A NOVEL CLASSIFIER FOR HANDWRITTEN NUMERAL RECOGNITION

Ying Wen, Pengfei Shi

Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, China

ABSTRACT

This paper presents a novel pattern classification approach–a kernel and Bhattacharyya distance based classifier which utilizes the distribution characteristics of the samples in each class. Bhattacharyya distance in the subspace spanned by the eigenvectors which are associated with the smaller eigenvalues in each class is adopted as the classification criterion. The smaller eigenvalues are substituted by a small value threshold in such a way that the classification error in a given database is minimized. Application of the proposed classifier to the issue of handwritten numeral recognition demonstrates that it is promising in practical applications.

***Index Terms*—** Pattern classification, character recognition, feature extraction

1. INTRODUCTION

Generally, a recognition system consists of preprocessing, feature extraction, classifier and postprocessing. For a large data set, the high dimensionality problem has to be solved first. PCA (Principal Component Analysis) is often used to linearly transform a high-dimensional input vector into a low-dimensional one whose components are uncorrelated [1]. However, PCA sometimes does not provide a satisfactory performance in classification, because PCA averages the characteristics of not only the between-class but also the within-class. PCA uses the global covariance while the within-class contribution to the recognition performance is not taken into account.

Recently, Bhattacharyya distance has investigated in pattern classification. C. Lee, et al reported that the accurate estimation of classification error becomes possible by using the Bhattacharyya distance [2]. Choi, et al depicted a feature selection approach based on Bhattacharyya distance in [3]. In this paper, we combine the Bhattacharyya distance with the kernel approach to propose a new pattern classification scheme, viz. a kernel and Bhattacharyya distance based classifier (KBD for short hereinafter). Bhattacharyya distance is regarded as an optimization criterion assuming the samples of database are subjected to the normal distribution. KBD makes full use of characteristics of each class distribution such as the class mean and covariance. Furthermore, we explore the relationship between the kernel and Guassian processes, and apply a kernel to Bhattacharyya distance to achieve a better recognition performance. In addition, we utilize a threshold to replace these smaller eigenvalues for the solution of non-existing inverse matrix for covariance matrix. In general, the threshold is selected so that the classification error in a given database is minimized. We apply the proposed classifier to the issue of handwritten numeral recognition and the experiment demonstrates it is promising in practical applications.

The rest of the paper is organized as follows: The proposed algorithm is depicted in Section 2. The experimental results of the proposed classifier are presented in Section 3, and the conclusion is drawn in Section 4.

2. ALGORITHM DESCRIPTION

2.1. Bhattacharyya distance

Suppose that a database has C classes, and $X = \{x_j; j = 1, 2, \cdots, n\}$ is a data set of ith class. The mean vector of ith class: $\mu_i = \frac{1}{n}\sum_{j=1}^{n} x_j$; the covariance matrix of the ith class: $\Sigma_i = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu_i)'(x_j - \mu_i)$. In the paper, the definitions of i, j fit for all equations, i.e.,$j = 1, 2, \cdots, n; i = 1, 2, \cdots, C$.

Supppose that $P(\omega) = \{P(\omega_i); i = 1, 2, \cdots, C\}$ is the probability of each class. When the distribution of the set is Gaussian distribution, the Bhattacharyya distance function is defined by:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i) \quad (1)$$

where, Σ_i^{-1} denotes the inverse matrix of the covariance matrix obtained from Σ_i. There exists theorthonormal matrix U_i , so that $\Sigma_i = U_i'\Lambda_i U_i$, where, Λ_i is the eigenvalues matrix of Σ_i ($\Lambda_i = diag\{\lambda_{i1}, \cdots, \lambda_{ik}, \lambda_{ik+1}, \cdots, \lambda_{id}\}$). We sort the eigenvalues (and their corresponding eigenvectors) so that $\lambda_{i1} > \cdots > \lambda_{ik} > \lambda_{ik+1} > \cdots > \lambda_{id}$. Then , Eq.(1) can be rewritten as:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)'(U_i'\Lambda_i U_i)^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Lambda_i| + \ln P(\omega_i) \quad (2)$$

2.2. Kernel and Gaussian Processes

For a finite set of linear variables $X = (x_1, \cdots, x_l)$, a Gaussian distribution (with zero mean) is specified by a symmetric

1-4244-1484-9/08/$25.00 ©2008 IEEE

positive definite covariance matrix $\Sigma = \Sigma(x_1, \cdots, x_l)$ with the corresponding distribution given by:

$$P_{\mathcal{F}\sim\mathcal{D}}[(f(x_1), \cdots, f(x_l)) = (y_1, \cdots, y_l)] \propto exp(-\frac{1}{2}y'\Sigma^{-1}y) \quad (3)$$

A Gaussian process is a stochastic process for which the marginal distribution for any finite set of variables is zero mean Gaussian. The (p, q) entry of Σ measures the correlation between $f(x_p)$ and $f(x_q)$, that is the expectation $E[f(x_p)f(x_q)]$, and hence depends only on x_p and x_q. There therefore exists a symmetric covariance function $K(x, z)$ such that $\Sigma(x_1, \cdots, x_l)_{pq} = K(x_p, x_q)$. All finite sets of input points of the covariance matrix is required positive definite, which nicely conforms to the defining property of a Mercer kernel given in [4], and hence we see that defining a Gaussian process over a set of variables indexed by a space X is equivalent to defining a Mercer kernel on $X \times X$. The definition of a Gaussian process by specifying the covariance function avoids explicit definition of the function class, and the prior over the functions in. Indeed one choice of function space is the class of linear functions in the space F of Mercer features:

$$X = (x_1, \cdots, x_l) \longrightarrow \Phi(X) = (\phi_1(x), \cdots, \phi_q(x), \cdots) \quad (4)$$

in the l_2 space defined by the weighted inner product given by

$$< \Psi \cdot \widetilde{\Psi} >= \sum_{q=1}^{\infty} \lambda_q \psi_q \widetilde{\psi_q} \quad (5)$$

The prior distribution over the weight vector Ψ is chosen to be an independent zero mean Gaussian in each coordinate with covariance in coordinate p equal to $\sqrt{\lambda_p}$. From the above statement, we can see the relationship between the kernel and Gaussian processes. One condition of Bhattacharyya distance is that the samples in a given database are subjected to Gaussian distribution, and then we try to find out the relationship between the kernel and Bhattacharyya distance. Here, we describe a property of the kernel. Let B be a symmetric positive semi-definite matrix. Consider the diagonalisation of $B = V'\Lambda V$ by an orthogonal matrix V, where Λ is the diagonal matrix containing the non-negative eigenvalues. Let $\sqrt{\Lambda}$ be the diagonal matrix with the square roots of the eigenvalues and set $A = \sqrt{\Lambda}V$. We therefore have:

$$\begin{aligned} x'\Sigma z &= x'V'\Lambda V z = x'V'\sqrt{\Lambda}\sqrt{\Lambda}V z \\ &= x'A'Az =< Az \cdot Ax >= K(x, z) \end{aligned} \quad (6)$$

where the inner product using the feature mapping A, $K(x, z)$ is the kernel. Based on the property of the kernel and the relationship between the kernel and Gaussian processes, the main part of Bhattacharyya distance becomes:

$$\begin{aligned} g_i(x) &= -\frac{1}{2}(x - \mu_i)'(U_i'{\Lambda_i}^{-1}U_i)(x - \mu_i) \\ &= K(\sqrt{\Lambda_i^{-1}}U_i(x - \mu_i), \sqrt{\Lambda_i^{-1}}U_i(x - \mu_i)) \end{aligned} \quad (7)$$

We employ the above discriminant to classification. The kernel K is selected as the polynomial or Gaussian kernel in the experiment, denoted by:

(1) Polynomial kernel: $K(x, z) = (< x, z > +C)^d$ where d is any positive integer and C is a constant

(2) Gaussian kernel: $K(x, z) = exp(-||x - z||^2/\sigma^2)$ In the experiment, the value of σ depends on the eigenvalues of each class.

2.3. Classification Approach Based on Kernel and Bhattacharyya Distance (KBD)

We calculate the covariance matrix of each class and obtain the corresponding eigenvalues and eigenvectors. If the data distribution is Gaussian distribution, the classification capability is optimal when the total eigenvectors are calculated. The smaller the eigenvalue, the better the classification performance. Because the smaller eigenvalue reflects the convergence of within-class, the corresponding eigenvector is more important. But, in practice, there exist some small eigenvalues that are close to zero, which causes a problem to calculate the inverse matrix of the covariance matrix Λ_i^{-1} and then Bhattacharyya distance classifier becomes invalid.

To deal with this problem, we utilize a threshold λ_0 to replace all of eigenvalues which are less than λ_0, i.e., $\lambda_{ij} = \lambda_0$, if $\lambda_{ij} \leq \lambda_0, j = k + 1, k + 2, \cdots, d, i = 1, 2, \cdots, C$. Then we obtain the new matrix $\widetilde{\Lambda_i}$:

$$\widetilde{\Lambda_i} = diag\{\lambda_{i1}, \cdots, \lambda_{ik}, \lambda_0, \cdots, \lambda_0\} \quad (8)$$

Although the value of λ_0 is very small, Σ_i^{-1} could be inversed and we can continue the downstreaming process. In this approach, we employ λ_0 to substitute all small eigenvalues of each class, i.e. the smallest eigenvalue of each class is same.

In fact, the samples in a given database are not always subjected to Gaussian distribution. In this situation, we can still employ the above method. For a given database, the value of the threshold λ_0 depends on the error rate of classification (see Eq.(9)).

$$\lambda_0 = \min_{\lambda_0}(error_{database}) \quad (9)$$

Based on the above statement, we combine Bhattacharyya distance with the Kernel method, and utilize a specified threshold which replaces small eigenvalues to solve the inverse matrix of the covariance matrix. We name this approach KBD hereinafter. If Gaussian kernel or Polynomial kernel is employed, it is called KBD-G and KBD-P, respectively. BD is used to stand for the classifier of Eq.(2).

The proposed classification approaches are based on Bayes discriminant. However, the optimal parameter is selected in such a way that the classification error in a given database is minimized, so our classification approach is the improved Bayesian classification, not strict Bayesian classification in some sense.

3. EXPERIMENTS AND RESULTS

3.1. Experiments on Handwritten Bangla Numeral Recognition

In order to verify the validity and performance of the proposed classifier, we apply it to the issue of the handwritten Bangla numeral recognition [5].

Figure 1 shows the samples of handwritten Bangla numerals. These numerals are acquired from live letters by the automatic letters sorting machine in the Dhaka mail processing centre of Bangladesh Post Office. We randomly select 30,000 samples as the training set and 15,000 samples as the test set. In this paper, our experiments are tested on a PC with Windows xp, Pentium 1.8G and 512 RAM. To evaluate the performance of our proposed pattern classification, apart from BD, KBD-P and KBD-G, three other classifiers are also tested for the purpose of comparison. They are Euclidean Distance, BP and SVM (Gaussian kernel).

Note that the six classifiers utilize PCA to lower the dimension to 100 for each input pattern. Table 1 presents the recognition rate and the processing speed of the six classifiers. It can be found that the recognition rate of Euclidean Distance is not good but the recognition time is the least. The recognition rates of the first three classifiers are lower than the rest. Obviously, the recognition time of the proposed classifiers are dramatically reduced while the high recognition rate is obtained. It can be seen that the classifiers proposed in this paper are superior to the others in the recognition performance.

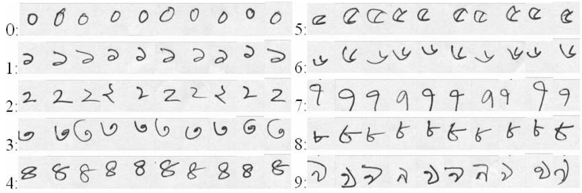

Fig. 1. Samples of handwritten Bangla numeral images(0-9)

Table 1. Performance comparison of six approaches

Classifier	Recognition rate	Recognition time
Euclidean Distance	85.83%	1.61ms
BP	89.01%	6.21ms
SVM	90.24%	12.41ms
BD	95.46%	7.43ms
KBD-P	96.85%	7.46ms
KBD-G	96.91%	7.52ms

In the above experiments, we compare the recognition results achieved by different classifiers. To further investigate the recognition performance of the proposed classifiers, we carry out some experiments to show the relations among the recognition performances such as the recognition rate, the error rate.

Figure 2 shows the relationship between the threshold λ_0 and the recognition rate. The recognition rate increases with the threshold increasing. But when the threshold is larger than 0.5, the recognition rate decreases. The figure shows that the threshold is not always small when the error rate is low. As the whole, the recognition result of KBD is superior to BD regardless of polynomial or Gaussian kernel to be adopted.

The proposed classifiers minimize the classification error in the training set, so they are obviously depend on the training set used. An experiment is carried out to investigate the influence of the number of training samples on the recognition performance. Figure 3 shows the recognition rate of the classifiers achieved by different numbers of training samples where λ_0 is set to be constant 0.25. It can be found that the recognition result becomes stable while the training set has more than 20,000 samples. From our experiments, we can see the average recognition rates of the classifiers are similar.

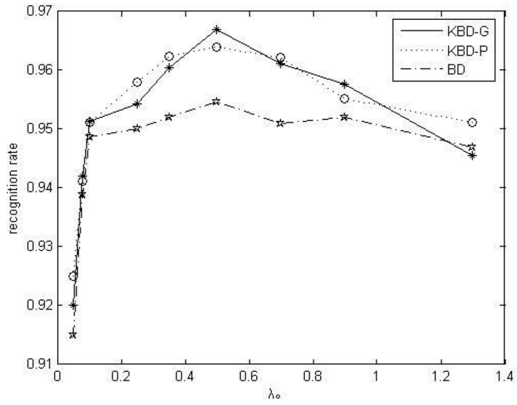

Fig. 2. Recognition rate vs. λ_0

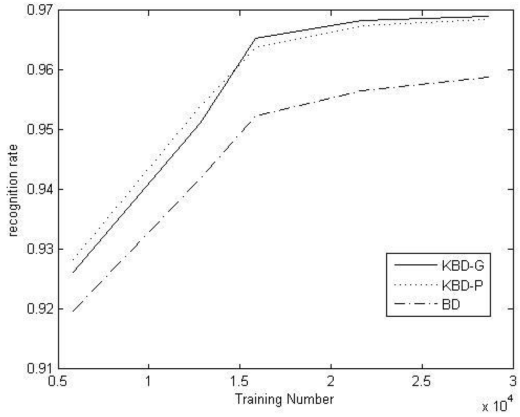

Fig. 3. Recognition rate vs. training number

3.2. Experiments on UCI and MNIST data sets

So far, we have described the results of the approaches for recognizing handwritten Bangla numerals. To test the feasibility of the proposed methods to other application, the approach is tested on two numeral databases. The first database is UCI data set which is consisting of 5620 handwritten numeral characters. A random procedure is used for partition of the database to the training and testing subsets, and 300 for training and the rest for testing. The original image of each numeral character has the size of 32 × 32 pixels. We compare our proposed classifier with some existing classifiers. The comparison is listed in Table 2, from which we can see that the recognition rates achieved by our classifiers are better and the proposed method scales well to a small database of handwritten digits.

The second database is the MNIST database. It contains 60,000 handwritten digit images for the classifier training and 10,000 handwritten digit images for the classifier testing. All digits have been size-normalized and centered in a 28 × 28 box. Some classifiers proposed in previous papers are used to test [6, 7, 8]. Figure 4 presents the performances of different algorithms tested on the MNIST database. The proposed classifier (KBD-P) is used on the original MNIST database and achieves a good performance with 1.8% error rate. The experimental result of LeNet5 is 0.95% error rate, one of the best classifiers on the market. Although the recognition performance achieved by our classifier is not the best, our classifier is efficient and not complex.

Table 2. Recognition rates of seven classifers on UCI data

Classifier	Recognition rate(%)
Linear classifier(1 layer NN)	93.39
K-NN	95.82
BP	96.85
SVM	97.97
BD	98.39
KBD-P	99.08
KBD-G	98.93

4. CONCLUSION

A new classifier based on the kernel approach and Bhattacharyya distance is proposed in this paper. KBD makes full use of characteristics of each class distribution such as the class mean and covariance while it does not average the covariance matrix of all classes. In all eigenvalues of each class, a small value threshold is used to substitute the smaller eigenvalues to overcome the problem of non-existing inverse matrix for covariance matrix so that the classification error in a given database is minimized. The experimental results have demonstrated its efficiency in both the recognition rate and the classification time.

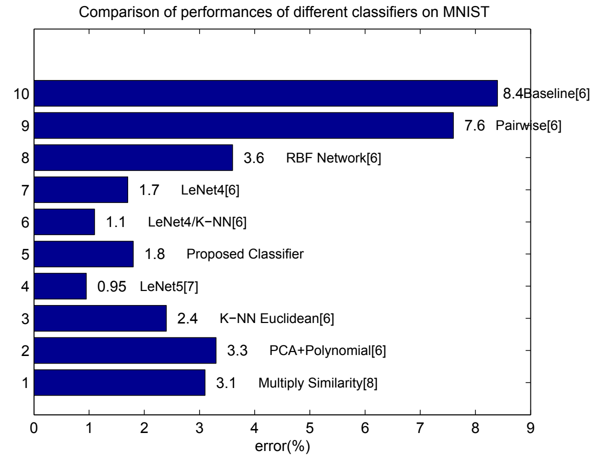

Fig. 4. Error rates of different classifiers on the test set of MNIST database. Each bar represents a classifie

Acknowledgements

This research is Supported by 863 Program (2006AA01Z119) and National Natural Science Foundation of China(60775009).

5. REFERENCES

[1] M.Partridge and R.A.Calvo, Fast Dimensionality Reduction and Simple PCA, Intelligent Data Analysis 2 (1998) 203-214.

[2] C. Lee and E. Choi, Bayes error evaluation of the Gaussian ML classifier, IEEE Trans. Geosci. Remote Sens. 38 (3) (2000) 1471-1475.

[3] E.Choi and C.Lee, Feature extraction based on the Bhattacharyya distance, Pattern Recognition Vol. 36,2003, pp.1703-1709.

[4] N.Cristianini and J.S.Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press.

[5] Y.Wen, Y.Lu and P.F.Shi, Handwritten Bangla numeral recognition system and its application to postal automation, Pattern Recognition, 40 (2007) 99-107.

[6] Y.LeCun, L.D.Jackel et al. Comparison of learning algorithms for handwritten digit recognition. In: Proceedings of the International Conference on Artificial Neural Networks, Paris,1995, pp. 53-60.

[7] Y.LeCun, L.Bottou, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998 (11), 2278-2324.

[8] S.W.Lee, Multilayer cluster neural network for totally unconstrained handwritten numeral recognition. Neural Networks 1995, (5), 783-792.