

Distance classifier ensemble based on intra-class and inter-class scatter

Yaqin Guo

Electrical and Energy Engineering College, Nantong Institute of Technology, Nantong, China

Abstract. Distance classifier ensemble method based on Intra-class and Inter-class Scatter is proposed in this paper. By Bootstrap technology, the training samples are sampled repeatedly to generate several subsample set, define Intra-class and Inter-class Scatter matrix with subsample set, train subsample set with scatter matrix, generate individual classifier. In the classifier ensemble, the results are integrated with the relative majority voting method. Experiment is tested on UCI standard database, the experimental results show that the proposed ensemble method based on Intra-class and Inter-class Scatter for distance classifier is effective, and it is superior to other methods in classification performance.

1 Introduction

In the field of pattern recognition, years of practical experience shows that, it is difficult to obtain satisfactory recognition performance by a the single method for complex pattern recognition problems[1-3]. However, the different classification methods are complementary to each other, and integrating multiple classifiers can reduce recognition errors and enhance robustness. So multiple classifier ensemble method has become a hot topic for researchers.

Generation and combination of individual classifier are concluded in the multiple classifier ensemble. In the individual classifier generation, Boosting[4]and Bagging[5] are important method. Using Boosting technology, the weight needs to be established for each sample in the training set, when misclassified probability of the sample is high, and increase sample weight, but the method is unstable. Using Bagging technology, individual classifier training samples are selected from the original training set. Training sample numbers is usually comparable to the original training set, and the training samples are selected repeatedly. So the original training set sample may appear multiple times in the new training set, or not appear at all. By reselecting the training set, the Bagging technique increases classifier integration difference degree, and improve the generalization ability.

By Bagging technology ,distance classifier ensemble method is proposed in this paper. The training samples can be sampled repeatedly to generate several different subsample sets, the subsample set is used to establish the Intra-class and Inter-class scatter matrix, train subsample set with scatter matrix, generate individual classifier. In the classifier ensemble, the results are integrated with the relative majority voting method. Experiment is tested on UCI standard database, the paper method is compared with minimum distance classifier integration method. In addition, the performance of single minimum distance

classifier and single minimum distance classifier based on Intra-class and Inter-class scatter are compared with the proposed method, the experimental results show that the proposed ensemble method is effective.

2 Minimum distance classifier based on intra-class and inter-class scatter

Set pattern classes is $\omega_1, \omega_2, \dots, \omega_c$, the number of pattern classes ω_i is n_i , $X_{is} \in \omega_i$ ($i = 1, 2, \dots, c$, $s = 1, 2, \dots, n_i$) is training sample, and eigenvectors is $X_{is} = (x_{is}^1, x_{is}^2, \dots, x_{is}^d)^T$, the dimension is d , the mean vector of pattern classes ω_i is $m_i = (m_i^1, m_i^2, \dots, m_i^d)^T$ ($m_i = \frac{1}{n_i} \sum_{s=1}^{n_i} X_{is}$). The weigh is

$W = [w_1, w_2, \dots, w_d]^T \in R^d$. The population sample mean vector is $m = \frac{1}{c} \sum_{i=1}^c m_i$.

Define Intra-class scatter matrix S_i , define total Intra-class scatter matrix S_w .

$$S_i = \sum_{x \in \omega_i} (x - m_i)(x - m_i)^T \quad i = 1, 2, \dots, c \quad (1)$$

$$S_w = \sum_{i=1}^c S_i \quad (2)$$

Define Inter-class scatter matrix S_b .

$$S_b = \sum_{i=1}^c (m_i - m)(m_i - m)^T \quad (3)$$

In the classification, the same class is expected to be as compact as possible, and different classes are expected to be as dispersed as possible. Intra-class scatter is expected to be as small as possible, Inter-class scatter is expected to be as big as possible. So define Criterion function $J(W)$.

$$J(W) = \arg \max_w \frac{W^T S_b W}{W^T S_w W} \quad (4)$$

Equation (4) is generalized Rayleigh quotient. It can be solved by Lagrange multiplier method[6], define Lagrange function $L(W, \lambda)$.

$$L(W, \lambda) = W^T S_b W - \lambda(W^T S_w W - c) \quad (5)$$

Take the Equation (5) partial derivative with respect to W .

$$\frac{\partial(L(W, \lambda))}{\partial W} = S_b W - \lambda S_w W \tag{6}$$

Let the partial derivative be equal to 0, then get Equation (7).

$$S_b W = \lambda S_w W \tag{7}$$

Solve for W in equation (7), that is to solve the eigenvalue problem of general matrix. According to Lagrange multiplier method, get W^* .

$$W^* = S_w^{-1} \sum_{i=1}^c (m_i - m) \tag{8}$$

Let $X = (x_1, x_2, \dots, x_d)^T$ is a identified sample, define distance $d_i(x)$ from X to mean value $m_i (i = 1, 2, \dots, c)$.

$$d_i(x) = W^T P^{(i)}(x) \quad i = 1, 2, \dots, c \tag{9}$$

where $P^{(i)}(x) = [(x_1 - m_i^1)^2, (x_2 - m_i^2)^2, \dots, (x_d - m_i^d)^2]^T$. According to the principle of distance minimization, the recognition decision of identified sample X is as follows.

$$e(x) = \arg[\min_{1 \leq i \leq c} d_i(x)] \tag{10}$$

3 Classifier ensemble

3.1 Individual classifier generation

Let training set $TR = \{(x_i, y_i) \mid i = 1, 2, \dots, l\}$, by Bootstrap technology, samples are extracted randomly from TR to form subsets $TR_k \{TR_k \mid k = 1, 2, \dots, K\}$, samples may be selected repeatedly. Therefore, the samples x_i of training set TR may appear multiple times in the new subsample set, while the samples x_j may not appear at all. Subsample set numbers is comparable to the original training set, and the repeated training samples increase subsample set difference. According to the method in section 2, training the minimum distance classifier with the K subsample set, get K individual classifiers, then integrate classification results.

3.2 Individual classifier ensemble

In the classifier ensemble, absolute majority voting and relative majority voting are two types of voting method. In the absolute majority voting method, when the number of class j is the largest and more than half of the total votes, the result is class j. In the relative majority voting, when the number of class j is the largest, the result is class j.

The relative majority voting is used in the paper. Assuming sample X to be identified, the output result of the K classifier is $e_k(x)$. For classifiers $e_k(x) = i$, define a binary function.

$$T_k(x \in \varpi_i) = \begin{cases} 1, & \text{if } e_k(x) = i \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

Sum the binary function.

$$E_i(x) = \sum_{k=1}^K T_k(x \in \varpi_i) \quad (12)$$

The final integration decision of the classifier is equation (13).

$$E(x) = \underset{i}{\operatorname{arg\,max}} E_i(x) \quad (13)$$

4 Experiment result

4.1 Experiment data

To verify the effectiveness of the algorithm,UCI standard data[7] are used in the experiment, data set zoo, wine, wave and segment are selected. In order to verify the ensemble method effectiveness, the four selected data sets have multi-category large sample data sets and multi-category small sample data sets. The numbers of data sets, the number of class and properties are shown in Table 1.

Table 1. Data Sets.

| Data set | Data Set number | Class number | Attributes number |
|----------|-----------------|--------------|-------------------|
| zoo | 101 | 7 | 17 |
| wine | 178 | 3 | 14 |
| wave | 5000 | 3 | 22 |
| segment | 2100 | 7 | 19 |

4.2 Experiment result

In the experimental process, by Bagging technology , Bootstrap sampling from the original training set for 20 times, generate 20 subsample sets. Subsample set is randomly divided into 5 groups, two groups are used as test set, the other three groups are used as training sets. According to the Section 2 method, the three groups training sets were used to train classifier, and get 20 individual classifiers. Finally, the results are integrated by the method of relative majority voting. In the classifier integration test process, a five cross-validation method is used, so each group samples can be used as a test set.

To verify the effectiveness of the method, the paper method is compared with minimum distance classifier ensemble method(MDC-E). In addition, the performance of single minimum distance classifier(MDC) and single minimum distance classifier based on Intra-

class and Inter-class Scatter(IIC-MDC) are compared with the proposed method, the experimental results show that the proposed ensemble method is effective.

Data sets are tested by four different methods. The training set classification result are shown in Table 2. The test set classification result are shown in Table 3. From Table 2 and Table 3, we can see that the ensemble method is better than single classifier. The experiment results show that, the recognition rate of the proposed method in the paper has been further improved, no matter in training set or test set.

Table 2. Training set classification result.

| Method | zoo | wine | wave | segment | Average accuracy |
|---------------------|--------|--------|--------|---------|------------------|
| MDC | 0.8839 | 0.7227 | 0.8003 | 0.7390 | 0.7865 |
| IIC-MDC | 0.9787 | 0.9424 | 0.8205 | 0.8343 | 0.8952 |
| MDC-E | 0.8983 | 0.7274 | 0.8012 | 0.7259 | 0.7882 |
| The proposed method | 0.9804 | 0.9745 | 0.8571 | 0.8365 | 0.9121 |

Table 3. Test set classification result.

| Method | zoo | wine | wave | segment | Average accuracy |
|---------------------|--------|--------|--------|---------|------------------|
| MDC | 0.8486 | 0.7130 | 0.8005 | 0.7369 | 0.7748 |
| IIC-MDC | 0.9297 | 0.9319 | 0.8109 | 0.8331 | 0.8764 |
| MDC-E | 0.8649 | 0.7159 | 0.8013 | 0.7419 | 0.781 |
| The proposed method | 0.9631 | 0.9647 | 0.8541 | 0.8214 | 0.9008 |

5 Conclusion

Distance classifier ensemble method based on Intra-class and Inter-class Scatter is proposed in this paper. Firstly, by Bootstrap technology, the training samples can be sampled repeatedly to generate several subsample set. In the individual classifier generation, define Intra-class and Inter-class Scatter matrix with subsample set, train subsample set with scatter matrix, generate individual classifier. In the classifier ensemble, the results are integrated with the relative majority voting method. Experiment is tested on UCI standard database, the experimental results show that the proposed ensemble method is superior to other methods in classification performance.

In addition, the proposed method is based on the minimum distance classifier in the experiment, it is not limited to this, and the method in this paper is also applicable to other distance classifiers.

This work was financially supported by Nantong Science and Technology Project (GCZ19048), and supported by Nantong Polytechnic College Young and Middle-aged Scientific Research Training Project(ZQNGG209) , and supported by the Key Project of Nantong Polytechnic College Natural Science (2017002).

References

1. H. Pengfei, C. Maoguo, W. Tao. Face Recognition Algorithm Based on Improved Convolutional Neural Network and Ensemble Learning, *C. E.* **46**,2(2020).
2. W. Haochang, L. Yu, L. Bin, W. Min. Ensemble Feature Selection for Recognizing Co-Expression Patterns of Genes, *J.J.U.* **35**,5(2017).
3. L. Min, L. Hua, C. Maohua. An Adaptive Sub-fusion Integration Classification Method, *C. M. C.* **27**,4(2019).
4. Z. Liling, L. Chaofeng. Multiple-kernel learning based object tracking algorithm with Boosting and SVM. *C. E. A.* **54**,13(2018).
5. Z. Zonglin, W. Dangpin. An imbalanced data classification method based on Probability threshold Bagging. *C. E. S.* **41**,6(2019).
6. L. Renli, B. Yaqin. A splitting augmented Lagrangian method embedding in the BB method for solving the sparse logistic problem. *O. R. T.* **23**,2 (2019).
7. Frank A, Asuncion A. UCI machine learning repository [http://www.ics.uni.edu/mlearn/ML Repository](http://www.ics.uni.edu/mlearn/ML_Repository). CA: University of California, S. I. C. S. ,(2010).