorem 7 *Under assumption 1*

$$Var(\mu_i^x) = Var(\hat{p}_1) - \frac{1}{4}\beta_i^2 - E[Var[|\hat{p}_1 - Z_{i1}||\, Training\ Data]]$$

Unfortunately the last two terms are unknown. However they are both positive so this does provide an upper bound for $Var(\mu_i^x)$. This area is the subject of future research.

4 Conclusion

The L1 PaCT 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 we randomly sample from the empirical distribution for a fixed data set. There you are trying to estimate the variability of some parameter. Your 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 $|\hat{p}_j - Z_{ij}|$ to estimate $\boldsymbol{\mu}^x$ and error's caused by $\boldsymbol{\mu}^x$ itself. In both cases the first sort of error will reduce rapidly and it is the second type we are really interested in. It is apparent (based on empirical evidence) that the variability of $\boldsymbol{\mu}^x$ is lower than that of directly estimating $\boldsymbol{\beta}^x$ using our PiC.

References

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

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. (1996) Experiments with a new boosting algorithm, "Machine Learning : Proceedings of the Thirteenth International Conference", July, 1996

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

James, G. and Hastie, T. (1997) A Generalized Prediction Error Decomposition

Kohavi, R. and Wolpert, D.H. (1996) Bias Plus Variance Decomposition for Zero-One Loss Functions, ftp starry.stanford.edu/pub/ronnyk/biasVar.ps

Tibshirani (1996) Bias, Variance and Prediction Error for Classification Rules, Dept. of Statistics, University of Toronto, Technical Report