2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics

A Weighted Minimum Distance Classifier Based on Relative Offset

Yuxing Wang, Yulian Jiang*
College of Electrical & Information Engineering, Southwest Minzu University
Chengdu, China
e-mail: 1098781541@qq.com, 987587478@qq.com

***Abstract*—In many classifiers based on supervised learning, the Minimum Distance Classifier (MDC) is widely used because of its simplicity and effectiveness. But it has the disadvantage of the low accuracy on classification. To improve the performance of the MDC, this paper proposed the Weighted Minimum Distance Classifier based on Relative Offset (WMDCRO). Under the condition of the Fisher criterion, this method defined the weights of each category attribute in terms of the relative distance. It analyzed and judged the degree of clustering about data in the feature space by using the relative offset. The experiment results show that the accuracy of this classifier is higher than that of the traditional MDC. Compared with the Naive Bayes classifier, the classification stability of this classifier is better.**

Keywords-relative offset; Fisher criterion; minimum distance classifier; weight; median

I. Introduction

Classification is a major research topic in supervised learning. Many classification methods and techniques are used to establish classification models, such as Decision Trees, Neural Networks, K-NearestNeighbor, Naive Bayes, Support Vector Machines, and Minimum Distance. Among these classifiers, the Minimum Distance Classifier (MDC) is fast and straightforward. It works very well on both basic and more complex recognition problems, and it is widely used in the fields of image processing, text classification, and face recognition [1-6]. However, one major drawback of this classifier is that ignoring the distribution of samples and the correlations among species may result in poor classification accuracy.

Domestic and foreign research on the MDC is more extensive, various methods and techniques have been proposed to enhance the performance of the MDC. One of the key ideas is to select more effective distance measures, such as the Standardized Euclidean distance, the Manhattan distance, the Mahalanobis distance, and so on. Lin proposed a new weighted minimum distance classifier, which uses the discriminately power and variance of features. The weights increase the interclass separability while they decrease the intraclass dissimilarity [7]. Another key idea is to trim the attributes or the datasets. The paper [8] presented a new approach for recognizing faces by combining principal component analysis (PCA) and MDC. The paper [9] presented a new KNN algorithm for classification. This method improves the classification accuracy by reducing the training set to relative equal density based on weighted distance and give weight to the attributes of time series.

In summary, domestic and foreign scholars have done some research on how to improve the performance of the MDC. However, there are still few studies on the method of enhancing the classification accuracy of the classifier based on the relative offset.

The main limitation of the MDC is that it is insensitive to differences in variance among categories. In this paper, we first introduce and analyze the concept of the relative offset, then establish a new formula based on the relative offset and the Fisher criterion [10-13]. By weighting the distances between the sample and the center vectors with proposed formula, a Weighted Minimum Distance Classifier based on Relative Offset (WMDCRO) was finally proposed. The experiment results show that the WMDCRO effectively improves the classification accuracy compared with the traditional MDC. When compared with the Naive Bayes classifier, the classification stability of this classifier is better.

II. Minimum Distance Classifier

The minimum distance classifier is simple and fast. It uses the distances between the input data and a set of center vectors in the feature space as the classification criterion. A new sample is then classified by finding the class that has the cluster with the minimum distance (Euclidean) to the new sample.

Assuming that the sample k is an n-dimensional vector, that is, the sample k has n kinds of attributes, and it can be expressed as $x^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, ..., x_n^{(k)})$. The number of the classes is m. $M_i = (m_{i1}, m_{i2}, ..., m_{in})$ is the center vector of the class i. It is gained by calculating the arithmetic mean of all attributes in each class. $d(M_i, x^{(k)})$ is the Euclidean distance between the sample k and the i-th center vector. When $d_i = \min\{d(M_i, x^{(k)})\}$, the sample k can be classified as the class i.

The MDC is simple in principle, easy to understand, and fast in calculation. However, it considers neither the variance within the category (the distribution of the samples) nor the covariance between categories (the correlation between different categories). So it is unreasonable in practical applications. The currently used solutions are as follows: (1) Reducing the attributes used for classification. (2) Trimming the training set, that is, deleting the samples in the training set that has little effect on the classification and then the remaining training samples are retained for classification. (3) Weighting the distance, by increasing the range constraints of class attributes, the accuracy of the MDC might be improved.

III. RELATIVE OFFSET

Concerning that an attribute of a category always has two opposite states, and the distribution of individuals always fluctuates around the normal value. Of course, due to the problem of genetic mutation and evolution, the normal value of this attribute may change. For different categories, under the same attribute, their normal values are always dissimilar.

Assuming that the normal value of the attribute A is c, and its fluctuation range is (c-a, c+b), x is the value of test sample, and the relative offset R can be expressed as formula (1).

$$R = \begin{cases} \frac{c-x}{a} & x < c \\ 0 & x = c \\ \frac{c-x}{b} & x > c \end{cases} \tag{1}$$

The relative offset R is a ratio, and its range is [-1, 1]. For the normal value c, we can choose the median, the mode, and the arithmetic mean, etc. However, in the data mining work, it seems difficult for us to predict the trend of data distribution. We need to discuss how to choose the right normal value before classifying.

For the arithmetic mean, it reflects the average level of a dataset. Although the arithmetic mean is often used to report the central tendencies, it is not a robust statistic, meaning that it is greatly influenced by extreme values in the dataset. For skewed distributions, the representativeness of the arithmetic mean might be affected.

For the mode, it reflects the concentration of a dataset, because it is calculated by counting, it is not susceptible to extreme values in the dataset, and it always indicates one of the most common tendencies.

For the median, it reflects the intermediate level of the dataset. Since it is obtained by sorting, changes in some data have less effect on the median. Besides, it is unaffected by the maximum or minimum value in the dataset. This property improves the representativeness of the median, and the median may be a better description of the central tendency. For skewed distributions, the representativeness of the median might be affected. However, considering the stability of the classifier, it is reasonable to choose the median as the normal value.

In addition, for continuous data, we suggest that the integral method can be used to obtain the normal value by averaging the area of the curved trapezoid to the bottom edge.

To better explain the relative offset, we choose the median as the normal value and use the relative offset to measure the quality of the classification. From the Fisher criterion, we can infer that better classification results are characterized by the largest interclass separability and the smallest intraclass dissimilarity. For an attribute A, we construct a function F(R) to measure the degree of divisibility of the training samples, and then weight the attribute. Such as formula (2).

$$F(R) = \frac{\mathrm{var}(me_i)}{\max\{\mathrm{average}(\sum_{j=1}^{p} |R_{ij}|)\}} \tag{2}$$

In this formula, p is the number of training samples in each class, me_i is the median of the i-th class. For the numerator, the variance of me_i is used to measure the degree of interclass separability. For the denominator, firstly, the mean of the absolute value of the relative offset is calculated and then take the maximum of these values to measure the intraclass dissimilarity.

It can be informed from the formula (2) that the larger value of F(R) represents the larger interclass separability and the smaller intraclass dissimilarity. And the classification work might be easily achieved by setting an appropriate threshold.

We use the iris dataset in the UCI database as an example. There are three kinds of irises in the dataset: Setosa, Versicolor, Virginica, and their four attributes are Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width. Table I shows part of this dataset. And the F(R) values of four attributes are calculated and then shown in Table II.

TABLE I. PART OF THE IRIS DATASET

class	*Sepal.Length/cm*	*Sepal.Width/cm*	*Petal.Length/cm*	*Petal.Width/cm*
setosa	*5.1*	*3.5*	*1.4*	*0.2*
setosa	*4.9*	*3*	*1.4*	*0.2*
setosa	*5.4*	*3.9*	*1.7*	*0.4*
versicolor	*7*	*3.2*	*4.7*	*1..4*
versicolor	*6.4*	*3.2*	*4.5*	*1.5*
versicolor	*5*	*2*	*3.5*	*1*
virginica	*5.8*	*2.7*	*5.1*	*1.9*
virginica	*6.5*	*3.2*	*5.1*	*2*
virginica	*7.2*	*3*	*5.8*	*1.6*

TABLE II. F(R) VALUES OF FOUR ATTRIBUTES

Attributes	*Sepal.Length*	*Sepal.Width*	*Petal.Length*	*Petal.Width*
F(R)	*1.42759*	*0.26046*	*11.62446*	*1.96656*

We can infer from the Table II that F(R) values of the Petal.Length and the Petal.Width are larger. Therefore, when classifying irises, the Petal.Length and the Petal.Width are supposed to be assigned larger weights.

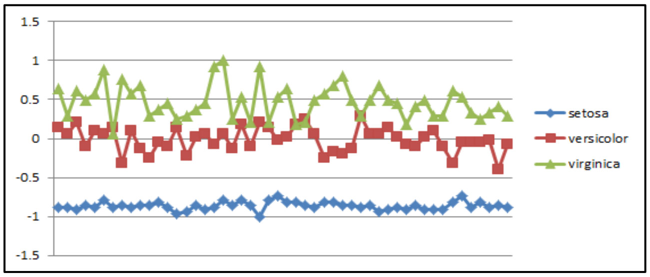

Figure 1. The relative offset distribution of the irises under the Petal.Length.

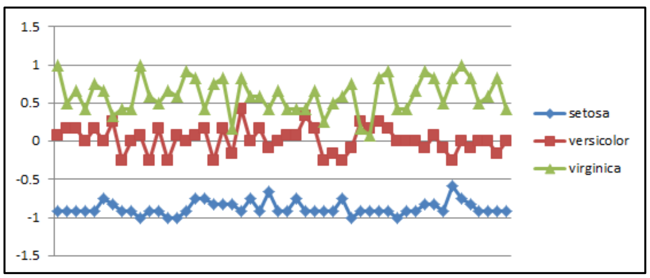

Figure 2. The relative offset distribution of the irises under the Petal.Width.

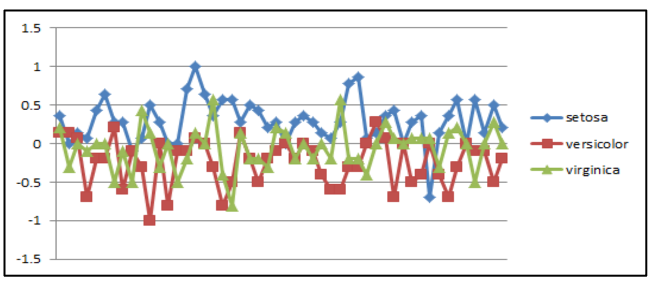

Figure 3. The relative offset distribution of the irises under the Speal.Width.

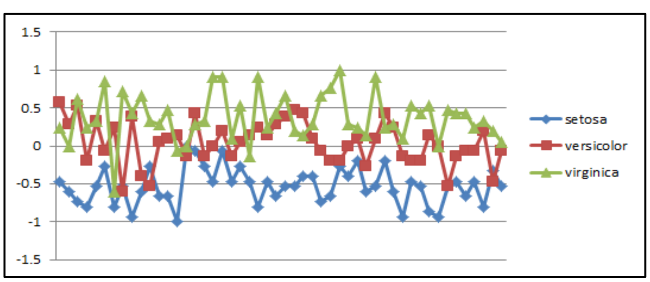

Figure 4. The relative offset distribution of the irises under the Speal.Length.

Figure 1 to Figure 4 visually show the distributions of irises under four attributes. It can be informed that most irises can be easily classified by setting some appropriate thresholds under the Petal.Length and the Petal.Width. However, for the Sepal.Width and the Speal.Length, it seems infeasible to classify three kinds of irises by setting some thresholds.

IV. Weighted Minimum Distance Classifier Based on Relative Offset

A. Training Steps

Step 1, divide the samples of the known categories into m sets. $a^{(i)} = (a_1^{(i)}, a_2^{(i)}, ..., a_n^{(i)})$ represents the sample in category A.

Step 2, under the first attribute, calculate the normal value, the maximum value, and the minimum value for each category. Here, the median is selected as the normal value. Of course, according to the actual situation, other parameters can also be selected, such as the mode and the arithmetic mean.

Step 3, repeat the step 2 to calculate remaining attributes. For each category, select the normal value of each attribute to generate its center vector $M_i = (m_{i1}, m_{i2}, ..., m_{in})$.

Step 4, calculate the weight ω_n of each attribute in the training set. According to the formula (1) of the relative offset and the formula (2) of the F(R) value, then the F(R) values of n kinds of attributes in the training set are obtained. Finally, normalize the F(R) values to obtain the weight of each attribute.

B. Classification Steps

Step 1, assuming that the test sample k in the test set is an n-dimensional vector, and the sample k can be expressed as $x^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, ..., x_n^{(k)})$.

Step 2, calculate the weighted distances between the test sample k and the center vectors separately and then obtain the set of the weighted distances $Q_i (i = 1, 2, ..., m)$. Finally, the minimum value of these distances is selected as the classification result and the calculation process of Q_i is expressed as follow:

$$Q_i = \sqrt{\sum_{j=1}^{n} \omega_j (x_j^k - m_{ij})^2} \tag{3}$$

V. Experiment and Analysis

In order to verify the validity of the WMDCRO, an experiment is designed. We chose datasets of different sizes and attributes, and the experimental data are from the Iris dataset, Glass dataset, Wine dataset and Weather dataset in Weka software and UCI database. Table III lists some information about these data, such as the number of samples, the number of classes, and the number of attributes (excluding class attributes).

The purpose of this experiment is to compare the classification accuracy of the minimum Euclidean distance classifier, the Naive Bayes classifier, and the WMDCRO in different datasets. In order to test the accuracy of three classifiers better, each classifier is tested ten times on the randomly divided datasets and takes the average of ten test results as the accuracy rate for each classifier. Table IV lists the results of this experiment. And in the last line, the average classification accuracy of each classifier on four datasets is also listed.

TABLE III. DESCRIPTION OF THE DATASETS

Datasets	*Number of samples*	*Number of classes*	*Number of attributes*
Iris	*150*	*3*	*4*
Glass	*214*	*6*	*9*
Weather	*14*	*2*	*4*
Wine	*178*	*3*	*13*

TABLE IV. EXPERIMENT RESULTS

Datasets	*Naive Bayes classifier*	*MDC*	*WMDCRO*
Iris	*96.02%*	*95.86%*	*96.87%*
Glass	*48.59%*	*50.53%*	*55.35%*
Weather	*56.37%*	*64.47%*	*68.45%*
Wine	*83.34%*	*74.31%*	*78.04%*
Average accuracy	*71.08%*	*71.29%*	*74.67%*

Table IV indicates that the average accuracy of the WMDCRO is about 3.38% higher than that of the MDC, and about 3.59% higher than that of the Naive Bayes classifier. In addition, it is worth noting that due to the increase of the number of the attributes, the classification accuracy of the three classifiers decreased. From Iris dataset to Glass dataset, when the number of attributes increases by five, the accuracy of the WMDCRO decreased by 41.52%, the accuracy of the Naive Bayes classifier decreased by 47.43% and the accuracy of the MDC decreased by 45.33%. In other datasets, WMDCRO also shows good stability.

For the Naive Bayes classifier, although it has stable classification efficiency and is suitable for incremental training, it assumes that the attributes are independent of each other. This assumption is often inappropriate in classification works. For the classification efficiency, the Naive Bayes classifier performs well for small-scale data processing and can handle multi-classification tasks. However, when dealing with large-scale data, because it needs to calculate conditional probability distribution, the time complexity will be very high, and there is no doubt that it will bring a big load to the computer.

For the WMDCRO, it not only considers the distribution of samples within the same category but also considers the relationship between different categories under the same attribute. Finally, the different attributes are linked by weighting. In addition, the classification accuracy of the WMDCRO is relatively less affected when the number of attributes is large. Compared with the traditional minimum Euclidean distance classifier and the Naive Bayes classifier, the WMDCRO has better stability and classification accuracy.

VI. CONCLUSION

In this paper, we first construct a new formula based on the relative offset and the Fisher criterion to weight the attributes. Then present a new weighted minimum distance classifier WMDCRO. Four kinds of datasets are used to test this classifier. Experiment results show that the WMDCRO effectively improves the classification accuracy compared with the traditional MDC. And the classification stability of WMDCRO is better than that of the Naive Bayes classifier. However, the median is selected as the normal value in this paper. For different types of data, whether there is a better normal value selection method is a problem to be further studied. Besides, for the datasets with large attributes, how to combine the dimensionality reduction algorithm with this classifier is also the next research content.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Funds for the Central Universities, Southwest Minzu University (NO.2016NZYQN07).

REFERENCES

[1] V. Khirodkar, R. Saha, M. M. Sardeshmukh and R. Borse, "Employing minimum distance classifier for emotion recognition analysis using EEG signals," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, 2017, pp. 1-9.

[2] G. Bhaumik, T. Mallick, K. S. Chowdhury and G. Sanyal, "Analysis and Detection of Human Faces by Using Minimum Distance Classifier for Surveillance," 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, Kochi, Kerala, 2010, pp. 265-267.

[3] R. Li, S. Zhang and X. Yi, "Scene recognition algorithm based on multi-feature and weighted minimum distance classifier for digital hearing aids," 2016 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, 2016, pp. 34-39.

[4] S. Senda, M. Minoh and I. Katsuo, "A fast algorithm for the minimum distance classifier and its application to Kanji character recognition," Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Quebec, Canada, 1995, pp. 283-286 vol.1.

[5] W. Shi, B. Xue, S. Guo, D. Y. T. Goh and W. Ser, "Obstructive Sleep Apnea Detection Using Difference in Feature and Modified Minimum Distance Classifier," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, 2018, pp. 1-4.

[6] N. M. Al-Obaidi and M. M. Al-Jarrah, "Statistical median-based classifier model for keystroke dynamics on mobile devices," 2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC), Beirut, 2016, pp. 186-191.

[7] H. Lin and A. N. Venetsanopoulos, "A weighted minimum distance classifier for pattern recognition," Proceedings of Canadian Conference on Electrical and Computer Engineering, Vancouver, BC, Canada, 1993, pp. 904-907 vol.2.

[8] S. Bag and G. Sanyal, "An efficient face recognition approach using PCA and minimum distance classifier," 2011 International Conference on Image Information Processing, Shimla, 2011, pp. 1-6.

[9] S. Xu, Q. Luo, H. Li and L. Zhang, "Time Series Classification Based on Attributes Weighted Sample Reducing KNN," 2009 Second International Symposium on Electronic Commerce and Security, Nanchang, 2009, pp. 194-199.

[10] J. Min and Y. Chai, "A PCNN improved with fisher criterion for infrared human image segmentation," 2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, 2015, pp. 1101-1105.

[11] S. Cao, Z. Hou, L. Wang and Q. Zhu, "Kernelized Fuzzy Fisher Criterion based Clustering Algorithm," 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, Hong Kong, 2010, pp. 87-91.

[12] F. Song and H. Li, "Discriminant face images taking both the feature correlation and Fisher criterion into account," 2010 Sixth International Conference on Natural Computation, Yantai, 2010, pp. 3305-3308.

[13] Q. Xu and J. Kong, "A new segmentation algorithm with the fisher criterion function," 2009 ISECS International Colloquium on Computing, Communication, Control, and Management, Sanya, 2009, pp.63-66.