CLASSIFICATION USING DIRICHLET PRIORS WHEN THE TRAINING DATA ARE MISLABELED

Robert S. Lynch, Jr. [*]

Naval Undersea Warfare Center
Newport, RI 02841
lynchrs@npt.nuwc.navy.mil

Peter K. Willett [†]

University of Connecticut
Storrs, CT 06269
willett@mailhost.engr.uconn.edu

ABSTRACT

The average probability of error is used to demonstrate performance of a Bayesian classification test (referred to as the *Combined Bayes* Test (CBT)) given the training data of each class are mislabeled. The CBT combines the information in discrete training and test data to infer symbol probabilities, where a uniform Dirichlet prior (i.e., a noninformative prior of complete ignorance) is assumed for all classes. Using this prior it is shown how classification performance degrades when mislabeling exists in the training data, and this occurs with a severity that depends on the value of the mislabeling probabilities. However, an increase in the mislabeling probabilities are also shown to cause an increase in M^* (i.e., the best quantization fineness). Further, even when the actual mislabeling probabilities are known by the CBT, it is not possible to achieve the classification performance obtainable without mislabeling.

1. INTRODUCTION

In this paper, performance of a Bayesian classification test (referred to as the *Combined Bayes* Test (CBT)) is illustrated given the training data of each class are mislabeled. The CBT combines the information in discrete training and test data to infer symbol probabilities which are assumed to have, for each class, a prior uniform Dirichlet distribution (i.e., a noninformative prior representing complete ignorance).

Here, the term "discrete" means that data used to represent each class can take on one of M possible values. This data may have arisen naturally in its M-level form, or it may have been derived by quantizing feature vectors. Also, for each class, there are certain labeled realizations of this (M-valued) data, and this is referred to as "training" data. That is, there are N_k realizations under class k and N_l realizations under class l.

Now, with the situation of interest, the training data of each class are assumed to be made up of two parts. That is, a correctly labeled part and a mislabeled part. Specifically, the N_k (N_l) training data of class k (l) consist of N_{kk} (N_{ll}) correctly labeled observations occurring with probability $1-\alpha_k$ ($1-\alpha_l$), and a remaining N_{kl} (N_{lk}) mislabeled observations occurring with probability α_k (α_l). With this, it is assumed that $N_{\vec{y}}$ unlabeled "test" data are observed, which are to be tested by a classifier. Therefore, the problem addressed here is to illustrate, using a formula for the average probability of error ($P(e)$), the effect that mislabeled training data has on classifying unknown test data.

In previous work, performance of the CBT was examined theoretically using $P(e)$, and correctly labeled training data. In particular, $P(e)$ was investigated as a function of the number of discrete symbols used, M (i.e., the quantization fineness). A minimum point of $P(e)$ was found given a fixed amount of training data and test data [7]; that is, a quantization fineness referred to as M^*, which is related to the *Hughes phenomenon* [3] of pattern recognition [4]. Further, in addition to this work, performance of the CBT has been compared to other classification tests, [6, 9], and it has been successfully applied to data reduction [8].

2. CLASSIFICATION WITH MISLABELED TRAINING DATA

2.1. Combined Multinomial Model

It is assumed that there exists a pair of probability vectors, $\vec{p_k}$ and $\vec{p_l}$, the i^{th} elements of which denote the probability of a symbol of type i being observed under the respective classes k and l. The fundamental model for this testing method is thus formulated based on the number of occurrences of each discrete symbol being an i.i.d. multinomially distributed random variable. Therefore, the joint distribution for the frequency of occurrence of all training and test data with the test data, $\vec{y}$, a member of class k is given by

[*] Supported by a Naval Undersea Warfare Center In-House Laboratory Independent Research grant.

[†] Supported by ONR under contract N00014-97-1-0502, and by AFOSR under contract F49620-97-1-0198.

$$
\begin{aligned}
&f\left(\vec{x}_{kk}, \vec{x}_{lk}, \vec{x}_{ll}, \vec{x}_{kl}, \vec{y} | \vec{p}_k, \vec{p}_l, H_k; \alpha_k, \alpha_l\right) = \\
&\quad N_{kk}!N_{lk}!N_{ll}!N_{kl}!N_{\vec{y}}! \prod_{i=1}^{M} \frac{p_{k,i}^{x_{kk,i}+x_{lk,i}+y_i}\, p_{l,i}^{x_{ll,i}+x_{kl,i}}}{x_{kk,i}!x_{lk,i}!x_{ll,i}!x_{kl,i}!y_i!} \\
&\quad \times \frac{N_k!}{N_{kk}!N_{kl}!} \left(\alpha_k\right)^{N_{kl}} \left(1-\alpha_k\right)^{N_{kk}} \\
&\quad \times \frac{N_l!}{N_{ll}!N_{lk}!} \left(\alpha_l\right)^{N_{lk}} \left(1-\alpha_l\right)^{N_{ll}} \qquad (1)
\end{aligned}
$$

where [1]

$k, l \in \{$target, nontarget$\}$, and $k \neq l$;
H_k is the hypothesis defined as $\vec{p}_{\vec{y}} = \vec{p}_k$;
M is the number of discrete symbols;
$x_{kk,i}$ is the number of occurrences of the i^{th} symbol in the correctly labeled training data for class k;
N_{kk} $\left\{N_{kk} = \sum_{i=1}^{M} x_{kk,i}\right\}$ is the total number of occurrences of the M symbols in the correctly labeled training data for class k;
$x_{kl,i}$ is the number of occurrences of the i^{th} symbol in the mislabeled training data for class k, and which belong to class l;
N_{kl} $\left\{N_{kl} = \sum_{i=1}^{M} x_{kl,i}\right\}$ is the total number of occurrences of the M symbols in the mislabeled training data for class k;
$x_{k,i} = x_{kk,i} + x_{kl,i}$ is the number of occurrences of the i^{th} symbol in all training data for class k;
N_k $\left\{N_k = N_{kk} + N_{kl} = \sum_{i=1}^{M} x_{k,i}\right\}$ is the total number of occurrences of the M symbols in all training data for class k;
y_i is the number of occurrences of the i^{th} symbol in the test data;
$N_{\vec{y}}$ $\left\{N_{\vec{y}} = \sum_{i=1}^{M} y_i\right\}$ is the total number of occurrences of the M symbols in the test data;
$p_{k,i}$ $\left\{\sum_{i=1}^{M} p_{k,i} = 1\right\}$ is the probability of the i^{th} symbol conditioned on the test data being an element of class k.

2.2. Combined Bayes Test

Rather than assuming that $\vec{p}_k$ and $\vec{p}_l$ are simply unknown parameters to be estimated (and use a combined generalized likelihood ratio test [6]), our approach here is to give them prior distributions. Nothing a priori is known about the probability vectors, and hence the appropriate prior is one of complete ignorance; the uniform Dirichlet.[2]

The first step in developing the CBT for mislabeled training data is to apply the Dirichlet,

$$
f\left(\vec{p}_k\right) = (M-1)!\delta\left(1 - \sum_{i=1}^{M} p_{k,i}\right) \qquad (2)
$$

to the formula of (1) under each class k and l, and then integrate, respectively, over the *positive unit-hyperplane* resulting in

$$
\begin{aligned}
&f\left(\vec{x}_{kk}, \vec{x}_{lk}, \vec{x}_{ll}, \vec{x}_{kl}, \vec{y} | H_k; \alpha_k, \alpha_l\right) = \\
&\quad \frac{\left((M-1)!\right)^2 N_{kk}!N_{lk}!N_{ll}!N_{kl}!N_{\vec{y}}!}{\left(N_{kk} + N_{lk} + N_{\vec{y}} + M - 1\right)! \left(N_{ll} + N_{kl} + M - 1\right)!} \\
&\quad \times \prod_{i=1}^{M} \frac{\left(x_{kk,i} + x_{lk,i} + y_i\right)! \left(x_{ll,i} + x_{kl,i}\right)!}{x_{kk,i}!x_{lk,i}!x_{ll,i}!x_{kl,i}!y_i!} \\
&\quad \times \frac{N_k!}{N_{kk}!N_{kl}!} \left(\alpha_k\right)^{N_{kl}} \left(1-\alpha_k\right)^{N_{kk}} \\
&\quad \times \frac{N_l!}{N_{ll}!N_{lk}!} \left(\alpha_l\right)^{N_{lk}} \left(1-\alpha_l\right)^{N_{ll}} \qquad (3)
\end{aligned}
$$

Continuing, formula (3) is now expressed in terms of the complete training data vectors, $\vec{x}_k$ and $\vec{x}_l$. This is accomplished by appropriately substituting the definitions $\vec{x}_k = \vec{x}_{kk} + \vec{x}_{kl}$ and $\vec{x}_l = \vec{x}_{ll} + \vec{x}_{lk}$ into formula (3), followed by summing over all arrangements of mislabeled training data, yielding

$$
\begin{aligned}
&f\left(\vec{x}_k, \vec{x}_l, \vec{y} | H_k; \alpha_k, \alpha_l\right) = \sum_{\vec{x}_{kl}=\vec{0}}^{\vec{x}_k} \\
&\quad \sum_{\vec{x}_{lk}=\vec{0}}^{\vec{x}_l} f\left(\vec{x}_k - \vec{x}_{kl}, \vec{x}_{lk}, \vec{x}_l - \vec{x}_{lk}, \vec{x}_{kl}, \vec{y} | H_k; \alpha_k, \alpha_l\right) \\
&\qquad (4)
\end{aligned}
$$

Using this result, the CBT is then given by the ratio of (4) to its analogous formula under class l (i.e., conditioned on H_l), and it appears as

$$
\frac{f\left(\vec{x}_k, \vec{x}_l, \vec{y} | H_k; \alpha_k, \alpha_l\right)}{f\left(\vec{x}_k, \vec{x}_l, \vec{y} | H_l; \alpha_k, \alpha_l\right)} \underset{H_l}{\overset{H_k}{\gtrless}} \tau \qquad (5)
$$

where, for minimizing the probability of error, the decision threshold τ is equal to $P\left(H_l\right)/P\left(H_k\right)$.

2.3. Probability of Error

Letting $z_k = f\left(\vec{x}_k, \vec{x}_l, \vec{y} | H_k; \alpha_k, \alpha_l\right)$ (see formulas (4) and (5) above), the average probability of error for the CBT is defined as

[1] In the following notation k and l are exchangeable.

[2] The uniform Dirichlet results when the parameters of this distribution are set to unity [1]. Note, it has been suggested in [5] that a better prior to use, given unknown true statistics, is the Dirichlet with its parameters set to one half (also see, [2]).

$$\begin{aligned} P\left(e\right) = P\left(H_k\right) P\left(z_k \le \tau z_l \mid H_k\right) \\ + P\left(H_l\right) P\left(z_k > \tau z_l \mid H_l\right) \end{aligned} \tag{6}$$

It is necessary to show the first term of (6) only as the second term is similar except for conditioning on H_l. Thus, ignoring $P\left(H_k\right)$, the first term of (6) is given by

$$\begin{aligned} P\left(z_k \le \tau z_l \mid H_k\right) = \\ \sum_{\vec{y}} \sum_{\vec{x}_k} \sum_{\vec{x}_l} \mathcal{I}\left(z_k \le \tau z_l\right) f\left(\vec{x}_k, \vec{x}_l, \vec{y} \mid H_k; \alpha_k, \alpha_l\right) \end{aligned} \tag{7}$$

where $f\left(\vec{x}_k, \vec{x}_l, \vec{y} \mid H_k; \alpha_k, \alpha_l\right)$ was defined in formula (4) above, and $\mathcal{I}(x)$ is the indicator function.

3. RESULTS

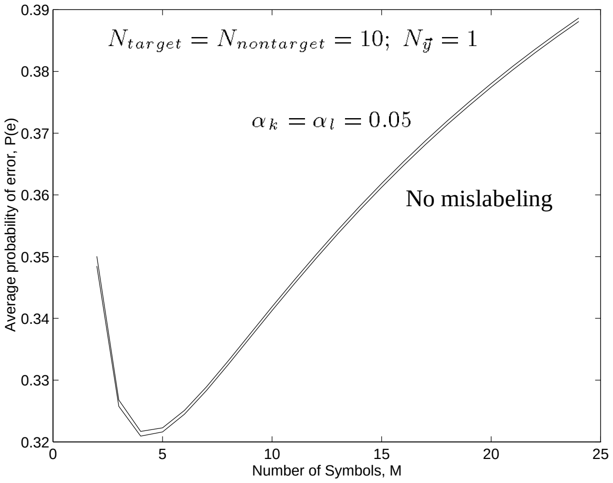

Fig. 1. CBT performance with $\alpha_k = \alpha_l = 0.05$.

Figure 1 above shows the average probability of error for the CBT when the true mislabeling probabilities are given by, $\alpha_k = \alpha_l = 0.05$. Also, for comparing performance, $P(e)$ is shown given there is no mislabeling in the training data. Notice that $P(e)$ is plotted as a function of M, and there are ten samples of training data for each class and one test observation. Additionally, in this and subsequent figures the threshold $\tau = 1$, meaning that $P\left(H_k\right)$ and $P\left(H_l\right)$ are both 0.5.

Observe in Figure 1 that in both cases $P(e)$ starts out decreasing with increasing M and is minimum at a point we call M^*, and in this case $M^* = 4$. With this, also notice that for M greater than M^* $P(e)$ steadily increases. This dependence of $P(e)$ on M reflects the fact that given a fixed amount of training and test data, a prior quantizing fineness exists which yields, on average, the "best" classification performance .[3] But, it can also be seen in Figure 1 that when mislabeling exists, performance begins to degrade in that $P(e)$ increases (although not by much for these small mislabeling probabilities; see Figures 2 and 3 below). However, it was also found that when the training data of both classes are mislabeled with the probabilities shown above, performance is identical for the CBT given it contains (equal for both classes) any mislabeling probabilities specified by the range, $0 \le \alpha_k = \alpha_l < 0.5$. Thus, in this situation, and as it turns out in all cases presented here, the mislabeling probabilities assumed by the test are irrelevant as long as they are not 0.5 or higher.

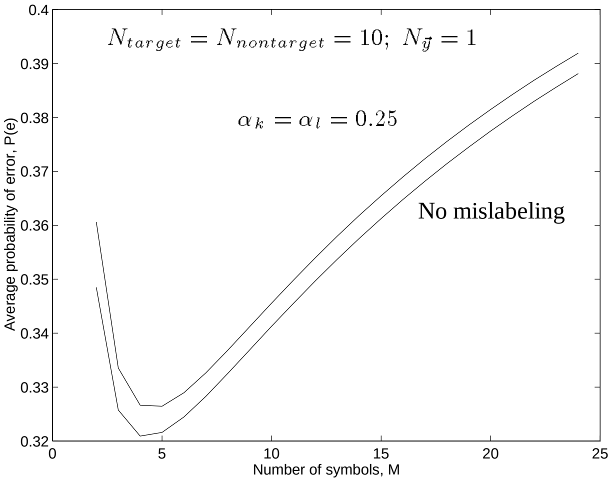

Fig. 2. CBT performance with $\alpha_k = \alpha_l = 0.25$.

In Figure 2 $P(e)$ is shown for the CBT when the true mislabeling probabilities are $\alpha_k = \alpha_l = 0.25$. With this, as in Figure 1 $P(e)$ is shown when there is no mislabeling in the training data, and again there are ten samples of training data for each class and one test observation.

Notice in this figure, as compared to Figure 1, $P(e)$ has increased given these larger mislabeling probabilities. But, more importantly, M^* has increased from 4 to 5. In other words, best classification performance occurs for situations which a priori require more discrete symbols. Further, and as mentioned above, this performance remains consistent when the CBT contains any mislabeling probabilities (equal for both classes) less than 0.5.

Figure 3, on the next page, shows $P(e)$ for the CBT when the true mislabeling probabilities are $\alpha_k = \alpha_l = 0.45$. Also, $P(e)$ is shown for the no mislabeling case, and the training and test data configurations are the same as in Figures 1 and 2. Observe $P(e)$ has increased even further with these mislabeling probabilities, and M^* is now equal to 6.

[3] This was previously pointed out in [4], and also see, [10].

This, and the results shown in the previous two figures indicate that M^* tends to increase with the mislabeling probabilities. However, this is only true for mislabeling probabilities less than 0.5. That is, performance of those cases with mislabeling probabilities greater than 0.5 mirrors the performance of those cases with mislabeling probabilities less than 0.5. In other words, performance with the mislabeling probabilities equal to 0.25 is essentially the same as when the mislabeling probabilities are equal to 0.75.

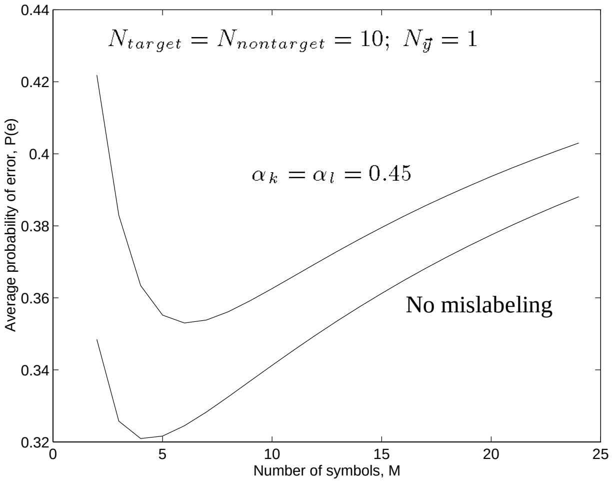

Fig. 3. CBT performance with $\alpha_k = \alpha_l = 0.45$.

4. SUMMARY

In this paper, the effect that mislabeled training data has on classification performance was demonstrated given there is no knowledge of the underlying discrete symbol probabilities, nor of the mislabeling probabilities. In general, it was shown that both the probability of error and the optimum quantization fineness, M^*, increase with the mislabeling probabilities.

Also, and not shown here, previous results obtained for the CBT with no mislabeling [6, 7] imply that $P(e)$ can be reduced in all cases shown above if the number of test observations is increased (i.e., $N_{\vec{y}} > 1$). With future work, this issue is to be explored further.

5. REFERENCES

[1] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, Wiley, New York, 1994.

[2] L. L. Campbell, "Averaging Entropy," *IEEE Trans. on Information Theory*, vol. 41, no. 1, January 1995, pp. 338-339.

[3] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, Inc., Boston, 1990.

[4] G. F. Hughes, "On the Mean Accuracy of Statistical Pattern Recognizers," *IEEE Trans. on Information Theory*, vol. 14, no. 1, January 1968, pp. 55-63.

[5] R. E. Krichevsky and V. K. Trofimov, "The Performance of Universal Encoding," *IEEE Trans. on Information Theory*, vol. 27, no. 2, March 1981, pp. 199-207.

[6] R. Lynch and P. Willett, "Classification With a Combined Information Test," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1996, pp. 3061-3064.

[7] R. Lynch and P. Willett, "Discrete Symbol Quantity and the Minimum Probability of Error for a Combined Information Classification Test," *Proceedings of the 35^{th} Annual Allerton Conference on Communication, Control, and Computing*, September 1997.

[8] R. S. Lynch, Jr. and P. K. Willett, "Bayesian Classification and Data Driven Quantization Using Dirichlet Priors," *Proceedings of the 32^{nd} Annual Conference on Information Sciences and Systems*, March 1998.

[9] R. S. Lynch, Jr. and P. K. Willett, "Testing the Statistical Similarity of Discrete Observations Using Dirichlet Priors," *Proceedings of the 1998 IEEE International Symposium on Information Theory*, August 1998.

[10] J. M. Van Campenhout, "On the Peaking of the Hughes Mean Recognition Accuracy: The Resolution of an Apparent Paradox," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 8, no. 5, May 1978, pp. 390-395.