4N−06

A Study on Feature Selection and its Application to the Recognition of Handwritten Japanese Characters

Mutalifu Abulimiti Manabu Ichino
Graduate School of Science and Engineering, Tokyo Denki University

1. Introduction

Feature selection is an important task in the design of pattern classifiers since the success of a classifier depends on the kind of classifier and also on the features the classifier operates on. It has been the focus of interest for quite some time, and much works have been done and applied into various fields. These have shown that feature selection is essentially important in any pattern classification problem. The main goal of feature selection is to select a subset of n features from the given set of d features, n<d, without significantly degrading the performance of the recognition system. To achieve this, we eliminate a feature if it gives us little or no additional information beyond that subsumed by the remaining features. This is what we call redundant features. In many applications, like the recognition of handwritten Japanese characters, the size of a data set is so large that it might not work well before removing unwanted features. Decrease in the number of redundant features drastically reduces the running time of the algorithm.

We approach a simple and general method of feature selection, called active feature selection, by describing its application to the offline recognition of handwritten Japanese characters. The main idea is trying to actively select effective features from the extracted feature vector of both unknown patterns and learning patterns, then carry out the recognition process by only using the selected features. By selecting features from the stable part of a character and from the most distinctive part of a class, we can reduce the effect of the irregularities of strokes of handwritten characters.

2. Theoretical Examination

Automatic recognition of patterns will be worthwhile in practice if the number of patterns to be recognized, the required speed, and the required rate of recognition are too great for available human effort.

To get a high rate of recognition, most researchers try to increase the number of features. But experiments indicates that the increase in the number of features not always lead the increase of recognition rate. It is because there is a tradeoff between the generality of class description and the interclass distinguish-ability [1]. Here, distinguish-ability indicates the ability of deciding an unknown pattern belongs to which pattern class. If the interclass distinguish-ability is high, it means we can describe the samples belongs to the class in details. The generality of class description indicates the ability of describing the general samples in a pattern class. If the generality of class description is high, it means we can describe the characteristics in which many samples in the class sharing with. So, In fact, the increase in the number of features will also lead to the increase of redundant features, the ability relative to an unknown pattern will be down. So it is impossible always increase the recognition rate by simply increase the number of features.

3. Active Feature Selection

Recognition is the assignment of a pattern, or at least part of a pattern, to a class. For example, in the case of handwritten characters, the pattern class “A” generally contains a very large number of somewhat different patterns that we call the samples of the class. They will differ from each other to a limited extent in shape, orientation and so on. There may be a mathematical definition of a pattern class of circles, but practical pattern classes of characters, fingerprints, photographs, and so on, have no known analytical definition. Furthermore, it is generally impractical to define a pattern class by storing all the constituent patterns, because there are far too many. What we actually do is to hypothesize that a particular definition, which will yield tolerably low recognition error rates in practice. A popular hypothesis has been that for each pattern class we can choose (defining) one specimen pattern that serves as a reference pattern. In the case of character recognition we call it as dictionary pattern. An unknown pattern can be cross-correlated with all of the

reference patterns and assigned to the class of the reference pattern that yields the highest correlation score.

4. An Experimental Study

Now we explain the active feature selection method by describing its application to the recognition of handwritten characters. The main algorithm is trying to select the features that are judged to be effective, from the feature vector of a test sample and recognizing the test sample by only using the selected features. For instance, to distinguish "は", "ぱ" and "ば" from each other, we can easily just compare the right-upper part of them, instead of comparing the whole structure. In this case, we try to select the features extracted from the right upper part of these characters as effective features and using only the selected features to apply the recognition process. The effective features selected will actively change according to the different test samples. That is why this method is called active feature selection method.

We try to actively select the effective features from the feature vector of both test samples and training samples. By using the information obtained from dictionary pattern, we can find out the most stable part of each class and select the features extracted from that part of the character as the effective features too. Then carry out the recognition process by only using the selected features. By selecting features from the stable part of a character and from the most distinctive part of a class, we can reduce the effect of the irregularities of strokes.

On the other hand, most researchers try to extract features by using several feature extraction methods simultaneously. We also test this method by compounding feature extraction methods [2]. In this case recognition rate is higher than that in the case of just using one kind of feature extraction method. It can be considered to be that they compensate for weakness of each other. But, there is almost no difference between the recognition rate in the case of using two kinds of methods and in the case of using three or more.

The algorithm was tested with 71 classes of Hiragana characters (160 samples/class) in ETL8 except the small characters "っ", "ゃ", "ゅ", "ょ". Among the 160 samples of each class, we took half (80) of the samples as training samples and the rest as test samples. The directional element feature was used to extract features. The extracted feature vector is a 196 dimensional vector. The result of the experiment is as shown in Figure.1.

Rate of recognition obtained was 90.8% and the number of the features used in the recognition process is 7.4. In comparison with the case of before applying the active feature selection, the rate of recognition increased by 2% and the number of the features decreased by 91.1%. This indicated that if we could select the effective features, it is possible to increase the recognition rate and decrease the number of features at the same time. This proved the effectiveness of the proposed algorithm.

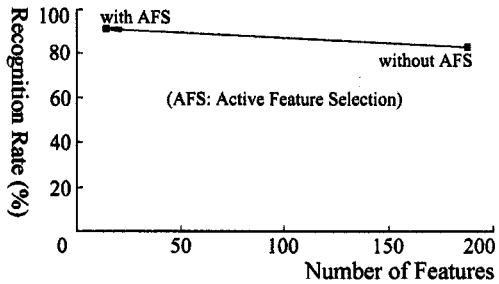

Fig.1.Recognition Rate and the Number of the Features

4. Concluding Remarks

We presented an algorithm for pattern classification based on active feature selection. The main achievement of the new algorithm is that if a large number of features are decreased, it achieves not only faster classification, but also higher reliable recognition.

Encouraging results with the new algorithm have been obtained and we want to continue the work of improvement to get better results.

References

[1] M. Ichino, H. Yaguchi, *An apparent simplicity appearing in pattern classification problems*, Pattern Recognition 33 (2000), 1467-1474

[2] M. Ablimit, M. Ichino, *Active Feature Selection Method and its Application to the Recognition of Handwritten Character Recognition*, OSDA 2000, Brussels, Belgium, 2000.