LETTER *Special Section on Information-Based Induction Sciences and Machine Learning*

MQDF Retrained on Selected Sample Set

Yanwei WANG[†a)], ***Student Member***, **Xiaoqing DING**[†b)], ***and*** **Changsong LIU**[†c)], ***Nonmembers***

SUMMARY This letter has retrained an MQDF classifier on the retraining set, which is constructed by samples locating near classification boundary. The method is evaluated on HCL2000 and HCD Chinese handwriting sets. The results show that the retrained MQDF outperforms MQDF and cascade MQDF on all test sets.

key words: *MQDF retraining, sample selection, sample number regularization, Chinese handwriting recognition*

1. Introduction

Offline Chinese handwriting recognition is considered as one of the most challenging problems in pattern recognition. It has been extensively studied in the past decade due to its wide applications and commercial needs. There are two dominant kinds of recognition methods namely generative method and discriminative method.

Among various generative methods, modified quadratic discriminant function (MQDF) [1] is one of the most representative one with excellent performance yet low computational complexity. However for free style handwriting, with large variability in character shapes and appearances samples do not satisfy Gaussian distribution strictly. As a result, MQDF estimated with maximum likelihood could not achieve the optimal classification performance. On one hand, the improvement could be made by adjusting MQDF parameters under discriminative principles, such as minimum classification error (MCE) [2]. Due to large computational complexity on large scale classification tasks, MCE leaves covariance matrix of MQDF not adjusted. That is bound to result in suboptimal discriminative learning. Hereafter, MCE based MQDF is improved through combination with compound Mahalanobis function (CMF) [3]. On the other hand, multi-Gaussian model shows capability to deal with non-Gaussian distribution problem. Gaussian mixture model (GMM) [4] is common used. It has been theoretically proved [5] that GMM could approach to any probability distribution if there are enough mixtures. As mixtures increase in the number, parameters of GMM will be doubled and redoubled. In practice, samples obtained are limited. It will lead to over fitted problem since too many parameters estimated on relatively small sample set. Fu [6] has proposed discriminative cascade MQDF, in which two MQDFs are trained and cascaded together. It has improved performance of MQDF effectively but still could not avoid small sample size problem.

Different from generative methods, support vector machine (SVM) [7], [8] avoids probability distribution assumption and directly learns classification boundary. It performs excellent in small category classification tasks by solving a quadratic programming problem. However, SVM is time consuming, especially in high dimensional feature space. Although Platt [9] has proposed a low complexity version of SVM, in large category classification problems computational complexity and model storages are still unacceptable.

Discriminative model's successful application in small category classification indicates that samples near classification boundary could provide important references for determining classification boundary. Based on this idea, this letter has retrained an MQDF denoted as rMQDF on a selected retraining set. Samples close to classification boundary are selected and added to the retraining set, and samples not selected will be discarded. The number of selected samples is regularized by a constant in order to combat against performance degradation resulting from small sample size problem. Compared with traditional MQDF, rMQDF assimilates discriminative information at the cost of recognizing the whole training set. Compared with cascade MQDF classifiers, rMQDF avoids small sample size problem by regularization technique and has low recognition complexity.

2. Offline Chinese Character Recognition System

The block diagram of retraining an MQDF is shown in Fig. 1. The traditional MQDF based recognition system consists of discriminant analysis, MQDF parameter estimation with maximum likelihood. Apart from traditional parts, recognizing training set, discriminant analysis and retraining MQDF are added to new system. The former is designed to help to construct a retraining set and the later two aim to learn a boosted MQDF classifier. The new added processing makes it possible that MQDF is improved with low recognition complexity. Firstly, discriminative information has been obtained by recognizing training set and applied in sample selection. Secondly, sample distance measure are consistent. The classifier used for sample selection is of the same type as the retrained one since one sample recognized

Manuscript received December 28, 2010.
Manuscript revised March 23, 2011.

[†]The authors are with State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.
a) E-mail: wangyw@ocrserv.ee.tsinghua.edu.cn
b) E-mail: dxq@ocrserv.ee.tsinghua.edu.cn
c) E-mail: lcs@ocrserv.ee.tsinghua.edu.cn

DOI: 10.1587/transinf.E94.D.1933

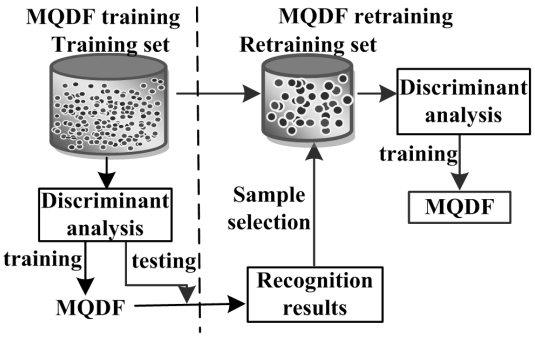

Fig. 1 Block diagram of offline Chinese character recognition system.

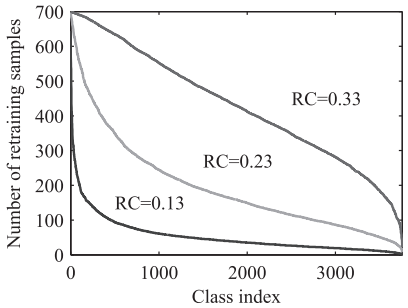

Fig. 2 Number of retraining samples selected from HCL2000 training set with different RC.

by different type classifiers will get different measurements. This ensures that selected samples gain consistent importance for rMQDF. Thirdly, rMQDF could be learned more accurately since it is trained on samples close to classification boundary yet traditional MQDF does not take classification performance in consideration. Last but not the least, no extra recognition complexity is added to rMQDF. Training complexity of a MQDF classifier is denoted as O_{tr} and recognition complexity O_{ts}. Both cascade MQDF and rMQDF have trained two MQDF classifiers thus they have the same training complexity equaling to $2O_{tr}$. For recognition, cascade MQDF recognizes a sample twice by two level MQDFs and integrates the recognition results. Therefore its recognition complexity is $2O_{ts}$. rMQDF gets approximately the same recognition complexity as baseline MQDF equaling to O_{ts}.

3. Retraining Set Construction

The retraining set is constructed through sample selection. A baseline MQDF is estimated on training set and recognizes the training set. It outputs distance based recognition results, based on which samples are selected according to the following principles.

(1) Samples misclassified. These samples has come over classification boundary and mixed with samples of the other class. They are determined simply according to recognized labels and groundtruth.

(2) Samples recognized correctly but close to classification boundary. For some samples, recognition results are correct but prone to be contaminated by noise and parameter estimation bias since these samples are excessively close to the classification boundary. General recognition confidence [10], an effective measurement for reliability of a recognition result is engaged as Eq. (1) to filter these samples.

$$RC = 1 - d_1/d_2 \tag{1}$$

Where d_1 and d_2 are distance measurements of the first two candidates and satisfy $d_1 < d_2$. From a statistical point of view, the higher the confidence is, the higher probability that one sample is recognized correctly with. The confidence approaches to zero under the condition that $d_1 \approx d_2$. It means the sample is recognized unstable and locates near classification boundary. Generally, a constant threshold TH is predefined and once RC is smaller than TH then the sample will be added to retraining set. This rule is the same as the one used in [6], in which determination of an effective TH is also detailed.

(3) Sample number regularization. The number of selected samples according to the forgoing two rules is denoted as N_b. It differs greatly for different classes as illustrated in Fig. 2. Some classes have less than 100 samples only. For a specific class, if N_b is too small, MQDF would be over fitted and deteriorates severely. To conquer this problem, retraining sample number of each class is regularized by a constant N_{min}. The constant can be regarded as the least number of samples required to train a relatively robust statistical classifier. For a specific class, if $N_b < N_{min}$, it indicates at least $N_{min} - N_b$ samples are required additionally. Removed the preselected N_b samples from training set, the rest samples are sorted according to general recognition confidence in ascending order. Then the preceding $N_{min} - N_b$ samples will be picked out and added to retraining set. If $N_b > N_{min}$, it means there are relatively enough samples, no further processing needs to be done. Noted that N_{min} should be carefully determined. To some extent, it has close relation to variability in training samples of the same class and will affect the overall recognition accuracy. For bad writing style samples, large variability presents in character shapes and appearances. Therefore more training samples are needed to cover different sample patterns. For good writing style samples, sample variability is small. In the limited case if all training samples are the same then most of the samples are redundant and any one of them could represent all sample patterns. Empirically, N_{min} could be set relatively smaller for good writing style samples and larger for bad writing style samples.

4. Experiments

The rMQDF is evaluated on Handwritten Character Library 2000 (HCL2000) and HCD, two Chinese handwriting sets and compared with baseline MQDF and cascade MQDF. HCL2000 is an already published character set [11]. It contains 3755 simplified Chinese character classes and has 1000 subsets in total. Subsets from xx001 to xx700 are used as training set and subsets from hh001-hh300 are testing set. HCD sample sets are divided into 10 subsets and denoted from HCD1 to HCD10. HCD4 and HCD9 are set up as test sets, containing 100×3755 characters and 3755 characters respectively. The other subsets are used as a training set, which contains a total of 1877×3755 characters.

392 dimensional gradient feature [12] is extracted from each normalized character image and projected onto low dimensional subspace by linear discriminant analysis (LDA) [13] and heteroscedastic linear discriminant analysis (HLDA) [14]. The MQDF classifier is denoted as MQDF(nFD, nTD), where nFD is compressed feature dimensionality and nTD is the truncation dimension of covariance matrix. In cascade MQDF, the first level classifier is the baseline MQDF(128,32), and the second level classifier is MQDF(144,44). In order to compared with cascade MQDF, the baseline MQDF has been used to recognize training set and exports distance based recognition results. Retraining set is constructed by principles mentioned previously. N_{min} is optimized by testing serials N_{min}s on training set. On retraining set MQDF(144,44) is trained with maximum likelihood estimation.

4.1 Recognition Results

On HCL2000, the retraining set is constructed by $RC = 0.23$ and $N_{min} = 300$. Several recognition algorithms have been estimated on HCL2000 set. Table 1 lists the important recognition results, where LDP is an abbreviation for local discriminant projection. The results show that rMQDF outperforms the other algorithms and obtains the highest accuracy up to 98.74% and 98.85% based on LDA and HLDA respectively. Compared with improved version of MCE based MQDF, rMQDF gets relatively higher recognition accuracy.

For HCD, the retraining set is selected by $RC = 0.13$ and $N_{min} = 600$. The experimental results are shown in Table 2. It indicates that on both of HCD test sets, rMQDF outperforms cascade MQDF and baseline MQDF. On HCD4, baseline MQDF has already get the high accuracy. The accuracy is improved again by cascade MQDF and rMQDF. On HCD9, compared with baseline MQDF, relative error rate has been decreased up to 35.39% and 32.21% for LDA and HLDA respectively.

Table 1 Important results on HCL2000 dataset.

Author	Discriminant analysis/learning	Method	Recognition rate (%)
Liu [15]	-	Matching	>92.00
Zhang [17]	LDP	Nearest Neighbor	97.53
Long [16]	LDA	compact MQDF	98.12
Fu [6]	LDA	MQDF	98.54
Liu [3]	LDA/MCE	MQDF+CMF	98.56
Fu [6]	LDA	Cacade MQDF	98.70
Proposed	LDA	rMQDF	**98.74**
Proposed	HLDA	rMQDF	**98.85**

4.2 Performance Improvement Investigation

This experiment investigates how the performance of baseline MQDF is improved. Recall from the process of retraining an MQDF, LDA and MQDF parameter estimation both have close relation to the performance enhancement. In this section, they are performed respectively on training set and retraining set. Different combinations of them are employed to train three MQDF classifiers. The first one is a case of baseline MQDF, and LDA and MQDF are both learned on training set. For the second MQDF, LDA is learned on training set while MQDF on retraining set. And the third one is rMQDF. In all tests, nFD = 144, nTD = 44 and the other parameters are specified the same as above section. Three experiments are recorded as A, B and C respectively.

HCL2000 test set and HCD4 are regulated sample sets, thus from a point of statistical view samples salsify independent identically distribution on the whole. As shown in Fig. 3, results on the both sets show that performance improvements mainly comes from MQDF retraining. HCD9 is a set of free writing style and takes on more non-Gaussian characteristics. LDA and MQDF trained on retraining set both improved the performance but the latter gets more apparently. As well known, the baseline MQDF under maximum likelihood estimation treats each sample in the same way. Samples usually do not strictly satisfy Gaussian distribution. The non-Gaussian part of samples account for only a small part of training set and thus gain less concerns in training process. These samples are prone to be recognized unstable or misclassified. The sample selection scheme to some extent is probable to make the assumed distribution

Table 2 Results on HCD dataset.

Test set	Discriminant analysis	Baseline MQDF (%)	Cascade MQDF (%)	rMQDF (%)
HCD4	LDA	98.14	98.32	**98.36**
HCD4	HLDA	98.28	98.38	**98.43**
HCD9	LDA	89.29	91.96	**93.08**
HCD9	HLDA	91.40	93.18	**94.17**

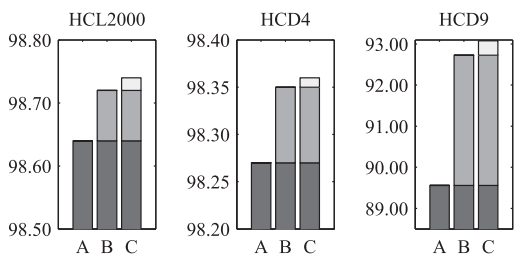

Fig. 3 Comparisons of three training schemes of MQDF on HCL and HCD test sets.

moved towards these samples. In other words, rMQDF concerns more about these samples. The results suggest that the MQDF trained on retraining set contributes dominantly to the performance improvement. It also reveals that to some extent, rMQDF is more adaptable to non-Gaussian distribution.

5. Conclusion

This letter improved performance of MQDF by retraining it on the selected retraining set. It is an alternative way attempting to tackle non-Gaussian distribution problem. On all character sets the retrained MQDF classifier obtained a high recognition performance with low recognition complexity. In future work, more theoretical investigation and systematic analysis will be carried out.

Acknowledgments

This work was supported by the National Basic Research Program of China (973 program) under Grant No. 2007CB311004 and the National Natural Science Foundation of China under Grant Nos. 60933010.

References

[1] F. Kimura, K. Takashina, and S. Tsuruoka et al., "Modified quadratic discriminate functions and the application to Chinese character recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.PAMI-9, no.1, pp.149–153, Jan. 1987.

[2] R. Zhang, X.Q. Ding, and H.L. Liu, "Discriminative training based quadratic classifier for handwritten character recognition," Int. J. Pattern Recognition and Artificial Intelligence, vol.21, no.6, pp.1035–1046, 2007.

[3] H.L. Liu and X.Q. Ding, "Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes," ICDAR, pp.19–23, Seoul, Korea, Sept. 2005.

[4] O.D. Richard, P.E. Hart, and D.G. Stork, Pattern Classification, Second Edition, John Wiley & Sons, New York, 2000.

[5] A.K. Jain, P.W. Robert, and J.C. Mao, "Statistical pattern recognition: A review," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.1, pp.4–37, Jan. 2000.

[6] Q. Fu, X.Q. Ding, T.Z. Li, and C.S. Liu, "An effective and practical classifier fusion strategy for improving handwritten character recognition," ICDAR, pp.1038–1042, Paraná, Brazil, Sept. 2007.

[7] L. Wang and J. Duan, "A novel multi-class cluster SVM for handwritten Chinese character recognition," Second International Symposium on Intelligent Information Technology Application, pp.291–295, Shanghai, China, Dec. 2008.

[8] J.X. Dong, B.A. Krzyźak, and Y.S. Ching, "An improved handwritten Chinese character recognition system using support vector machine," Pattern Recognit. Lett., vol.26, no.12, pp.1849–1856, Sept. 2005.

[9] J.C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Technical Reports: MSR-TR-98-14, Jan. 1998.

[10] X.F. Lin, X.Q. Ding, M. Chen, R. Zhang, and Y.S. Wu, "Adaptive confidence transform based classifier combination for Chinese character recognition," Pattern Recognit. Lett., vol.19, no.10, pp.975–988, Aug. 1998.

[11] H.G. Zhang, J. Guo, G. Chen, and C.G. Liu, "HCL2000-a large-scale handwritten Chinese character database for handwritten character recognition," ICDAR, pp.286–290, Barcelona, Spain, July 2009.

[12] S. Meng, Y. Fujisawa, and T. Wakabayashi et al., "Handwritten numeral recognition using gradient and curvature of gray scale image," Pattern Recognit., vol.35, no.10, pp.2051–2059, Oct. 2002.

[13] R.A. Fisher, "The statistical utilization of multiple measurements," Ann. Eugenics., no.8, pp.376–386, 1938.

[14] H.L. Liu and X.Q. Ding, "Improve handwritten character recognition performance by Heteroscedastic linear discriminant analysis," ICPR, pp.880–883, Hong Kong, China, 2006.

[15] X.B. Liu, Y.D. Jia, and M. Tan, "Geometrical-statistical modeling of character structures for natural stroke extraction and matching," IWFHR, pp.205–209, La Baule, France, Oct. 2006.

[16] T. Long and L.W. Jin, "Building compact mqdf classifier for large character set recognition by subspace distribution sharing," Pattern Recognit., vol.41, no.9, pp.2916–2925, Sept. 2008.

[17] H.G. Zhang, J. Yang, W.H. Deng, and J. Guo, "Handwritten Chinese character recognition using local discriminate projection with prior information," ICPR, pp.1–4, Florida, USA, Dec. 2008.