PERGAMON Journal of the Franklin Institute 336 (1999) 809–819

Journal *of* The Franklin Institute

www.elsevier.nl/locate/jfranklin

Bayesian classification using a noninformative prior and mislabeled training data

Robert S. Lynch Jr[a,*,1], Peter K. Willett[b,2]

[a] *Environmental Acoustics, Signal Processing & Analysis Branch, Naval Undersea Warfare Center, Newport, RI 02841, USA*

[b] *Department of Electrical Engineering, University of Connecticut, Storrs, CT 06269, USA*

Received 26 December 1997; accepted in revised form 19 January 1999

Abstract

The average probability of error is used to demonstrate the performance of a Bayesian classification test (referred to as the Combined Bayes Test (CBT)) when 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 the CBT, classification performance is shown to degrade when mislabeling exists in the training data, and this occurs with a severity that depends upon the mislabeling probabilities. With this, it is shown that as the mislabeling probabilities increase M^*, which is the best quantization fineness related to the *Hughes phenomenon* of pattern recognition, also increases. Notice, that even when the actual mislabeling probabilities are known by the CBT it is not possible to achieve the classification performance obtainable without mislabeling. However, the negative effect of mislabeling can be diminished, with more success for smaller mislabeling probabilities, if a data reduction method called the Bayesian Data Reduction Algorithm (BDRA) is applied to the training data. Published by Elsevier Science Ltd.

Keywords: Dirichlet; Discrete data; Multinomial; Feature selection

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

[*] Corresponding author. Tel.: + 1-401-832-8663; fax: + 1-401-832-7477

E-mail address: lynchrs@npt.nuwc.navy.mil (R.S. Lynch)

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

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

0016-0032/99/$20.00 + 0.00 Published by Elsevier Science Ltd.
PII: S0016-0032(99)00006-X

corrupted by mislabeling. 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. These data may have arisen naturally in an M-level form, or they may have been derived by quantizing feature vectors. Also, for each class, there are certain labeled realizations of the (M-valued) data, and this is referred to as "training" data. That is, in the binary hypothesis cases considered here, 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: a correctly labeled part, and a mislabeled part. Specifically, the $N_k(N_l)$ training data of class $k(l)$ consists 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 (i.e., belonging to the other class) occurring with probability $\alpha_k(\alpha_l)$. With this, it is assumed that N_y unlabeled "test" data are observed, which are to be classified. Therefore, the problem addressed here is to illustrate, based on the average probability of error ($P(e)$), the effect that mislabeled training data has on classifying the unknown test data.

Previously, classification 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 and test data [1]. That is, associated with this minimum point a quantization fineness (M^*) was found, which is related to the Hughes phenomenon of pattern recognition [2] (see also, [3]). Additionally, performance of the CBT has been compared to other classification tests [4], and it has been successfully applied to data reduction (i.e., feature selection) [5]. In the latter case, the Bayesian Data Reduction Algorithm (BDRA) was developed, and its ability to improve performance with mislabeled training data is demonstrated in Sections 4 and 5 below.

2. Classification with mislabeled training data

2.1. Combined multinomial model

With this model, it is assumed that there exists a pair of probability vectors, $\mathbf{p}_k$ and $\mathbf{p}_l$, the ith 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, $\mathbf{y}$, a member of class k is given by [6],

$$f(\mathbf{x}_{kk}, \mathbf{x}_{lk}, \mathbf{x}_{ll}, \mathbf{x}_{kl}, \mathbf{y}|\mathbf{p}_k, \mathbf{p}_l, H_k; \alpha_k, \alpha_l) = N_{kk}!N_{lk}!N_{ll}!N_{kl}!N_{\mathrm{y}}! \prod_{i=1}^{M} \frac{p_{k,i}^{x_{kk,i}+x_{lk,i}+y_i} p_{l,i}^{x_{ll,i}+x_{kl,i}}}{x_{kk,i}!x_{lk,i}!x_{ll,i}!x_{kl,i}!y_i!}$$

$$\times \frac{N_k!}{N_{kk}!N_{kl}!}(\alpha_k)^{N_{kl}}(1-\alpha_k)^{N_{kk}} \frac{N_l!}{N_{ll}!N_{lk}!}(\alpha_l)^{N_{lk}}(1-\alpha_l)^{N_{ll}}, \tag{1}$$

where[3] $k, l \in \{class\ 1, class\ 2\}$, and $k \neq l$, H_k is the hypothesis defined as $\mathbf{p}_{\mathrm{y}} = \mathbf{p}_k$, M is the number of discrete symbols, $x_{kk,i}$ is the number of occurrences of the ith symbol in the correctly labeled training data for class k, $N_{kk}\{N_{kk} = \sum_{i=1}^{M} x_{kk,i}\}$ is the number of correctly labeled training data for class k, $x_{kl,i}$ is the number of occurrences of the ith symbol in the mislabeled training data for class k, appearing with probability α_k and belonging to class l, $N_{kl}\{N_{kl} = \sum_{i=1}^{M} x_{kl,i}\}$ is the number of mislabeled training data for class k, $x_{k,i} = x_{kk,i} + x_{kl,i}$ is the number of occurrences of the ith symbol in all training data for class k, $N_k\{N_k = N_{kk} + N_{kl} = \sum_{i=1}^{M} x_{k,i}\}$ is the total number of training data for class k, y_i is the number of occurrences of the ith symbol in the test data, $N_{\mathrm{y}}\{N_{\mathrm{y}} = \sum_{i=1}^{M} y_i\}$ is the total number of test data, $p_{k,i}\{\sum_{i=1}^{M} p_{k,i} = 1\}$ is the probability of the ith symbol for class k.

2.2. Combined Bayes Test (CBT)

An important aspect of the CBT is that rather than assume $\mathbf{p}_k$ and $\mathbf{p}_l$ are simply unknown parameters to be estimated (and use a combined generalized likelihood ratio test [7]), the approach here is to give them prior distributions [8]. We assume nothing is known a priori about the probability vectors and so we use an "ignorance" prior. One version of prior ignorance is provided by the uniform Dirichlet given by [2][4]

$$f(\mathbf{p}_k) = (M-1)!\mathscr{I}_{\{\sum_{i=1}^{M} p_{k,i} = 1\}}, \tag{2}$$

where $\mathscr{I}_{\{x\}}$ is the indicator function.

The first step in developing the CBT for mislabeled training data is to apply the Dirichlet to the formula of (1) under each class k and l, and then integrate with respect

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

[4] The uniform Dirichlet results when the parameters of this distribution are set to unity [1]. Note, it was found in [9] in relation to universal encoding that a better prior to use, given unknown true statistics, is the Dirichlet with its parameters set to one-half (see also [10]). However, this distribution does not treat each symbol probability equally so that the uniform Dirichlet is used here to represent complete ignorance about the underlying symbol probabilities of each class. Also, use of the uniform Dirichlet for this application is consistent with previous work (e.g. [2,11]).

to $\mathbf{p}_k$ and $\mathbf{p}_l$ over the *positive unit-hyperplane* resulting in

$$f(\mathbf{x}_{kk},\mathbf{x}_{lk},\mathbf{x}_{ll},\mathbf{x}_{kl},\mathbf{y}|H_k;\alpha_k,\alpha_l)=\frac{((M-1)!)^2N_{kk}!N_{lk}!N_{ll}!N_{kl}!N_{\mathbf{y}}!}{(N_{kk}+N_{lk}+N_{\mathbf{y}}+M-1)!(N_{ll}+N_{kl}+M-1)!}$$

$$\times\prod_{i=1}^{M}\frac{(x_{kk,i}+x_{lk,i}+y_i)!(x_{ll,i}+x_{kl,i})!}{x_{kk,i}x_{lk,i}!x_{ll,i}!x_{kl,i}!y_i!}$$

$$\times\frac{N_k!}{N_{kk}!N_{kl}!}(\alpha_k)^{N_{kl}}(1-\alpha_k)^{N_{kk}}\frac{N_l!}{N_{ll}!N_{lk}!}(\alpha_l)^{N_{lk}}(1-\alpha_l)^{N_{ll}}. \tag{3}$$

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

$$f(\mathbf{x}_k,\mathbf{x}_l,\mathbf{y}|H_k;\alpha_k,\alpha_l)=\sum_{\mathbf{x}_{kl}=\vec{0}}^{\mathbf{x}_k}\sum_{\mathbf{x}_{lk}=\vec{0}}^{\mathbf{x}_l}f(\mathbf{x}_k-\mathbf{x}_{kl},\mathbf{x}_{lk},\mathbf{x}_l-\mathbf{x}_{lk},\mathbf{x}_{kl},\mathbf{y}|H_k;\alpha_k,\alpha_l). \tag{4}$$

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

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

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

2.3. Probability of error

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

$$P(e)=P(H_k)P(z_k\leqslant\tau z_l|H_k)+P(H_l)P(z_k>\tau z_l|H_l). \tag{6}$$

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 $P(H_k)$, the first term of (6) is given by

$$P(z_k\leqslant\tau z_l|H_k)=\sum_{\mathbf{y}}\sum_{\mathbf{x}_k}\sum_{\mathbf{x}_l}\mathscr{I}_{\{z_k\leqslant\tau z_l\}}f(\mathbf{x}_k,\mathbf{x}_l,\mathbf{y}|H_k;\alpha_k,\alpha_l), \tag{7}$$

where $f(\mathbf{x}_k,\mathbf{x}_l,\mathbf{y}|H_k;\alpha_k,\alpha_l)$ was defined in formula (4) above.

3. Results using the CBT

Fig. 1 contains an average probability of error curve, $P(e)$ (plotted as a function of the number of discrete symbols, M), for the CBT given the true mislabeling probabilities are given by, respectively, $\alpha_k=\alpha_l=0.0$, 0.05, 0.15, 0.25, 0.35, and 0.45. Notice that results are based on ten samples of training data for each class, and one observation of

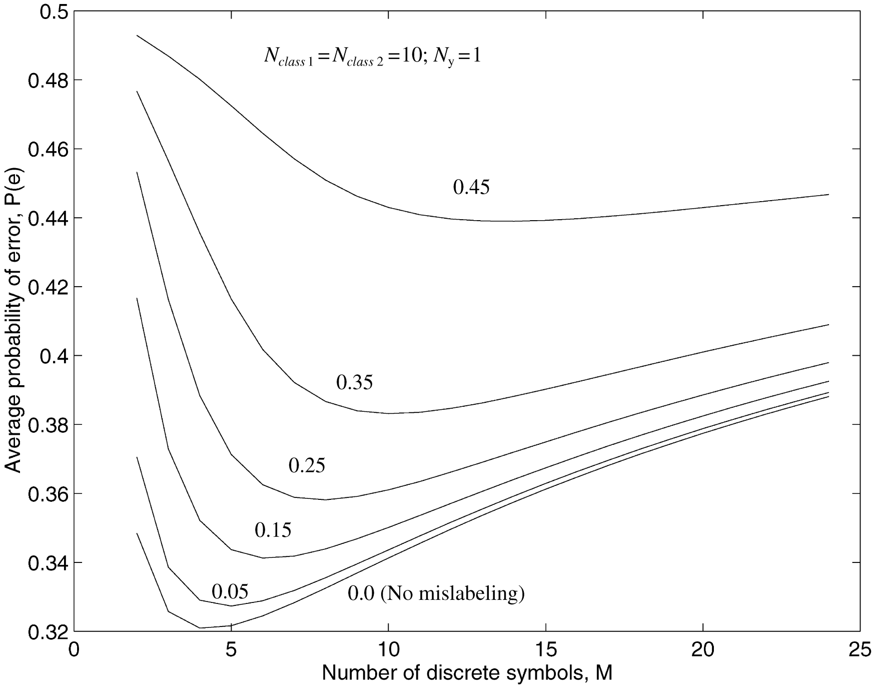

Fig. 1. Performance of the CBT with various mislabeling probabilities.

test data. Additionally, the decision threshold $\tau = 1$. In all cases of Fig. 1, observe that $P(e)$ starts out decreasing with increasing M and is minimum at a point referred to as M^*,[5] and 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 [2,1]. However, as the mislabeling probabilities are fixed with larger values overall performance begins to degrade in that $P(e)$ increases. Also, it can be seen that accompanying this degradation in performance is an increase in the quantity M^*. That is, for the mislabeling probabilities in Fig. 1 given by 0.0, 0.05, 0.15, 0.25, 0.35, and 0.45, M^* has the respective values of 4, 5, 6, 8, 10, and 12.[6] Intuitively, an increase in the mislabeling probabilities causes the classes to become similar, so that for best classification performance more information (i.e., a finer quantization) is required.

With these findings, it was found that if the mislabeling probabilities assumed for the training data (i.e., for the z_k and z_l of formula (6)) take on any values within the range, $0 \leqslant \alpha_k = \alpha_l < 0.5$, identical results are produced for all cases in Fig. 1. In other

[5] For example, when there is no mislabeling in the training data $M^* = 4$.

[6] Notice that even when $\alpha_k = \alpha_l = 0.45$, a best quantization fineness exists.

words, when testing it does not matter if the CBT of formula (5) contains the true mislabeling probabilities as long as they are not assumed to be 0.5 or higher (which would indicate a CBT that is testing as if most of the training data of each class is more likely to belong the other class). This aspect of the CBT's performance is attributed to the averaging which occurs in formula (4) over all possible orderings of the mislabeled training data, coupled with placement of the uniform (i.e., Dirichlet) prior on the symbol probabilities.

In Fig. 2, results from Fig. 1 are repeated (i.e., $N_y = 1$) for the mislabeling probabilities given by $\alpha_k = \alpha_l = 0.0$, 0.05, 0.15, and 0.25. Additionally, notice also shown (lower curves with an $*$) for the same mislabeling probabilities is the case involving two observations of test data (i.e., $N_y = 2$). It can be seen in this figure that when $N_y = 2$ performance improves for a given mislabeling probability as $P(e)$ is reduced (for more on this with correctly labeled training data, see [1]). With this, observe that as compared to the $N_y = 1$ case, increasing the number of test observations to $N_y = 2$ causes all associated values of M^*, where performance is best, to increase by one. That is, for the mislabeling probabilities given by 0.0, 0.05, 0.15, and 0.25, and when $N_y = 2$, M^* has the respective values of 5, 6, 7, and 9. Accompanying this increase in M^* is the associated increase in $P(e)$. However, it is apparent that the increase in $P(e)$ is relatively worse when $N_y = 2$ (i.e., the $P(e)$ curves are further apart). The reason this occurs is

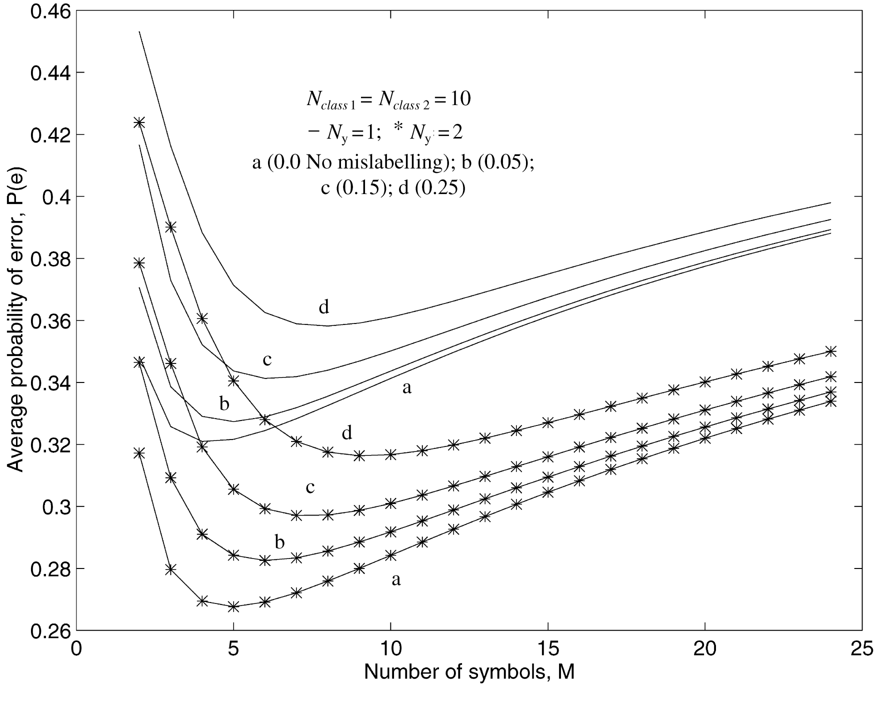

Fig. 2. Performance of the CBT with more test observations.

that although a greater number of test observations improves the estimation capability of the CBT, there also is more of a likelihood that a test observation will be of the same value as a mislabeled training datum.

4. Applying the BDRA to mislabeled training data

In this section a method of data reduction (i.e, feature selection) called the BDRA is applied to mislabeled training data. As used here, the BDRA demonstrates the degree to which the negative effect of mislabeling can be diminished by employing a suboptimal algorithm to train on the data. The BDRA is based on the CBT and its performance was previously described in [5] at classifying, and reducing, feature vectors containing binary valued components. In that work, performance of the BDRA was shown to be superior to a neural network. Here, the BDRA is applied to feature vectors consisting of six binary valued features (i.e., $M = 64$), which are also mislabeled in the training data of each class according to the probabilities shown in Fig. 1.

The BDRA works by reducing the quantization fineness, M, of the training data to a level which minimizes the average conditional probability of error, $P(e|X)$ (the variable X represents the entire collection of training data from all classes). The formula for $P(e|X)$ is a fundamental component of the BDRA, and similar to formula (6) above it is given by[7]

$$P(e|X) = P(e|\mathbf{x}_k, \mathbf{x}_l) = \sum_{\mathbf{y}} \sum_{\mathbf{x}_k, \mathbf{x}_l} P(H_k)\mathscr{I}_{\{z_k \leqslant z_l\}} f(\mathbf{y}|\mathbf{x}_k, H_k) + P(H_l)\mathscr{I}_{\{z_k > z_l\}} f(\mathbf{y}|\mathbf{x}_l, H_l), \tag{8}$$

where for the cases considered involving only one observation of test data (i.e., $N_{\mathbf{y}} = 1$) $z_k = f(\mathbf{y}|\mathbf{x}_k, H_k) = (x_{k,i} + 1)/(N_k + M)$.

For binary valued feature vectors the BDRA is then implemented with the formula (8) and the following iterative steps.

(1) Using the initial training data with quantization fineness M (i.e., $M = 2^{N_f}$, where N_f is the number of features), use formula (8) to compute $P(e|X; M)$.

(2) Beginning with the first feature (selection is arbitrary), remove this feature from each class by summing (i.e., merging) the numbers of occurrences of those discrete symbols that correspond to its removal (i.e., for all classes simultaneously merge those quantized symbols containing a binary zero for the reduced feature with those containing a binary one).

(3) Use the newly merged training data (X') and the new quantization fineness ($M' = 2^{N_f - 1}$), and compute $P(e|X'; M')$.

(4) Repeat items two and three for all N_f features.

[7] Note, the notational descriptions of formula (1) apply to formula (8).

(5) From item four select the minimum of all computed $P(e|X';M')$ (ties are broken arbitrarily), and choose this as the new training data configuration for each class (this corresponds to permanently removing the associated feature).
(6) Repeat items two through five until the probability of error does not decrease any further, or $M' = 2$, and this defines the new quantization fineness.

Note, the BDRA is a "greedy" algorithm in that it chooses a best training data configuration at each iteration (see step (5) above) in the process of determining a best quantization fineness. A slightly better algorithm is to do a global search over all possible merges and corresponding training data configurations. However, a simulation study involving hundreds of independent trials revealed that only a few percent of the time did the "greedy" algorithm shown here produce results different than a global algorithm. Additionally, the overall average probability of error for the two algorithms differed by only an insignificant amount.

5. Results using the BDRA

In Fig. 3, performance of the BDRA is shown when the training data are mislabeled according to the probabilities specified in Fig. 1. The results in this figure are based on an average of one hundred independent trials. At each trial, a set of $M = 64$ true symbol probabilities, consisting of six independent bit probability pairs, were generated for both classes using Gaussian mixture distributions (see footnote 9 below). Additionally, results appear for two training data sizes of 25 and 100 samples, which were randomly generated at each trial from the true symbol probabilities.

Observe in Fig. 3 that $P(e)$ appears as a function of the mislabeling probabilities both with and without applying the BDRA[8] to the training data, and for the optimal test.[9] It can be seen in Fig. 3 that in all cases performance degrades with the severity of the mislabeling probabilities, which is analogous to the results in Fig. 1. However, for both training data sizes the BDRA is successful at improving overall classification performance (relatively less improvement, that is, less data reduction, occurs with more training data as the probability estimates are more accurate). But, in all cases it appears that the improvement diminishes rapidly as the mislabeling probabilities approach 0.5. On the other hand, with one hundred samples of training data and mislabeling probabilities of less than 0.1, performance is relatively close to optimal.

Fig. 4 shows the average number of features reduced from the training data of each class by the BDRA as a function of the true mislabeling probabilities. In this figure,

[8] Results shown for the BDRA were obtained by using its trained test statistic with the actual symbol probabilities.

[9] Optimal results are based on the test which knows all true symbol probabilities, and there is no mislabeling of the training data. Also, notice that the optimal error probabilities are relatively constant at 0.075, and this is due to them having been constrained to be $\geqslant 0.05$ and $\leqslant 0.1$. Further, a constraint on the optimal error probabilities was possible because the true symbol probabilities were created with Gaussian mixture distributions (for more on this see [5]).

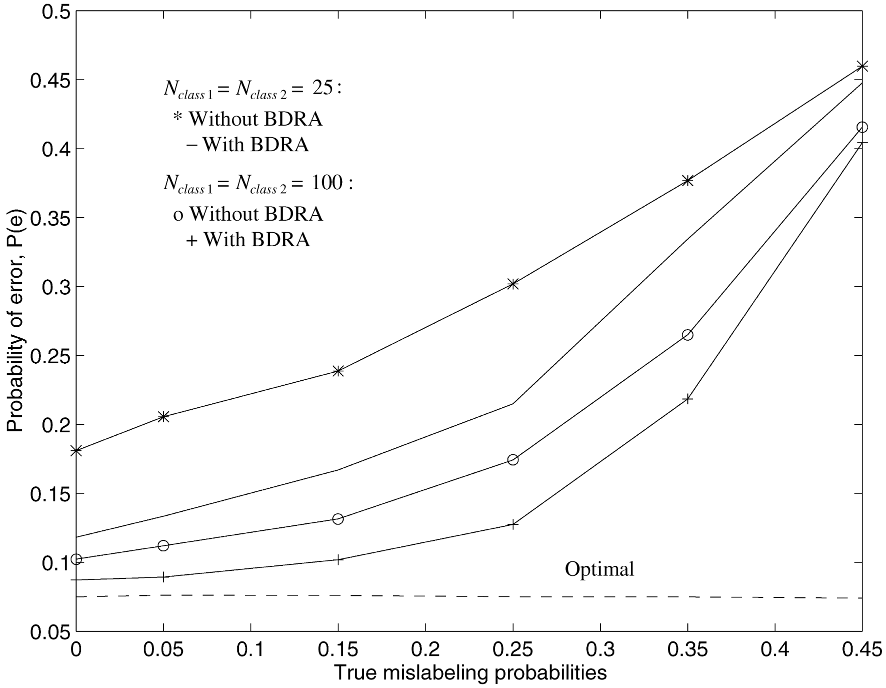

Fig. 3. Performance of the BDRA with the mislabeling probabilities of Fig. 1.

results appear for those training data sizes shown in Fig. 3, and are based on an average of one hundred independent trials. As expected from Fig. 1, overall it can be seen that the average number of features reduced (eliminated) from the training data of each class becomes less as the mislabeling probabilities increase (the increase in the number of features associated with a mislabeling probability of 0.05 is attributed to using only one hundred independent trials to obtain the results). Also, consistent with the results of Fig. 3, the BDRA appears to reduce a larger number of features when there is less training data, and this caused by relatively more uncertainty associated with the symbol probability estimates.

6. Conclusion

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. In general, it was shown that as the mislabeling probabilities increase, both the average probability of error and the optimum quantization fineness, M^*, increase. Additionally, it was found that $P(e)$ can be reduced if the number of test observations is increased to $N_y = 2$. However, the relative performance degradation

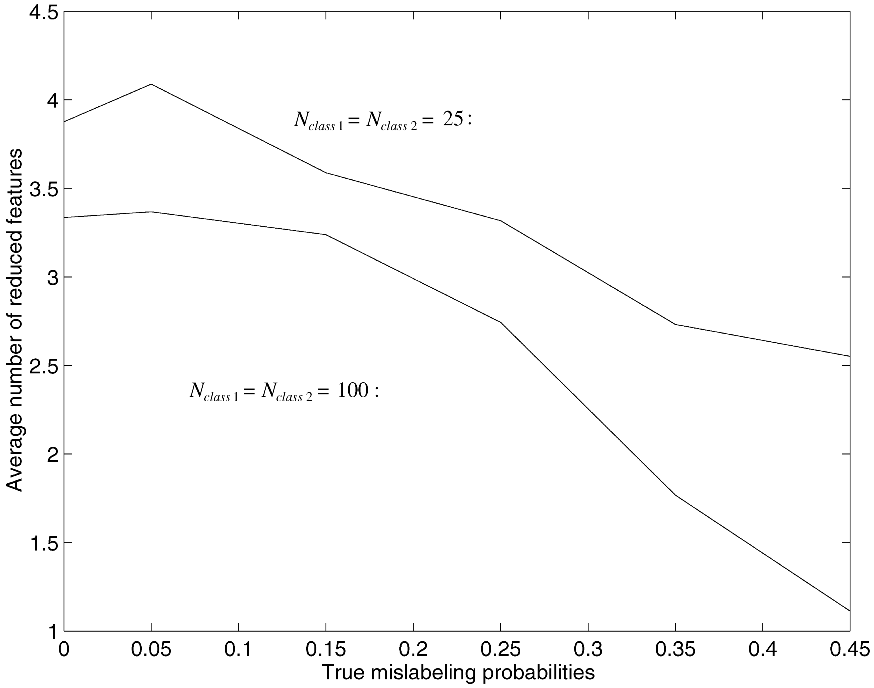

Fig. 4. Average number of features reduced from the training data of each class.

with mislabeling present is relatively larger then it is when $N_y = 1$, and this is due to an increased likelihood of the test data matching the mislabeled training data. Further, a method of data reduction (i.e., feature selection) called the BDRA was applied to training data corrupted by mislabeling, and results indicate that classification performance can be improved if the mislabeling probabilities are not too severe. But, the relative amount of improvement decreases with training data size as the symbol probability estimates become more accurate.

References

[1] R. Lynch, P. Willett, Discrete symbol quantity and the minimum probability of error for a combined information classification test, Proceedings of the 35th Annual Allerton Conference on Communication, Control, and Computing, September 1997.

[2] G.F. Hughes, On the mean accuracy of statistical pattern recognizers, IEEE Trans. Inform. Theory 14 (1) (1968) 55–63.

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

[4] R.S. Lynch Jr., P.K. Willett, Testing the statistical similarity of discrete observations using Dirichlet priors, Proceedings of the IEEE International Symposium on Information Theory, August 1998.

[5] R.S. Lynch Jr., P.K. Willett, Bayesian classification and data driven quantization using Dirichlet priors, Proceedings of the 32nd Annual Conference on Information Sciences and Systems, March 1998.

[6] R.S. Lynch Jr., P.K. Willett, Classification using Dirichlet priors when the training data are mislabeled, Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, March 1999.

[7] N. Merhav, Y. Ephraim, A Bayesian classification approach with application to speech recognition, IEEE Trans. Acoust. Speech Signal Process. 39 (10) (1991) 2157–2166.

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

[9] R.E. Krichevsky, V.K. Trofimov, The performance of universal encoding, IEEE Trans. Inform. Theory 27 (2) (1981) 199–207.

[10] L.L. Campbell, Averaging Entropy, IEEE Trans. Inform. Theory 41 (1) (1995) 338–339.

[11] J.M. Van Campenhout, On the peaking of the Hughes mean recognition accuracy: the resolution of an apparent paradox, IEEE Trans. Systems Man Cybernet. 8 (5) (1978) 390–395.