A New Local Mean-based Nonparametric Classification Method

Xiaoqin Zhang[1, a], Feng Liu[1, b]

[1]School of Computer and Information Science, Southwest University, Chongqing 400715, China.

[a]835357908@qq.com, [b]liuf@swu.edu.cn

Keywords: k-nearest neighbor; local mean-based nonparametric classifier; new local mean-based nonparametric classification method

Abstract: As an improved method of k-nearest neighbor classification, the local mean-based nonparametric classifier had the ability to resist the effects of noise and classify unbalanced data. When selecting the nearest k samples and calculating the distance between the test samples and the local mean-based vectors, it always used Euclidean distance. However, for multi-dimensional data, using Euclidean distance which focused on the difference of the value to determine whether two vectors was similar was not so accurate. To solve this problem, a new local mean-based nonparametric classification method was proposed in this paper. It used the cosine distance which focused more on the difference of the dimension to select the k nearest neighbors and compute the distance between the test samples and the local mean-based vectors. The new local mean-based nonparametric classification method was tested on the UCI datasets: Iris and Wine for different values of k in different test data set, the simulation results show that it outperforms the existing local mean-based nonparametric classifier.

Introduction

Among the classification methods, k-nearest neighbor (k-NN) which was initially proposed in 1968 by Cover and Hart determines the type of samples to be classified based on the categories of the k neighbors. It is an easy method to understand without training, and is more perfect on the side of theory [1]. But it has two major problems. First, when the distribution of samples is uneven, only caring the order of the first k nearest samples not considering the sample density may cause misjudgment, affecting the performance of classification. Second, it is the choice of k. If the value of k is too small, the number of neighbors to be obtained is also too small. This will not only reduce the classification accuracy, but also enlarge the disturbing of noise data. While the k value is too large, increased noise will cause lower classification performance.

With regard to these problems, many scholars in domestic and foreign also have studied [2-8]. For the first problem, we can homogenize the samples' distribution density to improve. Local mean-based nonparametric classifier (LMC) which was proposed by Y. Mitani et al. is a very good improved method. It has the ability to classify unbalanced data [9]. For the second problem, so far, cross- validation (CV) has been usually used to select an exact k value that is relatively good. However, these methods still exist two problems. First, LMC always used Euclidean distance. Since the Euclidean distance focus on the difference between values and is very sensitive to noise characteristics [10], it cannot well represent the similarity between multidimensional vectors. Second, the CV's reproducibility is poor, and it needs to obtain a suitable and precise value of k.

In order to improve the above problems, this paper proposed a new local mean-based nonparametric classification method (N-LMC). Firstly, this method used cosine distance to measure the similarity between sample data. Secondly, in order to better evaluate the N-LMC method, our experiments will test the classification accuracy of different k values to shield the selection of exact k value, then randomly select different test data set to verify the average classification accuracy of different k for several times. Finally, our method is tested on the UCI datasets [11, 12].

Related Theory

k-NN. The k-NN algorithm is a relatively mature approach in theory, the idea is very simple and intuitive, and easy to be quickly achieved. The basic idea of k-NN algorithm is: calculating the distance (as similarity) between sample x to be classified and each training samples depending on the distance function, then selecting k samples whose distances are the minimum as the nearest k neighbors of x, and finally x is assigned to the category that most of the nearest k neighbors belong to.

According to k-NN algorithm, we must first look for k neighbors, that is to say, we must find k-nearest training set samples. The similarity is generally computed by Euclidean distance. The Euclidean distance between two n-dimensional sample points a(x_{11},x_{12},…,x_{1n}) and b(x_{21},x_{22},…,x_{2n}) can be expressed as

$$dist(a,b)=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^2} \tag{1}$$

LMC. The local mean-based nonparametric classifier is a lazy, local, non-parametric classifier which was proposed by Y.Mitani and others in 2006. This method can be said to be another improved method of k-NN. The main idea is to calculate the local geometric center of each class as the nearest neighbor of the test sample. This geometric center is known as the "local mean-based vector". Specifically, in LMC, first is to select k samples by Euclidean distance in each class of training samples which are the nearest to test sample, and then use the k samples to calculate the local mean-based vector of each class. Since each class has a local mean-based vector, these vectors can be seen as a represent point of each category. Finally, this method classify the test sample into the category that the nearest local mean-based vector belongs to.

The test results of Y. Mitani et al. showed that LMC had better classification performance comparing with the classic classification algorithms k-NN, Parzen window and artificial neural network (ANN). Furthermore, this method also has the ability to resist the effects of noise and classify unbalanced data.

N-LMC

The new local mean-based nonparametric classification method proposed in this paper is an improved method on the basis of LMC. Firstly, when selecting the k nearest samples of test sample x, it uses cosine distance to measure the similarity between them. Secondly, the distance calculation between test sample x and local mean-based vector also uses cosine distance. The cosine distance between two n-dimensional sample points a(x_{11}, x_{12},…,x_{1n}) and b(x_{21},x_{22},…,x_{2n}) can be expressed as:

$$sim(a,b)=\cos\theta=\frac{\sum_{k=1}^{n}x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n}x_{1k}^{2}}\sqrt{\sum_{k=1}^{n}x_{2k}^{2}}} \tag{2}$$

The smaller $\cos\theta$ ($sim(a,b)$) is, the more dissimilar the two samples are. i.e., the larger $sim(a,b)$ is, the more similar the two samples are. Because the range of $\cos\theta$ is $[0,1]$, so the similarity of the two samples is $[0,1]$. Here normalization process is made to the similarity between the samples.

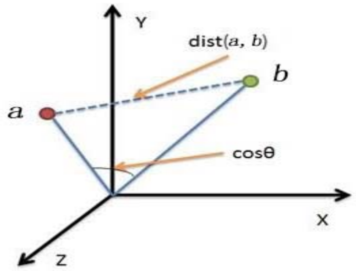

Fig. 1 Comparison of Euclidean distance and cosine distance

As can be seen from Fig. 1, Euclidean distance is to measure the absolute space distance between each point. It is of direct relevance with the position coordinates of each point. However, cosine distance is to measure the angle between the space vectors, which reflects more in the difference of direction, rather than position. Obviously, the cosine distance is more suitable to describe the similarity of multi-dimensional data.

After following the improvements above, the specific algorithm of the new local mean-based nonparametric classification method is described as follows:

Algorithm N-LMC (x, X, k)

/ * x is the test sample, X is the training sample set, k is the number of the nearest neighbor samples selected in each class * /

Step 1: Using cosine distance to select k nearest samples of the test sample x in each class ω_j, represented by $X_k(\mathrm{x}, \omega_j)$;

Step 2: Calculating the local mean-based vector for each class ω_j according to the selected k-nearest neighbor samples in the following method:

$$\bar{x}_j = \frac{1}{k}\sum_{i=1}^{k} x_i, x_i \in X_k(\mathrm{x}, \omega_\mathrm{j}) \tag{3}$$

Step 3: For each class ω_j, calculating the cosine distance between test sample x and the local mean-based vector $\bar{x}_j$:

$$sim(x, \bar{x}_j) = \cos\theta = \frac{\vec{x} \cdot \vec{\bar{x}}_\mathrm{j}}{|x||\bar{x}_\mathrm{j}|} \tag{4}$$

Step 4: Classifying the test sample x into the class ω_r which has the maximum $\cos\theta$ according to the value of $\cos\theta$ between test sample x and the local mean-based vector $\bar{x}_j$ in each class ω_j.

The test results of this method show that, whether to test corresponding classification accuracy of different k values or to test the average classification accuracy of different k values while randomly selecting test data, N-LMC has better classification performance comparing with k-NN and LMC.

Simulation Experiments

The experiments are conducted on two public data sets: Iris Data Set and Wine Data Set. The two data sets are UCI data sets. The basic information is in Table 1.

Table 1 Data sets for experiments

Data set	Total samples	Features	Categories	Test samples
Iris	150(50,50,50)	4	3	30
Wine	178(59,71,48)	13	3	45

The general idea of the experiment design is: in order to better verify the classification performance of the N-LMC proposed in this paper, we choose a range of k other than select an accurate k to contrast the classification accuracy of each method. Since the experiments randomly select test data and training data, the test results of a single experiment may be data dependent, we do experiments several times, each time selecting different test set and training set, finally compare the average classification accuracy of each method among these experiments.

Experiment 1. Related description of experiment 1 is below:

(1) Randomly selecting 120 samples (each category 40 samples) in Iris data set. These 120 samples are seen as the training set, and the remaining 30 samples (each category 10 samples) are test samples whose categories are unknown;

(2) Taking $\mathrm{k} \in [3,15]$, this experiment was repeated 10 times, comparing the average classification accuracy of different k for each method in 10 times.

First, selecting corresponding classification accuracy of different k for each method to compare, the result is shown in Fig. 2. Then, looping test 10 times, each time selecting different test samples so that we can better compare the classification results. Comparing the average classification accuracy of different k each time for the three methods. The experiment result is shown in Fig. 3.

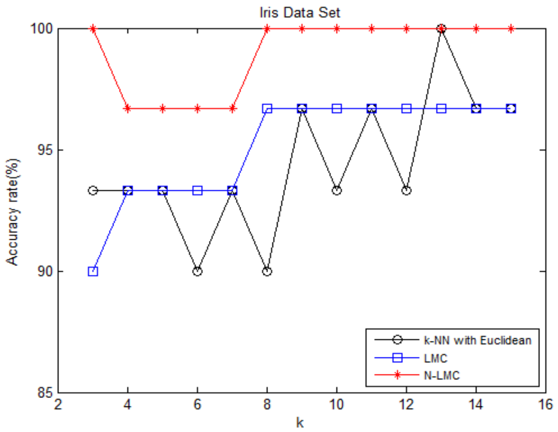

Fig. 2 Classification accuracy of different k

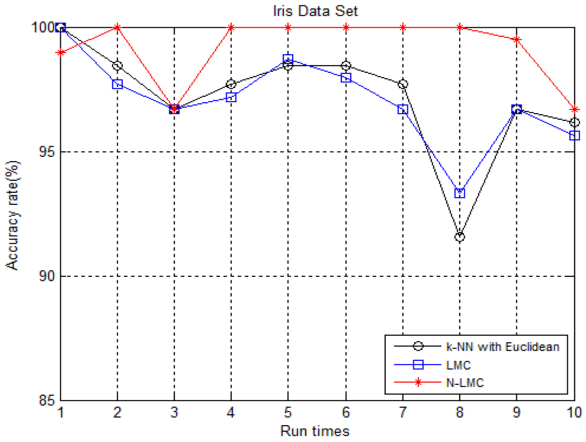

Fig. 3 Average classification accuracy of different k

As it can be seen from Fig. 2 and Fig. 3, on the Iris data set, when k takes from 3 to 15, the classification accuracy curve jumps largely in the experiments of k-NN. However, classification accuracy curve of N-LMC is relatively smooth and the accuracy rates are higher than k-NN and LMC. Furthermore, the average classification accuracy of different k values of N-LMC is generally higher. This illustrates that the classification performance of N-LMC proposed in this paper is better and more stable than the existing k-NN and LMC.

Experiment 2. Related description of experiment 2 is below:

(1) Randomly selecting 45 samples (each category 15 samples) in Wine data set. These 45 samples are test samples whose categories are unknown. The remaining 133 samples are seen as the training set (the first category 44 samples, the second category 56 samples, the third category 33 samples) ;

(2) Taking $k \in [3,20]$, repeating the experiment 20 times, comparing the average classification accuracy of different k for each method in 20 times.

First, selecting corresponding classification accuracy of different k for each method to compare, the result of this experiment is shown in Fig. 4. Then, looping test 20 times, each time selecting different test samples. Comparing the average classification accuracy of different k each time for the three methods. The experiment result is shown in Fig. 5.

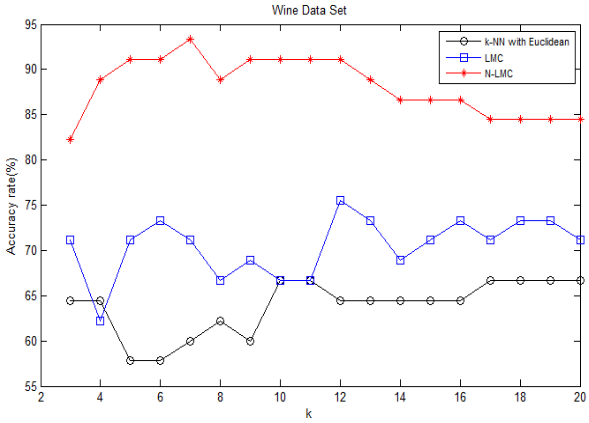

Fig. 4 Classification accuracy of different k

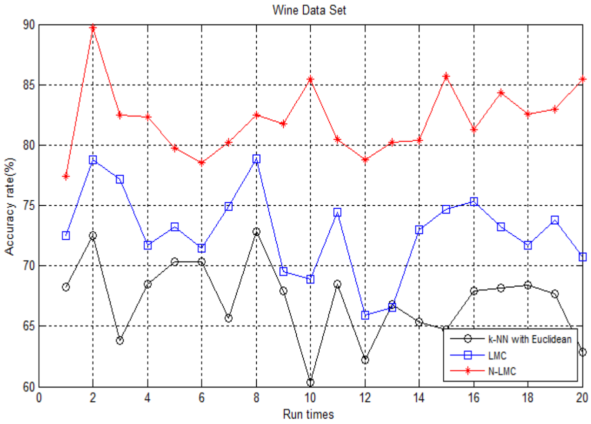

Fig. 5 Average classification accuracy of different k

As we can see in Fig. 4 and Fig. 5, on the Wine data set, when k takes from 3 to 20, the classification accuracy of N-LMC is significantly higher than the existing k -NN and LMC. Even if comparing the average value of classification accuracy for different k, the overall effect of N-LMC proposed in this paper is better than k-NN and LMC.

Conclusions

This paper proposes a new local mean-based nonparametric classification method. First, it uses the cosine distance which pays more attention to the difference between the dimensions to measure the similarity between the sample data. Then, it calculates the average classification accuracy of different k values by randomly selecting different test data each time. This not only shields the choice of precise k value but also can test whether the proposed method is suitable for different values of k. Finally, experiments are carried out on two public datasets. The results of experiment 1 and experiment 2 have indicated that the classification performance of the proposed N-LMC in this paper is better than k-NN and LMC. By comparing the experiment 1 and experiment 2, it can be found that on the Wine data set in which the number of sample features is a few more and the number of samples of each class are different, the advantage of N-LMC with respect to k-NN and LMC is more obvious than it is on the Iris data set. It proves that the proposed N-LMC is better suitable for the classification of multidimensional data.

Acknowledgements

Thanks for our corresponding author professor Feng Liu and the anonymous reviewers for their helpful suggestions. Furthermore, we thank Southwest University for the experiment platform they provided.

References

[1] T.M. Cover, P.E. Hart. Nearest neighbor pattern classification [J]. IEEE Trans, IT-13(1967)21-27.

[2] Y. Mitani, Y. Hamamoto. A local mean-based nonparametric classifier [J]. Patter Recognition Letter, 27(10), 2006, pp.1151-1159.

[3] Y.B. Sang. Study on classification algorithm based on k nearest neighbor [D]. Chongqing: Chongqing University, 2009.

[4] E.B. Du. Design and implementation of text classification algorithm based on improved k-NN [D]. Shanghai: Shanghai Jiao Tong University, 2010.

[5] G.H. Feng, J.X. Wu. Study progress of improved k-NN classification algorithm [J]. Information Technology, 56 (21), 2012, pp.97-100.

[6] N. Du. Research on fuzzy k-NN based on Dempster-Shafer theory and the application in fault diagnosis [D]. Shijiazhuang: Hebei Normal University, 2011.

[7] Z.Y. Zhang, Y.L. Huang, H.H. Wang. A new k-NN classification approach [J] .Computer Science, 35 (3), 2008, pp.170-172.

[8] X.Y. Wang, Z.O. Wang. An improved k nearest neighbor algorithm [J]. Journal of Electronics and Information Technology, 27 (3), 2005, pp.487-490.

[9] H.S. Li. Research of classification methods based on evidence theory [D]. Guangzhou: South China University of Technology, 2013.

[10] X.M. Bi. Review of k-NN algorithm [J]. Science and Technology Innovation Herald, 2009, 14:31.

[11] F. Lu, N. Du. A fuzzy-evidential k nearest neighbor classification algorithm [J]. Acta Electronica Sinica, 40 (12), 2012, pp.2390-2395.

[12] M. Lichman. UCI machine learning repository [EB/OL]. (2013) [2014-12-10] http://archive.ics.uci.edu/ml/.