A Novel Multistage Classification Strategy for Handwriting Chinese Character Recognition Using Local Linear Discriminant Analysis

Lei Xu, Baihua Xiao, Chunheng Wang and Ruwei Dai

Laboratory of Complex System and Intelligent Science
Institute of Automation, Chinese Academy of Sciences
Zhongguancun East Rd, No.95, Beijing, 100080, P.R. China
lei.xu@ia.ac.cn

Abstract. In this paper we present a novel multistage classification strategy for handwriting Chinese character recognition. In training phase, we search for the most representative prototypes and divide the whole class set into several groups using prototype-based clustering. These groups are extended by nearest-neighbor rule and their centroids are used for coarse classification. In each group, we extract the most discriminative feature by local linear discriminant analysis and design the local classifier. The above-mentioned prototypes and centroids are optimized by a hierarchical learning vector quantization. In recognition phase, we first find the nearest group of the unknown sample, and then get the desired class label through the local classifier. Experiments have been implemented on CASIA database and the results show that the proposed method reaches a reasonable tradeoff between efficiency and accuracy.

1 Introduction

Handwriting Chinese character recognition (HCCR) is one of the most challenging topics in the fields of pattern recognition. There are three main factors that make HCCR difficult:

- Handwriting styles vary widely among individuals so that the boundaries between different classes are very complicated.
- The statistical features for HCCR are usually highly dimensional and the number of classes is very large.
- The actual HCCR system should possess the ability to *batch process* large amounts of documents efficiently.

To improve the classification accuracy, we should employ nonlinear classifiers with complex structure and higher-dimensional feature. However, such classifiers often result in great requirement for both computation and storage. From an application point of view, we hope to reach an acceptable tradeoff between accuracy and efficiency.

Linear discriminant analysis (LDA), as a dimension reduction technique, is usually utilized to alleviate the computation burden and speed up the recognition process. Furthermore, the criterion of LDA aims at maximizing the between-class variance while

I. King et al. (Eds.): ICONIP 2006, Part II, LNCS 4233, pp. 31–39, 2006.

simultaneously minimizing the within-class variance, so that more accurate classification can be achieved because the samples are rearranged in the reduced feature space.

Multistage classification [1,2,3,4], as shown in Fig. 1, is another effective strategy for recognition problems on large class set. The coarse classification often utilizes *cheap* feature and structure such that less computation is required. The candidate set is selected according to the output of the coarse classifier, and consequently the most matching classifier is constructed (or selected) for fine classification. The fine classifier will perform detailed comparison and provide the final class labels.

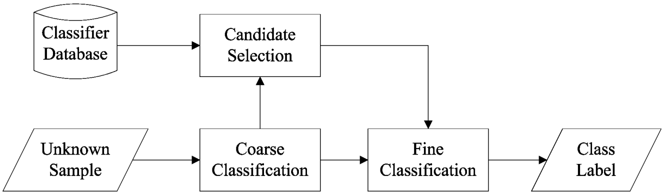

Fig. 1. Structure of multistage classification

It is natural to combine LDA and multistage classification for HCCR. In the coarse classification stage, all classes are involved, so that we perform LDA to obtain a global transformation matrix W by using all the training samples. In the fine classification stages, since only a part of classes are involved, such W isn't optimal any more and we need a local transformation matrix for each candidate set to provide locally discriminative information. However, theoretically speaking, for a system with n classes, the number of possible candidate sets is $2^n - 1$.

As is evident from the above discussion that we need an efficient candidate selection rule which can effectively exclude most of the redundant candidate sets, because we can't afford the burden of performing LDA online. Unfortunately, the existing rules, such as rank-based rule [3] and cluster-based rule [1,4], can not solve this problem.

In this paper, we present a novel multistage classification scheme for HCCR. The whole class set is divided into a set of subsets called groups, by clustering algorithms and extension rules. The adjacent groups overlap each other, and as a straightforward result, the nearest group to an unknown sample can be used as its candidate set. Since the number of groups is finite, we perform LDA for each group, so that the local discriminative features can be extracted and used for fine classification.

During the design process, we adopt a hierarchical learning vector quantization (LVQ) to improve the overall performance of our HCCR system. Firstly, we use the *global* LVQ to search for the most representative prototypes in the sense that the classification accuracy is highest. Such prototypes are used to initialize the desired groups. Secondly, we use the *group-based* LVQ to optimize the groups centroids such that the hit rate is highest. At last, after all groups have been decided, we use the *local* LVQ for each group to train the fine classifiers.

2 Group-Based Candidate Selection Rule

There are two key issues for the design of multistage classification. The first one is to improve the hit rate, which denotes the probability that the candidate set of an unknown sample's nearest region contains the correct class label. The second one is to improve the accuracy of each fine classifier. In our HCCR system, these problems are circumvented by group-based candidate selection rule and local linear discriminant analysis,respectively.

The goal of coarse classification is to decide the compact and accurate candidate set, and then construct the appropriate fine classifier. There basically exist two types of decision rules for candidate selection in the literature. Without loss of generality, we utilize the two-dimensional feature space to depict the related decision rules.

The first one is rank-based decision rule [3,5,6]. Namely, we select the classes with the highest scores based on the output of the coarse classifier and consequently construct the candidate set. An overwhelming advantage of this rule is that the fine classifier can always fix attention on the most confusable classes and exclude the irrelevant ones, as is shown in Fig. 2.

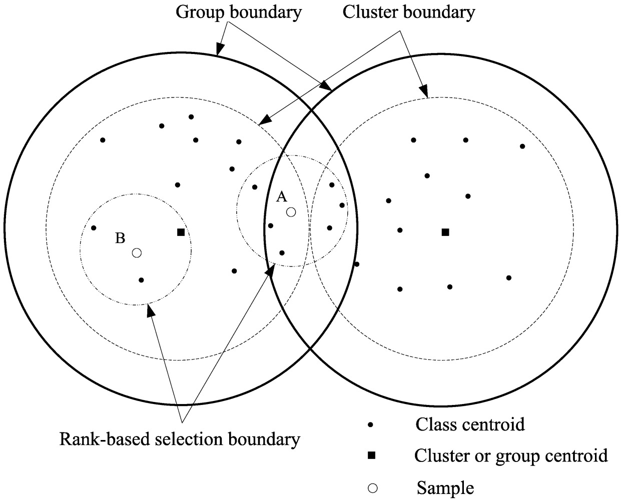

Fig. 2. Comparison of different decision rules

The second one is cluster-based decision rule [1,4]. We divide the whole class set into a number of unoverlapped clusters and utilize the cluster centroids to construct a distance-based coarse classifier. Then for an unknown sample, we choose *several* nearest clusters and use the related classes to build the candidate set. For example, sample *A* in Fig. 2 is located around the boundary between two clusters, so that both clusters should be selected to ensure high hit rate.

A common drawback of rank-based and cluster-based rule is that the number of the possible candidate sets is nearly infinite. Hence, there is little flexibility for designing an appropriate fine classifier for each group. For example, in [2], both the coarse and fine classification adopt nearest neighbor classifier. The only difference is that the fine classifier employs more samples for each class than coarse classifier.

In order to overcome this drawback, we propose the group-based decision rule. Groups are distinguished from clusters just because adjacent groups overlap each other. Such overlap is achieved by nearest-neighbor-based extension rule which will be described in the next section. For example, the boundary of the left group in Fig. 2 has been extended to the solid line. Under this circumstance, the nearest one group of an unknown sample may entirely contain its candidate set. Even though sample A is located around the boundary of the two clusters, its candidate set is still entirely contained in its nearest group (the left one).

Since the number of groups are definite, we can perform LDA for each of them, so that the most discriminative features can be extracted and the recognition rates of the fine classifiers can be greatly improved. More importantly, we can even employ different types of features and classifiers for each group.

Notes and Comments. If we excessively extend the groups, the overall hit rate will infinitely approach 100%. However, such high hit rare is meaningless because the corresponding group size will be very large. As a result, we should reach a tradeoff between the hit rate and the average size of the groups.

3 Hierarchical LVQ for Classifier Design

Considering the large class set and high-dimensional features, we employ distance-based nearest prototype classifiers (NPC(k,C)) in our HCCR system [7], where the parameter k represents the number of prototypes for each class and C is the number of classes. It is pertinent to point that our multistage classification scheme doesn't inherently prohibit other types of classifiers.

The methodology for NPC design can be found in [7,8,9]. In this paper, we employ the GLVQ algorithm due to its superior performance [7,10]. We will not describe this algorithm in detail in this paper.

3.1 Step 1: Prototype Abstraction Via Global LVQ

We first calculate the global transformation matrix W using the LDA algorithm in [11], and the rest work of step 1 and step 2 will be based on the lower-dimensional space.

The main task of step 1 is to initialize the group centroids. An intuitive choice is sample-based clustering algorithm. However, through preliminary experiments we have found that this approach suffers from slow convergence and is sensitive to the outliers in the training set. In order to overcome these drawbacks, we employ prototype-based approach. Namely, we design a NPC(K,C), and directly use the corresponding prototypes for clustering. Considering the different handwriting styles, $K > 1$ is necessary. The prototypes of each class are initialized using k-means clustering algorithm. Then the whole prototype set is trained by global LVQ algorithm, where *globe* means that the optimization process is based on the whole class set.

The NPC(*K,C*) designed in this step will play another role when a sample falls into the risk zone. In this condition, we utilize the cluster-based rule instead to decide the candidate set and extract the corresponding prototypes for fine classification.

3.2 Step 2: Group Extension and Group-Based LVQ

The task of step 2 is to extend the groups by nearest-neighbor (NN) rule and then optimize the centroids using supervised group-based LVQ. *Group-based* LVQ means that the objects of this optimization process are the group centroids.

To make the GLVQ algorithm meaningful, we have to define the group label of a sample before the training procedure. Since the adjacent groups overlap each other, a training sample may simultaneously belong to several groups, so that its group label is not unique.

Definition 1. *The group label* Z_t *of a sample* (x_t, y_t) *is the union of indices of the groups that contain its class label. Namely,* $Z_t = \{i \,|\, y_t \in G_i\}$.

The whole procedure for phase 2 is described as the pseudocode in Algorithm 1, where the function $NNeighbor(x, P, M)$ returns the indices of the M nearest neighbors of x in P, and G_i is the union of the indices of the classes that belong to group i. The elements in G_i are arranged in increasing order.

Algorithm 1. Pseudocode for step 2

```
Input: prototypes $P$   // $p_{(i-1)*K+1}, \cdots, p_{iK}$ belong to class $i$
Output: group $G_i$ and group centroids $g_i$,$1 \le i \le L$, where $L$
        is the number of groups
$\{g_1, \cdots, g_L\} = kmeans(P, L)$;
For $i = 1$ to $L$
    $G_i = \Phi$ (empty set);
End
For $i = 1$ to $CK$
    $q = \arg\min_j d(p_i, g_j)$;
    $\{b_1, \cdots, b_M\} = NNeighbor(p_i, P, M)$;
    For $j = 1$ to $M$
        $G_q = G_q \cup \{ceil(b_j/K)\}$;
    End
End
Repeat
    For $i = 1$ to $N$
        $k = \arg\min_{j \in Z_i} d(x_i, g_j)$;
        $l = \arg\min_{j \notin Z_i} d(x_i, g_j)$;
        $Update(g_k, g_l)$;      //using GLVQ algorithm [10]
    End
    $CalculateHitRate()$;
    $flag = IsConvergent()$;
Until $flag == True$
```

3.3 Step 3: Fine Classifier Design Via Local LVQ

After the groups have been decided, We should design the fine classifier for each group. Since the average size of these groups are much smaller than that of the whole class set, each group can be independently treated as a simple pattern system.

In our system, the fine classifier for each group is NPC(1). One prototype for each class is enough and can guarantee satisfying classification accuracy. We initialize the prototypes using the corresponding class means and adjust them by local LVQ, where *local* means that the optimization process is based on a subset. For each group, we repeat step 1 and obtain the local transformation matrix W_i and prototype set Q_i.

4 Recognition with Risk-Zone Rule

For an unknown sample, we first find its nearest group and then decide its class label within this group. However, as is mentioned above that we don't excessively extend each group to avoid too large group size. Therefore, if the sample is located just around the boundary of two groups, there still exist the possibility that its nearest group doesn't contain the correct class label, which will then result in incorrect classification.

To avoid such errors, we introduce the risk-zone rule. Namely, if a sample (x, y) falls into a *window*

$$min(\frac{d(x,g_k)}{d(x,g_l)}, \frac{d(x,g_k)}{d(x,g_l)}) > \eta$$

where g_k and g_l are two nearest group centroids and $0 < \eta < 1$ is a threshold, we will think that this sample is located in the risk zone. Under this circumstance, we utilize the cluster-based rule rather than the group-based one to select its candidate set. The whole algorithm for recognition is described as the pseudocode in Algorithm 2, where P_i is the subset of P that belongs to class i.

Note that the reasonable interval for η is $[0.93, 0.98]$. Although the hit rate will monotonously rise when η decreases, however, if η gets rather small, the group-based rule will degrade to the cluster-based one.

5 Experimental Results and Analysis

We conduct the experiments on a large handwriting Chinese character database collected by the Institute of Automation, Chinese Academy of Sciences. This database contains 3755 Chinese characters classes of the level 1 set of the standard GB2312-80. There are totally 300 samples for each class, and we randomly select 270 of them for training and the rest for testing. The experimental results provided in this section are all based on the test set. In preprocessing, each character image is linearly normalized to 64×64 size and then the 896-dimensional hierarchical periphery run-length (HPRL) feature is extracted.

We first perform an experiment to compare the performance of class-based rule and cluster rule. We utilize a nearest mean classifier (NMC) for coarse classification and a NPC(4,C') for fine classification, where C' is the size of the corresponding candidate set. In other words, after the candidate set is decided, we extract the prototypes that belong to the candidate set and construct the fine classifier. The experimental comparisons of these two rules can be seen from Fig. 3.

Compared with rank-based decision rule, cluster-based rule needs less computation in coarse classification, but the resulted candidate sets contain many redundant classes. On the contrary, the resulted candidate set of class-based rule is more compact and precise, so that the ultimate recognition rate is higher.

We conduct another experiment to validate the group extension method by varying the parameter M in algorithm 1. The resulted hit rate and the average group size are plotted in Fig. 4

Algorithm 2. Pseudocode for Recognition

```
Input: unknown sample $x$
Output: class label $y$
$x' = Wx$;
$\{a, b\} = NNeighbor(x', G, 2)$;    // $d(x', g_a) < d(x', g_b)$
If $\frac{d(x',g_a)}{d(x',g_b)} < \eta$
    $x'' = W_a x$;
    $q = NNeighbor(x'', Q_a, 1)$;
    $y = G_a(q)$;
Else
    $P' = \Phi$;  $G' = \Phi$;
    $\{b_1, \cdots, b_T\} = NNeighbor(x', G, T)$;
    For $i = 1$ to $T$
        $G' = G' \cup G_{b_i}$;
    End
    For $i = 1$ to $|G'|$
        $P' = P' \cup P_{G'(i)}$;
    End
    $q = NNeighbor(x', P', 1)$;
    $y = G'(ceil(q/4))$;
End
```

In application we choose $M = 33$ and then optimize the corresponding group centroids by group-based LVQ. The resulted hit rate is 98.91%. Moreover, if the risk zone rule is used with the threshold $\eta = 0.95$, the ultimate hit rate will exceed 99.47%.

The detailed results about the recognition rate and processing speed (millisecond per character) on the whole test set are listed in table 1.

From table 1 we can conclude that the proposed method with risk zone rule yields the best tradeoff between processing speed and classification accuracy. The superior recognition rate should be mainly ascribed to the local LDA which can extract the most discriminative features for the local subset. We also note that processing speed of the proposed method is much lower than that of the cluster-based method. The main reason is that the adjacent groups overlap each other so that their average size is larger. Furthermore, the second transformation before fine classification will also cost additional computation.

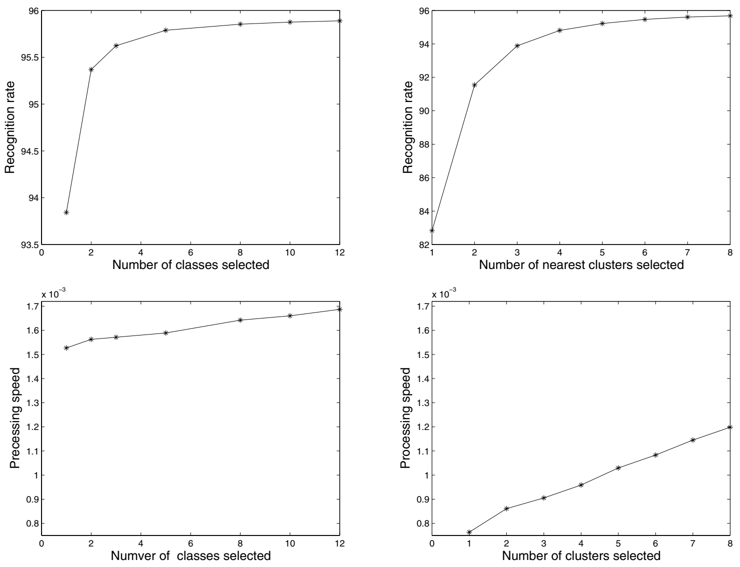

Fig. 3. The recognition rate and processing speed (seconds per character) (the left for rank-based one and the right for cluster-based one)

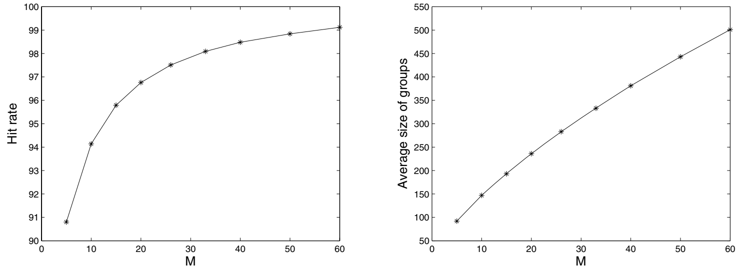

Fig. 4. The hit rate and the average group size for different M of algorithm 1

Table 1. Comparison of different methods

Classifier	Recognition rate	Processing speed
NMC	93.85	1.5269
NPC(4,3755) (trained by global GLVQ)	95.92	4.1811
class-based rule (top 8 classes selected)	95.85	1.6423
cluster-based rule (top 5 clusters selected)	95.22	**1.0297**
proposed method (without risk-zone rule)	97.30	1.8642
proposed method (with risk-zone rule,$\eta = 0.95,\ T = 3$)	**97.65**	1.7665

6 Conclusion

A novel multistage classification strategy for HCCR has been proposed in this paper. The basic idea of the proposed method is to divide the whole class set into overlapped groups such that the nearest group of a sample entirely contains its candidate set. Compared with conventional methods, our proposed method only allow one group for each unknown sample. Since the number of groups is finite, it is applicable to perform local LDA for each group. As a result, the most discriminative feature can be extracted and more accurate classification can be achieved within each group.

During the design phase, we utilize the hierarchical LVQ as a powerful tool to optimize the global prototypes, centroids and local prototypes. Experimental results show that this method can greatly improve the overall performance of our HCCR system.

Considering that the overall hit rate on test set is lower than 99%, we have introduced the risk zone rule. If samples fall into the risk zone, we use the cluster-based rule to construct the fine classifier. At the sacrifice of locally discriminative information, a large proportion of such sample can be correctly classified.

However, we find that even if the risk zone rule is utilized, the overall hit rate is still less than 99.5%. Hence, we should find more efficient group extension rules which will yield more compact groups and higher hit rate.

References

1. Liu, C.L., Mine, R., Koga, M.: Building Compact Classifier for Large Character Set Recognition Using Discriminative Feature Extraction. In: Proceedings of the Eighth International Conference on Document Analysis and Recognition. Volume 2., IEEE (2005) 846–850
2. Rodrguez, C., Soraluze, I., Muguerza, J.: Hierarchical Classifiers Based on Neighbourhood Criteria with Adaptive Computational Cost. Pattern Recognition **35**(12) (2002) 2761–2769
3. Liu, C.L., Nakagawa, M.: Precise Candidate Selection for Large Character Set Recognition by Confidence Evaluation. IEEE Transaction on Pattern Analysis and Machine Intelligence **22**(6) (2000) 636–642
4. Tseng, Y.H., Kuo, C.C., Lee, H.J.: Speeding up Chinese Character Recognition in an Automatic Document Reading System. Pattern Recognition **31**(11) (1998) 1601–1612
5. Horiuchi, T.: Class-selective Rejection Rule to Minimize the Maximum Distance between Selected Classes. Pattern Recognition **31**(10) (1998) 1579–1588
6. Ha, T.M.: The Optimum Class-selective Rejection Rule. IEEE Transaction on Pattern Analysis and Machine Intelligence **19**(6) (1997) 608–615
7. Liu, C.L., Nakagawa, M.: Evaluation of Prototype Learning Algorithms for Nearest-neighbor Classifier in Application to Handwritten Character Recognition. Pattern Recognition **34**(3) (2001) 601–615
8. Kuncheva, L.I., Bezdek, J.C.: Nearest Prototype Classification: Clustering, Genetic Algorithms, or Random Search? IEEE Transaction on Systems, Man and Cybernetics-Part C **28**(1) (1998) 160–164
9. Veenman, C.J., Reinders, M.J.T.: The Nearest Subclass Classifier: A Compromise between the Nearest Mean and Nearest Neighbor Classifier. IEEE Transaction on Pattern Analysis and Machine Intelligence **27**(9) (2005) 1417–1429
10. Sato, A., Yamada, K.: Generalized Learning Vector Quantization. In: Advances in Neural Information Processing Systems. Volume 7., MIT Press (1995) 423–429
11. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley (2000)