Feature Selection based on the Bhattacharyya Distance

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

Yun Q. Shi, Dongdong Fu
Dept. of Electrical and Computer Engineering
New Jersey Institute of Technology
Newark, NJ 07102, USA
shi@njit.edu

Abstract

This paper presents a Bhattacharyya distance based feature selection method, which utilizes a recursive algorithm to obtain the optimal dimension reduction matrix in terms of the minimum upper bound of classification error under normal distribution for multi-class classification problem. In our scheme, PCA is incorporated 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 and the steganalysis applications have demonstrated the effectiveness of our proposed method.

1. Introduction

The feature selection problem is to map the original high dimensional feature space into an optimum low one based on certain criterion [1] [2] [3]. Theoretically, the Bayes error is the best criterion to evaluate feature effectiveness for classification. However, it is very difficult to minimize Bayes error rate in an analytical way. Fortunately, the upper bound of the classification error can be obtained by the Bhattacharyya distance (denoted BD for short in this paper). One of feature selection schemes based on Bhattacharyya distance can be found in [4], but there is no ideal result achieved in that paper.

In this paper, based on the results achieved in [5], we further propose a fast Bhattacharyya distance feature selection scheme, which combines the recursive algorithm with a PCA pre-processing to mitigate the computational complexity.

The rest of the paper is organized as follows.

A recursive algorithm is derived in Section 2. A feature selection scheme combining PCA pre-processing and recursive algorithm is presented in Section 3. The experimental results of handwritten digit recognition with MNIST database [6] and steganalysis applications are presented in Section 4 and the conclusions are drawn in Section 5.

2. Bhattacharyya distance

The aim of feature selection is to find an $m \times n$ ($m < n$) 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 error probability:.

$$(\boldsymbol{Y})_m = (\phi^{\mathrm{t}})_{m \times n} (\boldsymbol{X})_n \qquad (1)$$

This procedure can be expressed in terms of Bayes classification error probability $\boldsymbol{P_e}$:

$$\phi = \arg \min_{\phi} (\boldsymbol{P_e}) \qquad (2)$$

Actually, Equation (2) is very hard to be solved in an analytical way. Fortunately, we can substitute the minimum upper bound of the classification error for the minimum Bayes error rate in that the former is relative simpler than the latter.

The upper bound of Bayes minimum error probability ε_{ij} for two classes i and j can be expressed as: ([1] p.47)

$$\boldsymbol{P}_{eij} \le \varepsilon_{ij} = \sqrt{\boldsymbol{P}(\omega_i)\boldsymbol{P}(\omega_j)} \int_{-\infty}^{\infty} \sqrt{\boldsymbol{P}(\boldsymbol{x} \mid \omega_i)\boldsymbol{P}(\boldsymbol{x} \mid \omega_j)}\, d\boldsymbol{x}$$

where $\boldsymbol{P}(\omega_i), \boldsymbol{P}(\omega_j), \boldsymbol{P}(\boldsymbol{x} \mid \omega_i), \boldsymbol{P}(\boldsymbol{x} \mid \omega_j)$ are the prior probabilities and conditional probabilities of classes i and j, respectively.

If the distribution of samples is unknown, the normal distribution provides a reasonable approximation. Under normal distribution assumption, the upper bound of classification error probability for two classes i and j, ε_{ij}, can be simplified as:

$$\varepsilon_{ij} = \sqrt{\boldsymbol{P}(\omega_i)\boldsymbol{P}(\omega_j)}\, \exp(-\mu(1/2)), \qquad (3)$$

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

$$\mu(1/2)=\frac{1}{8}\boldsymbol{tr}\left[\boldsymbol{W}_{ij}\boldsymbol{B}_{ij}\right]+\frac{1}{2}\ln\frac{\left|\boldsymbol{W}_{ij}\right|}{\left|\boldsymbol{W}_{i}\right|^{1/2}\left|\boldsymbol{W}_{j}\right|^{1/2}}$$

where $\mu(1/2)$ is Bhattacharyya distance. W_i and W_j are the expected within-class scatter matrices of classes *i* and *j*, respectively. The average of them is denoted as W_{ij}. B_{ij} is the between-class scatter matrix for classes *i* and *j*. They are defined as:

$$\boldsymbol{W}_{i}=\boldsymbol{E}\left[\left(\boldsymbol{X}-\boldsymbol{M}_{i}\right)\left(\boldsymbol{X}-\boldsymbol{M}_{i}\right)^{t}\right],$$

$$\boldsymbol{W}_{j}=\boldsymbol{E}\left[\left(\boldsymbol{X}-\boldsymbol{M}_{j}\right)\left(\boldsymbol{X}-\boldsymbol{M}_{j}\right)^{t}\right],$$

$$\boldsymbol{W}_{ij}=\left(\boldsymbol{W}_{i}+\boldsymbol{W}_{j}\right)/2$$

$$\boldsymbol{B}_{ij}=\left(\boldsymbol{M}_{i}-\boldsymbol{M}_{j}\right)\left(\boldsymbol{M}_{i}-\boldsymbol{M}_{j}\right)^{t}.$$

where M_i and M_j are mean vectors of classes *i* and *j*, respectively .

The upper bound of the classification error after dimension reduction, $\varepsilon_{ij\phi}$, and Bhattacharyya distance $\mu_{ij\phi}(1/2)$ can be expressed as:

$$\varepsilon_{ij\phi}=\sqrt{\boldsymbol{P}(\omega_i)\boldsymbol{P}(\omega_j)}\exp(-\mu_{ij\phi}(1/2))$$

$$\mu_{ij\phi}(1/2)=\frac{1}{8}\boldsymbol{tr}\left[\left(\phi^{t}\boldsymbol{W}_{ij}\phi\right)^{-1}\left(\phi^{t}\boldsymbol{B}_{ij}\phi\right)\right]+\frac{1}{2}\ln\frac{\left|\left(\phi^{t}\boldsymbol{W}_{ij}\phi\right)\right|}{\left|\left(\phi^{t}\boldsymbol{W}_{i}\phi\right)\right|^{1/2}\left|\left(\phi^{t}\boldsymbol{W}_{j}\phi\right)\right|^{1/2}}$$

For the multi-class problem, if there exist L normal distributed classes, which have equal prior probabilities, the upper bound of Bayes error probability in the reduced feature space can be expressed as the sum of that of every two-class pair (refer to [2] 1st Edition p.268):

$$\boldsymbol{P}(\omega_i)=\boldsymbol{P}(\omega_j)=\frac{1}{\boldsymbol{L}(\boldsymbol{L}-1)}$$

$$\varepsilon_{L\phi}=\sum_{i>j}^{L}\sum_{j=1}^{L}\varepsilon_{ij\phi}=\frac{1}{L(L-1)}\sum_{i>j}^{L}\sum_{j=1}^{L}\exp\left[-\mu_{ij\phi}(1/2)\right] \quad (4)$$

Differentiating $\varepsilon_{L\phi}$ with respect to transform matrix ϕ and setting the result to 0 (see [2] 2nd Edition pp. 566-568，formula A16 and A27), we obtain

$$\nabla_{\phi}(\varepsilon_{L\phi})=-\frac{1}{L(L-1)}\sum_{i>j}^{L}\sum_{j=1}^{L}\exp[-\mu_{ij\phi}(1/2)]\cdot\nabla_{\phi}(\mu_{ij\phi}(1/2))=0 \quad (5)$$

where

$$\nabla_{\phi}(\mu_{ij\phi}(1/2))=\left\{W_{ij}\left[\left(W_{ij}^{-1}B_{ij}+4I\right)\phi-\phi\left(\phi^{t}W_{ij}\phi\right)^{-1}\left(\phi^{t}B_{ij}\phi\right)\right]\cdot\left(\phi^{t}W_{ij}\phi\right)^{-1}-2W_{i}\phi\left(\phi^{t}W_{i}\phi\right)^{-1}-2W_{j}\phi\left(\phi^{t}W_{j}\phi\right)^{-1}\right\}/4$$

Because Equation（5）is a highly non-linear equation of the transform matrix ϕ, it is very hard to obtain the analytical solution. In order to solve this problem, we propose a recursive algorithm based on gradient method:

$$\phi_{r+1}=\phi_{r}-\lambda\nabla_{\phi}\left(\varepsilon_{L\phi}\right) \quad (6)$$

where λ is the step size and *r* the number of recursion.

The gradient of Bhattacharyya distance can be expressed in an analytical form. The gradient of the upper bound of classification error can also be obtained in an analytical form. Based on these facts, the gradient search method can be simplified as a recursive algorithm of transform matrix itself, namely,

$$\phi_{r+1}=\boldsymbol{f}\left(\phi_{r}\right) \quad (7)$$

This is much simpler in computation. The recursive algorithm continues until

$\left\|\phi_{r+1}-\phi_{r}\right\|<\delta$ (δ is a small threshold value).

3. PCA pre-processing and BD

Although the BD recursive algorithm can achieve the accurate solution of the feature selection matrix, it has limitations in practice because of the highly demanding computational cost. In order to solve this problem, we combine PCA pre-processing and the BD recursive algorithm together. In our feature selection scheme, a PCA pre-processing is first carried out to reduce the original dimension $\boldsymbol{n}_0$ to a middle dimension $\boldsymbol{n}$. Then, the recursive algorithm is used to do the rest of the dimension reduction, namely, from *n* to the final resultant dimension *m*. Let ϕ_{PCA} as the transform matrix of PCA pre-processing, ϕ_{BD} the transform matrix obtained by the recursive algorithm, then the full transform matrix combining PCA and BD (denoted by PCA+BD for simplicity in the rest of the paper) is:

$\phi^{t}=\phi_{BD}^{t}\phi_{PCA}^{t}$ or $\left(\phi^{t}\right)_{m\times n_0}=\left(\phi_{BD}^{t}\right)_{m\times n}\left(\phi_{PCA}^{t}\right)_{n\times n_0}$.

4. Experiments

Some experimental works are presented in this

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

section to verify the validity of the proposed scheme.

4.1. Experiments on 8-dimensional 4-class data

4.1.1. BD vs. PCA. The data used in this experiment are a set of test data with 8-dimensional 4-class case including only means and covariances from [2]. We generate totally 4000 samples, and 1000 samples in each class by MATLAB. A scaling factor F is introduced to control the relative distance between different classes $\boldsymbol{M}_k \Rightarrow (\boldsymbol{F} \cdot \boldsymbol{M}_k)$. For example, the between-class distance is unchanged when F=1. The between-class distance is increased when F=2 and the between-class distance is decreased when F=0.1. The bigger the F, the larger the between-class distance will be.

Figure 1 illustrates the performance comparison between PCA+ Bayes classifier and BD+ Bayes classifier. It can be seen that BD+ Bayes classifier outperforms PCA+ Bayes classifier for different F except very few cases.

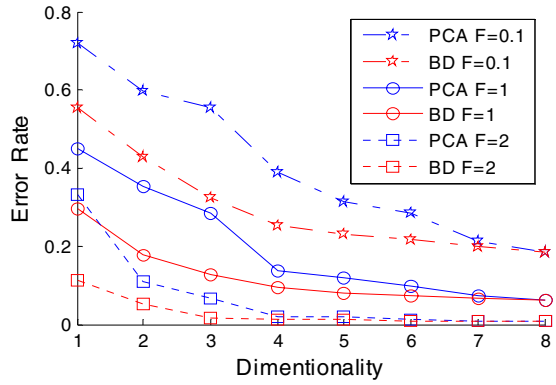


Figure1. Comparison of BD and PCA under different F

4.1.2. PCABD vs. BD, PCA and LDA. Figure 2 shows the performance comparison of different feature selection schemes.

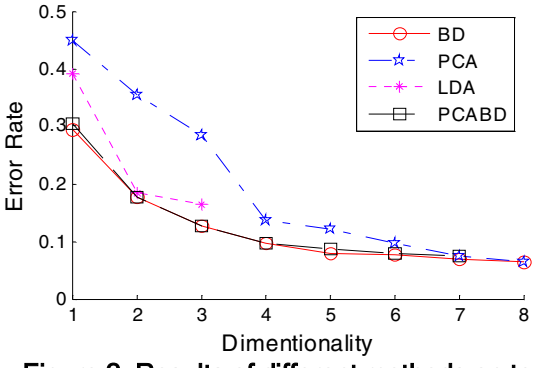


Figure 2. Results of different methods on test data

The following schemes are included: BD+ Bayes classifier (BD), PCA+ Bayes classifier (PCA), LDA+ Bayes classifier (LDA-Linear Discriminant Analysis, or FLD-Fisher Linear Discriminant), and PCA+BD+ Bayes classifier (PCABD). PCABD is just what we proposed in Section 3. It utilizes PCA to reduce the original dimension to some middle dimension firstly. It can be seen in Figure 2 that the error rate of PCABD is just slightly higher than that of BD. However, PCABD works much faster than BD. This is very important in practical applications, which have been shown in the following two experiments.

4.2. Experiments on Digit Recognition

In this experiment, the handwritten-digit recognition is carried out on the MNIST database [6]. The sizes of each image in the MNIST is 28×28, namely, the original dimensionality is 28×28 = 784. We divide the database into a training data set and a testing data set. There are 60,000 samples in the training set and 10,000 samples in the testing set.

Table 1 Time cost with different middle dimensions

Middle Dimension (n)	500	100	50	30
Time Consumption (Minutes)	54.8	4.2	1.2	~0

Table 1 lists the time consumption for different middle dimension choices (The experiment is running on a personal computer with a Pentium IV 1.8G CPU and 256M Rambus memory). In this experiment, the final resultant dimension is 30 and the middle dimensions by PCA pre-processing are set to be 500, 100, 50, and 30, respectively. The highest middle dimension is set to 500, because the rank of the scatter matrices is about 500. As we can see, the time cost is dramatically reduced after PCA pre-processing.

We have also compared the classification results of the following feature selection schemes: A: PCABD (PCA+BD+ Bayes classifier) which has been described in detail in Section 3. B: LDA. C: Pure PCA algorithm.

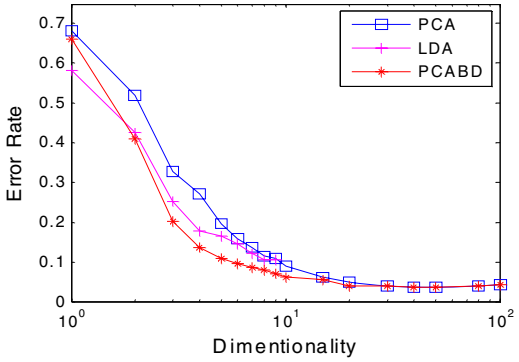


Figure 3. Results of different methods on MNIST

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

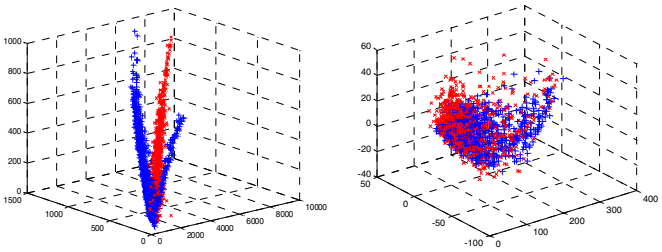
The middle dimension in PCABD is set to 100. The results are shown in Figure 3.

As we can see, the performance of proposed PCABD is obviously better than that of LDA and PCA in terms of lower classification error rates.

4.3 Experiments on Applications to Steganalysis

Another application of feature selection is to map the original high dimensional data into 2-D or 3-D space so that the data can be displayed without losing much original distribution information. The performance of 3-D display by the proposed BD method and PCA are compared in this experiment. As an example, we show the application to the steganalysis method proposed in [7], which extracts features (39-dimensionality) based on statistical moments of wavelet characteristics functions to determine whether any information is embedded in a cover image or not. In order to observe the results visually, we map the original 39-dimensional feature vectors into 3-D space for display.

This steganalysis example belongs to the two-class classification. The 3-D distributions of the feature vectors are shown in Figure 4 in which the original images are marked by "×"s in red (1096 original CorelDraw images), the stego-images by "+"s in blue (corresponding 1096 stego-images by least bit-plane (LSB) data hiding with the data embedding rate being 0.3 bpp (bits per pixel)).



(a) BD (b) PCA

Figure 4. Dimension Reduction

Without dimension reduction, the detection rate of the method proposed in [7] for LSB embedding is 94.0%. The detection rate of Bhattacharyya Distance dimension reduction is still as high as 87.0%, and visually separable. The detection rate of PCA drops dramatically to 67.7% and the two classes are no longer visually separable. This steganalysis display example demonstrates that our proposed BD method reflects the original distribution in a reduced dimensional space more reliably.

5. Conclusions

(1) This paper proposed a feature selection algorithm based on Bhattacharyya distance for multi-class classification problem, which result in a recursive algorithm to obtain the accurate solution of feature selection matrix.

(2) The analytical form of the error rate upper bound in terms of the gradient of transformation matrix is derived. Such that the gradient search algorithm can be simplified as a recursive algorithm of the transformation matrix itself.

(3) PCA pre-processing is integrated with Bhattacharyya distance feature selection in this paper. Our extensive experiments demonstrate that the proposed scheme accelerates the processing speed dramatically without sacrificing the high classification performance.

(4) The experiments have shown that our proposed BD method outperforms the PCA method in the sense that the detection rate does not reduce significantly from the high dimensional case.

6. References

[1] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd Edition, John Wiley & Sons, 2001.

[2] K. Fukunaga, "Introduction to Statistical Pattern Recognition", Academic Press, Inc., Boston, 1978（1st Edition)，1990（2nd Edition).

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

[4] E. Choi, et al, "Feature extraction based on the Bhattacharyya distance", Pattern Recognition Vol. 36 ,2003, pp.1703–1709

[5] 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.

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

[7] G. Xuan, Y. Q. Shi, J. Gao, D. Zou, C. Yang, Z. Zhang, P. Chai, C. Chen, W. Chen, "Steganalysis Based on multiple features formed by statistical moments of wavelet characteristic functions," Information Hiding Workshop (IH05), June 2005, Springer-Verlag, Vol. 3727 / 2005, pp. 262 - 277