MQDF Discriminative Learning Based Offline Handwritten Chinese Character Recognition

Yanwei WANG,Xiaoqing DING, Changsong LIU
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology, Tsinghua University
Department of Electronic Engineering
Beijing, China
{wangyw, dxq, lcs}@ocrserv.ee.tsinghua.edu.cn

Abstract—**This paper has proposed a discriminative learning method of modified quadratic discriminant function (MQDF) based on sample importance weights. Firstly, sample importance function is derived from distance based recognition results under bayes decision rule. It weights samples according to extended recognition confidence. On these weighted samples, parameters of MQDF are modulated indirectly by re-estimating the mean vector and covariance matrix. The proposed method is investigated and compared with other discriminative learning methods about MQDF on THU-HCD offline Chinese handwriting sets. The results show that the proposed method has improved the basic MQDF drastically and outperforms other methods compared.**

Keyword: MQDF discriminative learning;sample importance weight;offline Chinese character recognition;larage category classification.

I. INTRODUCTION

Offline Chinese handwritten character recognition is a challenging research area. The major difficulty stems from large variability in character shape, writing style, character scales and so many similar characters. Among the great many methods, modified quadratic discriminant function（MQDF）[1] is an excellent one and widely applied. To learn a basic MQDF classifier, it is assumed that samples are normally distributed with unknown mean and covariance matrix. The unknown parameters are generally estimated with maximum likelihood estimation (MLE).

Under the condition that the assumption is coincident with sample's real distribution and there are a sufficiently large number of samples, classification error could approaches to bayes error. Unfortunately, on one hand, samples always do not strictly satisfy Gaussian distribution thus the MQDF learned will be limited in description of all samples. On the other hand, samples available in real application are limited and furthermore from a statistical point of view, MLE does not take classification performance in consideration therefore it is doomed to fail in getting the optimal classification performance.

Many efforts have been devoted to improving performance of MQDF. Discriminative learning technique is one of the most important methods. Discriminative learning methods are divided into three categories. One, avoiding underlying probability assumption, constructs and modulates classification boundaries directly by minimizing empirical risk. Support vector machine (SVM) [2], adaboost [3] are two of the most representative methods and both of them could learn complex classification boundaries in feature space. They have been applied successfully in small scale classification problems such as alpha numeric identification [4],[5] and begin to be extended to large scale classification problem [6]. There is another kind of discriminative learning methods, directly modulating parameters of classifier, such as learning quadratic discriminant function (LQDF) [7]. To reduce computation complexity in large scale classification problem, usually only a part of parameters get modulated [8]. Apparently it would lead to suboptimal discriminative learning. Apart from these two kind methods, cascade MQDF [9] and modified boosting method [10] to some extent could be also categorized as discriminative learning methods in view of integration discriminant information in classifier learning. Besides the basic MQDF, cascade method gets extra MQDFs trained on selective subsets and fuses them in final recognition. In modified boosting, several MQDFs are trained by gradually enhancing weights of misclassified samples and reducing weights of correctively identified samples. Finally all MQDFs are integrated to construct a robust classification. In consideration of computation complexity, cascade MQDF model only fuses two MQDFs and modified boosting picks out the best MQDF to complete final recognition task. According to ideas of discriminative learning, it benefits classifier learning by enhancing weights of samples misclassified and reducing weights of sample recognized correctly because misclassified samples are important references for determining the optimal classification boundary. Some samples are recognized correctly, however they locate closely to classification boundary in feature space. Recognition of these samples are unstable and prone be contaminated by noise and parameter estimation bias. They could also provide references for determining classification boundary. That indicates that all samples close to classification boundary weights important no matter whether the recognition result is correct or incorrect.

A discriminative learning method for MQDF based on sample importance weights is proposed in this paper. MQDF parameters are indirectly modulated on weighted samples. The method is investigated on two free writing style Chinese handwriting sets and compared with cascade MQDF and modified boosting.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.222

The rest of this paper is organized as follows. Section 2 presents the proposed MQDF discriminative learning method. Section 3 and 4 detail sample importance and sample importance weight respectively. The experiments are then followed in section 5. Finally in Section 6 we summarize the paper.

II. MQDF Discriminative Learning

The basic idea of MQDF discriminative learning is to train MQDF by giving each sample an importance weight. It aims to adjust MQDF parameters indirectly to avoid large computation complexity and assimilates discriminative information at simultaneously. Block diagram of proposed discriminative learning method is illustrated in Fig. 1. Firstly, a basic MQDF is trained under MLE. It is employed to recognize all training samples and outputs recognition distance. Then on these distance based recognition results, sample importance function is derived under bayes decision rule. The sample importance means to what extent that the sample contributes to determining classification boundary. It is measured by extended recognition confidence. Details will be found in section 3 and 4. Importance function value of each sample is normalized to weight a sample. The mean and covariance matrix are re-estimated by (1)(2).

$$\mu_i = \sum_{j=1}^{N_i} \tilde{\pi}_{ij} x_j \tag{1}$$

$$\Sigma_i = \sum\nolimits_{j=1}^{N_i} \tilde{\pi}_{ij} (x_j - \mu_i)(x_j - \mu_i)^T \tag{2}$$

Where $\tilde{\pi}_{ij}$ is the normalized importance weight for j^{th} sample of class w_i. If all $\tilde{\pi}_{ij}$ s equal to each other, equation (1)(2) would reduce to MLE.

III. Sample Importance

In a classification task with C classes, $w_1, w_2, \cdots w_C$ are class labels. Under the bayes decision rule, a sample $x \in w_i$ is recognized as

$$w(x) = \underset{i=1,2,\cdots,C}{\arg\max}\, p(w_i \mid x)\,. \tag{3}$$

Where $w(x)$ is a class label indicating the recognition result. According to bayes formula, the a posteriori probability $p(w_i \mid x)$ could be transformed into product of the prior probability $p(w_i)$ and conditional probability $p(x \mid w_i)$. Without any prior knowledge, generally the prior probability of each class is assumed to be equally distributed thus the a posteriori probability is equivalent with the conditional probability. In discriminative learning process, recognition of a training sample $x \in w_i$ can be broken down into two stages. Firstly, to find the maximum conditional probability of x among its unreal label class distributions. w_{so} is the corresponding class label and defined as

$$w_{so} = \underset{j=1,2,\cdots,C,j\neq i}{\arg\max}\, (p(x \mid w_j))\,. \tag{4}$$

Secondly, to obtain the recognition result by comparison between $p(x \mid w_{so})$ and $p(x \mid w_i)$.

$$L(x) = \frac{p(x \mid w_{so})}{p(x \mid w_i)} = \begin{cases} \geq 1 & misclassified \\ < 1 & otherwise \end{cases} \tag{5}$$

If $0 \leq L(x) < 1$, it indicates that the sample is recognized correctly. If $L(x)$ is small then the sample is far from classification boundary and easy to be identified correctly. On the contrary if $L(x)$ approaches to 1, in other words that $p(x \mid w_{so}) \approx p(x \mid w_i)$, then the sample is prone to be misclassified as w_{so}. Apart from bayes error, it is largely due to parameter estimation bias under the condition of finite samples and sample's nonGaussian distribution. If $L(x) \geq 1$, it means that the sample has come over classification boundary and been misclassified.

Through the above analysis, it is confirmed that the value of $L(x)$ could reflects to what extent that the sample has been misclassified. Samples misclassified or recognized correctly but close to classification boundary are more important relatively. $L(x)$ and sample importance satisfy a monotonic function relationship. In this paper, the function is simply set as a deterministic function $\left(L(x)\right)^{\eta}$, Where η is a positive control constant. The importance function is derived in the following section.

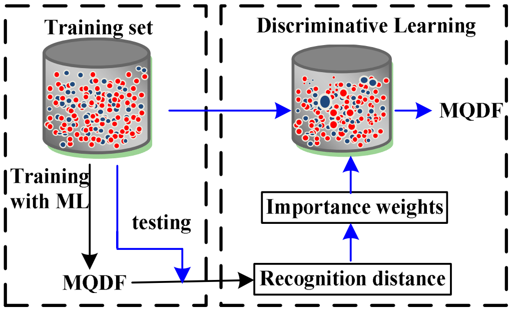

Figure 1. Block diagram of MQDF discriminative learning based on sample importance weights.

IV. Sample Importance Weight

As well known the conditional probability $p(x \mid w_i)$ and distance based recognition result d_i of MQDF satisfy

$$p(x \mid w_i) \propto e^{-d_i/2}\,. \tag{6}$$

Thus probability measure in $L(x)$ could be transformed into distance formation. Sample importance is defined as

$$\pi = \left(\frac{e^{-d_{so}/2}}{e^{-d_i/2}} \right)^{\eta} = e^{-\frac{1-d_i/d_{so}}{2/\eta d_{so}}} = e^{-\frac{1-d_i/d_{so}}{\sigma}} \tag{7}$$

Where $\sigma = 2/\eta d_{so}$ and could be regarded as a constant, which controls distribution variances of $1 - d_i / d_{so}$. When $1 - d_i / d_{so} > 0$, recognition result is correct and $1-d_i/d_{so}$ is defined as general recognition confidence[11], which is an effective measurement of recognition reliability. The equation $1-d_i/d_{so}=0$ corresponds to classification boundary between class w_{so} and w_i. If $1-d_i/d_{so}<0$, it means that the sample comes across the classification boundary and is misclassified. To avoid the negative effects on discriminative learning caused by these samples, $1-d_i/d_{so}$ is forced to zero. That means misclassified samples have been replaced at the boundary. Then the extended recognition confidence is formulated as follows

$$R=\begin{cases}1-d_i/d_{so} & if \quad w(x)=w_i \\ 0 & otherwise\end{cases} \tag{8}$$

Substitute (8) into (7), then sample importance function is expressed as $\pi(R)=e^{-R/\sigma}$, $R\in[0,1]$. σ and R both determine sample importance. Their characteristics and relationship with recognition performance are analyzed in the following. Distributions of R computed in different dimensional feature space are illustrated in Fig. 2. MQDF classifier is marked as MQDF(d ,k)，which indicates that the feature is compressed to d dimensional and k is the truncation dimensionality of covariance matrix. As illustrated, the number of misclassified samples is small and as the horizontal ordinate increases, statistics of R appear to have a decline trend after an initial ascent. As feature dimensionality increases the expected value of R decreases and approaches to zero. Recall from(8), computation of R has already integrated discriminative information.

Parameter σ fixes weighting mechanism. As shown in Fig. 3, importance functions are characterized by different values of σ. Determination of σ has significant relation to distribution of R. As illustrated in Fig. 3, if distribution of R is fixed, importance weight is a function of σ. In the limited case as $\sigma\to\infty$, all samples shared the same weights and naturally the trained MQDF performs the same as the one estimated by MLE. So the value of σ should not be too big. In the contrary limited case，if σ is too small, importance function is too steep over small R region. Recall from distribution of R, importance weights will be concentrated on only a few samples and most of samples get almost zero weights. Therefore it is a small sample size problem with very small σ. MQDF is a generative model estimated on samples. The sample number is closely related to recognition performance [12][13]. When faced with small sample size problem, minor eigenvalues of covariance matrix tend to be under estimated. Therefore σ is suggested not to be small. In this paper σ is studied experimentally.

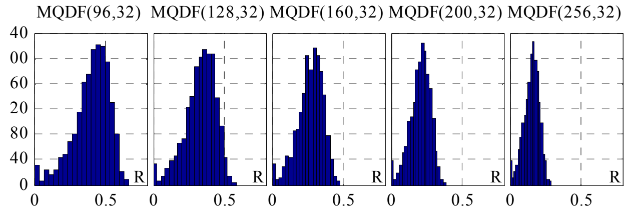

Figure 2. Histogram of R for the first character in GB2312-1980 level I set.

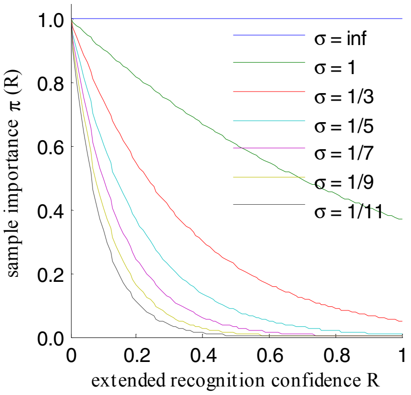

Figure 3. Sample importance function.

Let π_{ij} be sample importance of the j^{th} sample of class w_i， in order to keep the probability formation, sample importance weight is defined as the normalization of π_{ij}.

$$\tilde{\pi}_{ij}=\frac{\pi_{ij}}{\sum_l \pi_{il}} \tag{9}$$

V. EXPERIMENTS

The discriminative learning method proposed is investigated and compared with the other methods on THU-HCD sets, which is an offline Chinese handwriting character sample library. Each subset of THU-HCD database contains 3,755 simplified Chinese character classes. Training set contains 1877 subsets and testing sets contain two sets, denoted as A and B. They both are transformed from free style online Chinese handwriting and contain 170 and 100 subsets, respectively. Character samples are illustrated in Fig. 4.

Before feature extraction, every character image is normalized into 65×65 size. The normalized image is decomposed in to 8 directional templates. On each template, 7×7 features are extracted, thus in total 392 dimensional gradient features are obtained [14]. σ is optimized on the validation set, which is selected randomly from training sets. The basic MQDF is learned under MLE and the truncation dimensionality is set to 32.

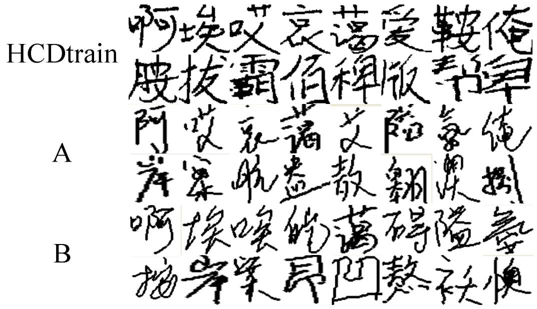

Figure 4. Samples of training set and test sets.

The first experiment investigates the relationship between accuracy and $1/\sigma$ on the validation set. Curves of σ to recognition accuracy are plotted as Fig. 5. As the figure shows that for the same classifier, as $1/\sigma$ increases, the recognition accuracy goes up and then shows a downward trend. As σ increases to a certain extent, samples obtain their most appropriate weights thus accuracy increases and gets to its peak. As σ continue to increase, weights have over concentrated on only a few samples. This leads to over fitted problem and give rise to performance degradation. For different MQDF classifiers, as feature dimensionality increases, the value of $1/\sigma$ corresponding to the highest recognition accuracy gradually rises. As mentioned previously, weights are dependent on both R and σ thus accuracy also has close relation to distribution of R. In higher dimensional feature space, distribution of R is more intent. Therefore, in higher feature space it needs a bigger $1/\sigma$ to alleviate the imbalance of sample importance weights.

The following experiment has investigated basic MQDF, cascade MQDF(C_MQDF), modified boosting(M_boosting) and the proposed method based on sample importance weight(SIW). The initial 392 gradient features are reduced to 96D, 128D, 160D, 200D, 256D by linear discriminant analysis (LDA)[15] respectively. Values of $1/\sigma$ are set as 7, 9, 11, 11, 11 for weighting functions correspondingly. Test results are listed in TABLE I and TABLE II.

The results show that all discriminative learning methods compared in the experiments gain the capability to improve performance of basic MQDF classifier. On test A, SIW has acquired the highest accuracy on all feature dimensionalities. Relative to basic MQDF performance, recognition accuracies have been increased by 11.22%，9.78%，9.55%，7.93% and 6.18% respectively. On set B, at 96D, cascade MQDF has achieved the highest accuracy. SIW gains a comparable accuracy with half recognition complexity of cascade MQDF. Except for $d=96$, SIW gives the highest recognition accuracies and outperforms both cascade MQDF and modified boosting. The results confirm that SIW is a promising method to improve performance of MQDF.

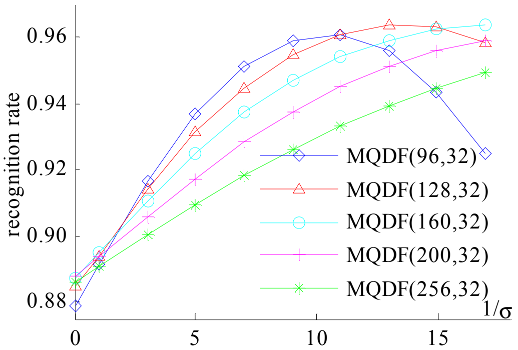

Figure 5. Relationship between $1/\sigma$ and recognition rate on validation set.

TABLE I. ACCURACY OF OFFLINE CHARACTER RECOGNITION ON A

Classifier	MLE (%)	C_MQDF (%)	M_boosting (%)	SIW (%)
MQDF(96,32)	78.37	89.46	88.46	**89.59**
MQDF(128,32)	79.24	88.58	88.09	**89.02**
MQDF(160,32)	79.56	86.88	87.11	**89.11**
MQDF(200,32)	79.61	84.68	85.82	**87.54**
MQDF(256,32)	79.37	82.01	84.24	**85.55**

TABLE II. ACCURACY OF OFFLINE CHARACTER RECOGNITION ON B

Classifier	MLE (%)	C_MQDF (%)	M_boosting (%)	SIW (%)
MQDF(96,32)	87.23	**88.93**	88.58	88.65
MQDF(128,32)	87.71	89.46	89.38	**89.59**
MQDF(160,32)	87.93	89.48	89.49	**89.79**
MQDF(200,32)	87.89	89.15	89.35	**89.70**
MQDF(256,32)	87.66	88.41	88.93	**89.22**

In conclusion, compared with other discriminative methods, the proposed method gains higher or comparable performance and lower recognition complexity. Furthermore, it provides a penetrating insight into the sample weighting mechanisms in classifier training process.

VI. CONCLUSION

This paper has proposed an MQDF discriminative learning method based on sample importance weight and compared with other outstanding discriminative learning methods. The results proved that rectifying classifier parameters based on sample importance weights could not only enhance classification performance but also reduce the computational complexity. Furthermore, the methods provide insights into the mechanisms of weighting samples in discriminative learning. The sample importance weights could be used in any other Gaussian distribution based models. The test sets are free writing style character sets, on which the recognition accuracy is much lower compared to the reports in literature. It means there is still a long way to go on this research. More theoretical investigation and systematic analysis will be carried out in future work.

ACKNOWLEDGMENT

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. 60772049, 60933010.

REFERENCES

[1] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, “Modified Quadratic Discriminate Functions and the Application to Chinese Character Recognition,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.9, Jan. 1987, pp. 149-153.

[2] C.J.C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining and Knowledge Discovery. Boston: Kluwer Academic Publishers, vol.2, 1998, pp. 955-974.

[3] J. Friedman, T. Hastie, and R. Tibshirani, “Additive Logistic Regression: a Statistical View of Boosting”, Annals of Statistics, vol. 28, 2000, pp. 337-407.

[4] Y. LeCun, L. Bottou and Y. Bengio, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol.86, Nov.1998, pp. 2278-2324.

[5] D. Decoste, B. Schölkopf, "Training Invariant Support Vector Machines," Machine Learning,. vol.46,2002, pp. 161-190.

[6] J.X. Dong, A. Krzyzak and C.Y Suen, "Fast SVM Training Algorithm With Decomposition on Very Large Datasets," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.27, Apr. 2005, pp. 603-618.

[7] C.L. Liu, H. Sako and H. Fujisawa, "Discriminative Learning Quadratic Discriminant Function for Handwriting Recognition," IEEE Trans. on Neural Networks, vol.15, Mar. 2004, pp. 430-444.

[8] H.L. Liu, X.Q. Ding, "Handwritten Character Recognition Using Gradient Feature and Quadratic Classifier with Multiple Discrimination Schemes," Proc. IEEE International Conference on Document Analysis and Recognition, IEEE Computer Society, Sep. 2005. pp. 19-23

[9] Q. Fu, X.Q Ding, T.Z. Li and C.S. Liu, "An Effective and Practical Classifier Fusion Strategy for Improving Handwritten Character Recognition," Proc. of International Conference on Document Analysis and Recognition, IEEE Computer Society, Sep. 2007, pp. 1038-1042.

[10] Q. Fu, X.Q. Ding, C.S. Liu, "Boosting Descriptive Model for Large Scale Classification and Its Application to Chinese Handwritten Character Recognition," Proc. of International Conference on Frontiers in Handwriting Recognition, 2008.

[11] X.F Lin, X.Q Ding and M. Chen, et al, "Adaptive Confidence Transform Based Classifier Combination for Chinese Character Recognition," Pattern Recognition Letters, vol.19, Aug. 1998, pp. 975-988.

[12] C. Sima, E.R. Dougherty, "The Peaking Phenomenon in the Presence of Feature-selection,". Pattern Recognition Letter, vol. 29, Jun. 2008, pp. 1667-1674.

[13] X.D Jiang, "Asymmetric Principal Component and Discriminate Analyses for Pattern Classification,". IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, May. 2009, pp. 931-937.

[14] C.L. Liu, K. Nakashima, H. Sako and H.Fujisawa, "Handwritten Digit Recognition: Investigation of Normalization and Feature Extraction Techniques", Pattern Recognition, vol. 37, Feb. 2004, pp. 265-279.

[15] C.R. Rao, "The Utilization of Multiple Measurements in Problems of Biological Classification," Journal of Royal Statistical Society B. vol. 10, 1948, pp. 159~203.