A Feature Selection Based on Minimum Upper Bound of Bayes Error

Guorong Xuan, Zhenping Zhang, Peiqi Chai
Dept. of Computer Science
Tongji University
Shanghai, P. R. China
grxuan@public1.sta.net.cn

Yun Q. Shi, Dongdong Fu
Dept. of Electrical & Computer Engineering
New Jersey Institute of Technology
Newark, NJ 07029, USA
shi@njit.edu

***Abstract*—This paper[1] presents a novel feature selection scheme based on the upper bound of Bayes error under normal distribution for the multi-class dimension reduction problem. The upper bound of Bayes error in the multi-class problem is represented by the sum of the upper bound of Bayes error of every two-class pair. In order to obtain an accurate solution of the feature selection transform matrix in term of the minimum upper bound of Bayes error, a recursive algorithm based on gradient method is developed. The principal component analysis (PCA) is used as a pre-processing to reduce the intractably heavy computation burden of the recursive algorithm. The superior experimental results on the handwritten digit recognition with the MNIST database demonstrate the effectiveness of our proposed method.**

Keywords—feature selection based on minimum error bound (FME); fast feature selection based on minimum error bound (FFME); PCA pre-processing; recursive algorithm; handwritten digit recognition

I. INTRODUCTION

The feature selection is one of the basic topics in pattern recognition, which plays an important role in multimedia signal classification. Generally speaking, the feature selection problem is to map the original high dimensional feature space into an optimum lower one based on certain criterion. Theoretically, the minimum Bayes error is the best criterion to evaluate feature effectiveness for classification. However, it is too complicated for the multi-class dimension reduction problem [1].

In practice, Mahalanobis distance feature selection may be a good method. The Linear Discriminant Analysis (LDA) [2] or so called Fisher Discriminant Analysis (FDA) [3] can be considered as the approximate schemes for Mahalanobis distance feature selection. But these feature selections have no direct relation with the Bayes error.

On the contrary, the Bhattacharyya distance can be used as an optimization criterion for feature selection which gives the upper bound of the Bayes error [1]. To the best of our knowledge, only approximate solutions of upper bound of Bayes error of normal distribution can be obtained by Bhattacharyya distance in the last decades except paper [4].

In this paper we achieve the accurate solution by a recursive algorithm. Now we prefer to use the term Feature Selection "based on Minimum Upper Bound of Bayes Error "in this paper instead of "based on Bhattacharyya distance", because the classification error can be shown correctly under multi-classes, see formula (7) in next section. We name this method as *Feature Selection based on Minimum Upper Bound of Bayes Error* (FME). To reduce the computational complexity of FME, a PCA pre-processing is used to speed up the recursive algorithm without sacrificing the high classification performance. We name this scheme as *Fast Feature Selection based on Minimum Upper Bound of Bayes Error* (FFME).

The paper is organized as follows. A multi-class feature selection algorithm based on the upper bound of Bayes error is formulated in Section II. A recursive algorithm is derived in section III. In order to improve the processing speed, a feature selection using PCA as pre-processing is also presented in Section IV. The experimental results of handwritten digit recognition with MNIST database [5] are presented in Section V and the conclusions are drawn in Section VI.

II. THE UPPER BOUND OF BAYES ERROR PROBABILITY OF MULTI-CLASS PROBLEM

A. *Two-class problem:*

The upper bound of Bayes error probability ε_{ij} and Bhattacharyya distance $\mu_{ij}(1/2)$ for two classes *i* and *j* with normal distribution can be obtained by:

$$\varepsilon_{ij} = \frac{1}{2}\exp\left[-\mu_{ij}(1/2)\right] \tag{1}$$

$$\mu_{ij}(1/2) = \frac{1}{8} tr\left[\tilde{W}_{ij}^{-1}\tilde{B}_{ij}\right] + \frac{1}{2}\ln\frac{\left|\tilde{W}_{ij}\right|}{\left|\tilde{W}_i\right|^{1/2}\left|\tilde{W}_j\right|^{1/2}} \tag{2}$$

$$\tilde{W}_i = \phi^t W_i \phi = \phi^t E\left[\left(X - M_i\right)\left(X - M_i\right)^t\right]\phi \tag{3}$$

$$\tilde{W}_j = \phi^t W_j \phi = \phi^t E\left[\left(X - M_j\right)\left(X - M_j\right)^t\right]\phi \tag{4}$$

$$\tilde{W}_{ij} = \phi^t W_{ij} \phi = \phi^t \left(\frac{W_i + W_j}{2}\right)\phi \tag{5}$$

[1] This research is supported partly by National Natural Science Foundation of China (NSFC) on the project (90304017), and by New Jersey Commission of Science and Technology via New Jersey Center of Wireless Networking and Internet Security (NJWINS).

$$\tilde{B}_{ij} = \phi^t B_{ij} \phi = \phi^t \left(M_i - M_j\right)\left(M_i - M_j\right)^t \phi \quad (6)$$

where ϕ is the feature selection transform matrix which maps the original feature space into a reduced feature space. M_i and M_j are the means of class *i* and *j*, respectively. W_i and W_j are the within-class scatter matrices for class *i* and *j* in the original *n*-dimensional feature space, respectively. $\tilde{W}_i$ and $\tilde{W}_j$ are the within-class scatter matrices for class *i* and *j* in the reduced feature space, respectively. $\tilde{W}_{ij}$ and $\tilde{B}_{ij}$ denote the average within-class scatter matrix and between-scatter matrix for class *i* and *j* in the reduced feature space.

B. Multi-class problem:

If there exist *L* normal distributed classes which have equal prior probabilities, the upper bound of Bayes error probability, ε , can be expressed as :

$$\varepsilon = \sum_{i>j}^{L}\sum_{j=1}^{L} \varepsilon_{ij} = \frac{1}{L}\sum_{i>j}^{L}\sum_{j=1}^{L} \exp\left[-\mu_{ij}\left(1/2\right)\right] \quad (7)$$

where ε_{ij} is the upper bound of Bayes error probability of every two-class pair as defined in last subsection.

III. Recursive Algorithm

The purpose of the proposed scheme is to find a $m \times n$ dimensional linear transformation matrix ϕ which maps the *n*-dimensional original feature space to the *m*-dimensional reduced feature space by minimizing the upper bound of Bayes error probability, where $m < n$.

Therefore the linear feature selection can be expressed as:

$$\phi^t X = Y \quad (8)$$

where *X* is the original n-dimensional vector and *Y* is the reduced m-dimensional vector.

In order to minimize the upper bound of Bayes error probability ε , we can obtain the optimal solution by:

$$\frac{\partial \varepsilon}{\partial \phi} = 0 \quad (9)$$

From the formula (see formula A16 and formula A27 at pp. 566-568 in [2]), we have:

$$\frac{\partial \varepsilon}{\partial \phi} = -\frac{1}{L}\sum_{i>j}^{L}\sum_{j=1}^{L} \exp\left[-\mu_{ij}\left(1/2\right)\right]\frac{\partial \mu_{ij}\left(1/2\right)}{\partial \phi} \quad (10)$$

where

$$\frac{\partial \mu_{ij}\left(1/2\right)}{\partial \phi} = \left\{W_{ij}\left[\left(C_{ij} + 4I\right)\phi - \phi\,\tilde{C}_{ij}\right]\tilde{W}_{ij}^{-1} - 2W_i\phi\tilde{W}_i^{-1} - -2W_j\phi\tilde{W}_j^{-1}\right\}\Big/4 \quad (11)$$

and $$C_{ij} = W_{ij}^{-1}B_{ij}, \tilde{C} = \tilde{W}_{ij}\tilde{B}_{ij} \quad (12)$$

In order to solve the equation (9), we propose a recursive algorithm based on gradient method as $\phi_{r+1} = f(\phi_r)$. Therefore, the procedure of the recursive algorithm to solve (9) with the step length λ for the r^{th} recursive step can be shown as:

$$\phi_{r+1} = \phi_r - \lambda\frac{\partial \varepsilon}{\partial \phi} \quad (13)$$

The recursive algorithm continues until

$$\| \phi_{r+1} - \phi_r \| < \delta \quad (14)$$

where δ is a small threshold value.

IV. PCA Pre-Processing

Although the recursive algorithm can get the accurate solution of the feature selection matrix, its computational burden is heavy and time consumption is high, especially when the dimension of the original feature space is large.

To overcome this problem, we further propose a *Fast Feature Selection Based on Minimum Upper Bound of Bayes Error (FFME)* in which PCA is applied as a pre-processing method before the recursive algorithm is used.

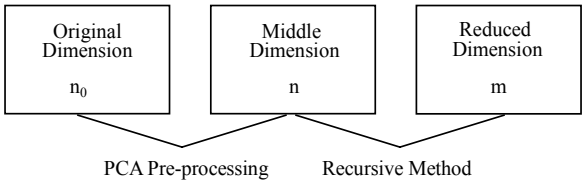

Fig. 1. An example of the proposed FFME method

Fig. 1 illustrates the block diagram of the proposed scheme. In this example, a PCA pre-processing is carried out to reduce the original dimension n_0 to a middle dimension n. Then, the recursive algorithm is used to do the rest of the dimension reduction, namely, from n to m.

Traditionally, the PCA feature selection finds the optimal feature selection transform matrix by maximizing the criterion J_{PCA} , namely:

$$A_P = \arg\max_A \left(J_{PCA}\right) \quad (15)$$

where

$$J_{PCA} = tr\left[A^t\left(W + B\right)A\right] \quad (16)$$

where *W* and *B* are the covariances of within-class scatter matrices and between-class scatter matrices respectively.

PCA feature selection is well known for its effective representation of the original pattern and not for classification. But in some cases, PCA feature selection is also effective to classification problem. Our analysis and experimental results show that when the between-class scattering is dominant over the within-class scattering, PCA feature selection performances very good for pre-processing.

Some reasons make PCA preprocessing effective. The first, the ranks of the scatter matrixes of the samples are usually not full. We can't use the recursive algorithm directly. The scatter matrixes might be full after PCA pre-processing used. The second, PCA pre-processing can get rid of the noises effectively, sometimes can enhance the robustness of the algorithms. Finally, our extensive experiments show that when the between-class scattering of the original patterns are dominant over the within-class scattering, the data after PCA

pre-processing can still preserve enough classification information. Therefore, PCA pre-processing can improve the computational speed dramatically without compromising the classification performance of the feature selection algorithm if we choose the middle dimension appropriately.

V. EXPERIMENTS

A. *Experiment 1*

The data used in this experiment are adopted from [1]. It is an 8-dimensional 4-class case ($n_0 = 8$).

In this experiment, we investigate the performance of PCA as a function of the between-class distance. To do this, we simply multiply every mean vector by a scaling coefficient F while keep all within-class scatter matrixes fixed. In this way, we can adjust the between-class distance conveniently. For example, the between-class distance is unchanged when F=1. The between-class distance will increase to 4 times when F=2. And so on. The bigger F is, the larger between-class distance is.

Fig. 2 and 3 show the results for different resultant dimension cases. The black vertical lines at the center of the figures indicate F=1. We find out, for both cases, the performance of PCA becomes better and better while F increases. When F>5，the performance difference between PCA and the proposed algorithm is really trivial. It means that when the between-class distances in this kind of data are large, it is reasonable for us to utilize PCA as pre-processing without sacrificing the performance.

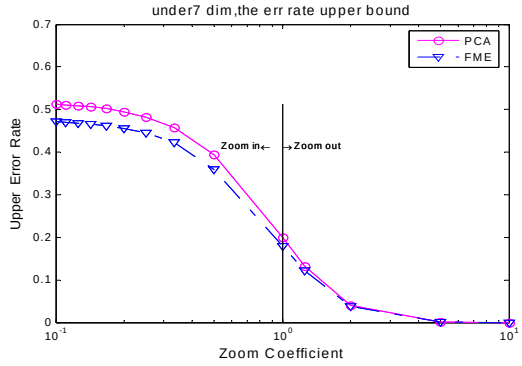

Fig. 2. Performance comparison between PCA and FME for different between-class distances (reduced from 8D to 7D).

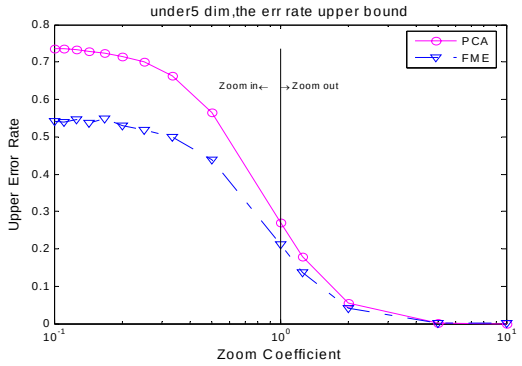

Fig. 3. Performance comparison between PCA and FME for different between-class distances (reduced from 8D to 5D)

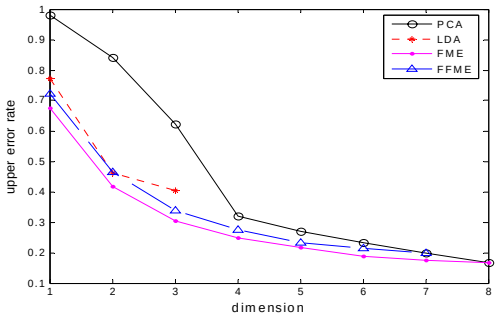

Fig. 4. Comparison of different feature selection schemes.

Fig. 5. The samples of the handwritten-digits in the MNIST database.

In Fig. 4, we show the comparison of different feature selection schemes, the data are also adopted from [1]. We compare our proposed algorithms with other feature selection schemes for different resultant space dimensions. The horizontal axis in the figure stands for resultant space dimension and the vertical axis is the upper bound of Bayes error. As we can see, the error rates of the proposed algorithms are lower than those of other schemes. We also observe that the difference in this kind of data between the algorithms with (FFME) and without (FME) PCA pre-processing is very small.

B. *Experiment 2 (FFME performance evaluated by the upper bound of Bayes error)*

In this and next experiment, the handwritten digit recognition is applied on the MNIST database [5]. There are totally 60,000 digit images from 0 to 9 in MNIST. The size of each digit image is 28×28 (n_0 = 28×28 = 784). Therefore, this experiment is a 784-dimensional 10-class case (n_0=784). The dimension of the resultant subspace is set to 30 for classification (m=30). Fig. 5 illustrates the original expected vectors of these handwritten digits

The middle dimensions by PCA pre-processing are set to 500, 400, 300, 200, 100, 50, 40 and 30, respectively. The highest middle dimension is 500, because the rank of the scatter matrixes is 500. The recursive step size we used in the experiment is $\lambda = 0.01$ in (13). Actually, In FFME, we reduce from n=500 to a middle dimension by FME and further reduce from a middle dimension to n=30 by PCA.

The experiment is running on a personal computer with a Pentium IV 1.8G CPU and 256M Rambus memory. The experimental results are shown in Table I. As we can see, if we choose appropriate middle dimension, the proposed FFME scheme gives an excellent tradeoff between the classification performance and processing speed. For example, when the middle dimension n is set to 100, the upper bound of error rate achieved by the proposed algorithm with PCA pre-processing is only a little bit larger than that of the pure recursive algorithm ((1.05-0.98) / 0.98 = 7%), but the time consumed by

the pure algorithm is about 202 times more than that consumed by the algorithm with PCA pre-processing. Therefore, the proposed algorithm with PCA pre-processing gives an effective tradeoff between the upper bound of Bayes error rate and the computational complexity, which makes it more practical.

TABLE I. "COMPUTATIONAL RESULTS"(THEORETICAL UPPER BOUND OF BAYES ERROR)

Middle Dimension (n)	Upper bound Error (%)	Time Consumption (Minutes)
500	0.98	17×60
400	0.98	69
300	0.98	40
200	1.00	19
100	1.05	5
50	1.36	2
40	1.54	1
30	1.77	~ 0

C. *Experiment 3(FFME performance evaluated by practical results)*

TABLE II.METHODS COMPARISION AND TIME CONSUMPTION

Methods / Dimension or Time	A (FME)	B (FFME)	C (LDA)	D (PCA)
Original 500D, Middle 100D	Recursive	PCA Pre-processing	no Recursive	no Recursive
Middle:100D, Resultant:9D	Recursive	Recursive	no Recursive	no Recursive
Time Consumption	~8 Hours	~5 minutes	~0.5 minutes	~0.2 minutes

TABLE III. "STATISTICAL RESULTS" (ACTUAL CORRECTION RATE OF CLASSIFICATION %)

Methods / Patterns	A (FME)	B (FFME)	C (LDA)	D (PCA)
0	97.02	96.34	95.70	94.10
1	95.48	94.64	93.96	96.14
0	97.02	96.34	95.70	94.10
1	95.48	94.64	93.96	96.14
2	92.98	93.10	88.94	87.44
3	89.32	89.40	84.32	83.68
4	93.22	93.86	89.78	83.02
5	91.40	91.20	83.34	87.18
6	94.76	94.46	93.50	93.72
7	92.62	92.76	95.56	89.06
8	88.50	89.06	82.84	82.12
9	90.96	90.14	86.48	73.36
Average	92.63	92.50	88.94	86.98

In experiments 1-2, the upper bound of the classification error is used as the criterion for the evaluation of the performance of the classification after feature selection. The results in these experiments show that the proposed scheme outperforms PCA and LDA in term of upper bound of the classification error. However, the relation between the Bhattacharyya distance and the upper bound of Bayes error is derived assuming that the samples are with normal distributions. In practice, this assumption can not be always satisfied. To evaluate the performance of the proposed algorithm for these cases, we conduct the following classification experiment for the samples in MNIST database.

In this experiment, we select 5000 samples in each of the 10 digit-patterns randomly. Four feature selection schemes are applied for classification: *A*. FME; *B*. FFME (FME with PCA pre-processing); *C*. LDA; *D*. PCA only. All the methods reduce the dimension of the original sample space into a 9-dimentional space first (10 classes, there are at most 9-dimensional features are preserved in LDA). Then, a Bayes classifier is utilized for classification. The results are shown in table II and III. The experiment is running on a personal computer with a Pentium IV 1.8G CPU and 256M Rambus memory.

Although the samples in MNIST database do not completely satisfy the normal distribution assumption, the results in table II and III show that the proposed feature selection schemes (FME and FFME) outperform the LDA and PCA in term of true correction rate of classification. At the same time, the correction rate of the recursive algorithm with PCA pre-processing (FFME) is a little bit lower (~0.13% dropped) than that of the pure recursive algorithm (FME), but the time consumption decreases significantly. Therefore, the PCA pre-processing makes the proposed algorithm more practicable.

VI. CONCLUSIONS

1. This paper proposed a feature selection algorithm based on the upper bound of Bayes error for multi-class classification problems. The upper bound of Bayes error probability in the multi-class problem is represented by the sum of the classification error probabilities of every two-class pair while the latter is obtain by the two-class Bhattacharyya distance.

2. This paper presented a recursive algorithm to find the accurate solution for the feature selection transform matrix based on upper bound of Bayes error probability.

3. This paper proposed a fast feature selection based on minimum error bound (FFME) in which PCA is applied as a pre-processing method to reduce the original dimension to a middle dimension before the recursive algorithm is used.

4. The experimental results have shown that FFME speed up FME significantly without sacrificing the performance of classification.

REFERENCES

[1] K. Fukunaga, "Introduction to Statistical Pattern Recognition", 2nd Edition, Academic Press, Inc., Boston, 1990.

[2] K. Torkkola. "Discriminative Features for Text Document Classification". Pattern Analysis and Applications, 6(4), February 2004, pp.301–308.

[3] T. Hastie, R. Tibshirani, and A. Buja, "Flexible discriminant analysis by optimal scoring," Journal of the American Statistical Association, vol. 89, 1994, pp. 1255-1270.

[4] G. Xuan, P. Chai, M. Wu, "Bhattacharyya Distance Feature Selection", 13th International Conference on Pattern Recognition, Aug. 25-29, 1996, Vienna, Austria., pp.195-199.

[5] MNIST database for Hand written digits recognition, Yann LeCun, NEC Research Institute, (http://yann.lecun.com/exdb/mnist/)