An Improved kNN Based on Class Contribution and Feature Weighting

HUANG Jie[1,2], WEI Yongqing[3,*],YI Jing[2,4] and LIU Mengdi[1,2]
[1]School of Information Science & Engineering, Shandong Normal University, Jinan ,Shandong,250358,China
[2]Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan 250358, China;
[3]Basic Education Dept. ,Shandong Police College, Jinan , Shandong,250014,China
[4]School of Computer Science & Technology, Shandong Jianzhu University, Jinan ,Shandong,250014,China
*corresponding author e-mail:13864155372@163.com

***Abstract*—Aiming at the problem that the kNN algorithm is susceptible to the choice of k-nearest neighbors and the method of class judgment, this paper propose a kNN algorithm based on class contribution and feature weighting called DCT-kNN . Firstly, using traditional kNN to calculate accuarcy of original dataset and of the data lack of each dimension feature successively.Then by comparing two accuarcies to weight the feature and to calculate the weighted distance,by which the k-nearest neighbors are obtained.Finally,by using class contribution which combines the number of k-nearest neighbors and their mean distance ,the final labels of the samples are obtained. The comparison experiment of UCI datasets showed a certain degree of improvement in classification accuracy of the proposed method.**

Keywords- kNN,feature weighting,class contribution

I. INTRODUCTION

With the advent of the big data age, there is an urgent need for more efficient categorization in order to obtain the information that is needed, valuable and concise from the massive data.In the field of data mining, there are many urgent learning methods such as Bayesian classification[1], decision tree classification[2] and neural network classification[3] ,and lazy learning method based on analogy like kNN,which is widely used in text classification and pattern recognition because of its property like simplicity and operation.kNN algorithm is a classic analogy-based classification algorithm, it does not establish a classification model as SVM, it stores all the training samples until the test samples are classified .The basic process is as following: when a test sample is given,kNN algorithm searches for the n-dimensional pattern space of the training data, and finds the k training samples closest to the sample to be sorted by a certain distance measure,and finally the category is judged to be the class that has the most nearest neighbors of k-nearest neighbors.However, there are many problems: both the selection of k and distance formula and the uneven distribution of the sample set will have impacts on the classification accuracy.

In the paper,we present the DCT-kNN algorithm, which improves the kNN by changing algorithm itself. Firstly, the feature is weighted by the accuracy rate associated with itself to obtain the k-nearest neighbors,and then the samples are classified by considering both the number of samples and the class contribution of the sample distance.

II. KNN

A.The basic process of kNN

kNN concludes two steps mainly.First,find the k-nearest neighbors:calculating the the euclidean distance between two points or tuples $X1=(x_{11},x_{12},...,x_{1n})$, $X2=(x_{21},x_{22},...,x_{2n})$,the formula is as following:

$$dist(X1,X2)=\sqrt{\sum_{i=1}^{n}(x_{i1}-x_{i2})^2} \tag{1}$$

In the actual operation, in order to prevent the wide gap of the weigh of characteristics that have different initial value ranges , the value of feature is normalized as follows:

$$v'=\frac{v-\min_A}{\max_A-\min_A} \tag{2}$$

$\max_A$、$\min_A$ are the maximum and minimum values of the feature A respectively,and v'、v represent the normalized and non-normalized eigenvalues, respectively.

Secondly,getting the classfication according to the category of k proximity points:

$$C_X=\arg\max_{j\in l}\sum_{y\in X_k}I(C_y=j) \tag{3}$$

X_k is the k-nearest neighbors including y, C is the label;the return value is 1 when the label of the y is equal to j,that is,the value of $I(\cdot)$ is "true" , otherwise, $I(\cdot)=0$

B.Methods of improving accuracy of kNN

In view of the two steps and problems of kNN, scholars have put forward a lot of solutions to improve the accuracy of kNN[9].Authors proposed methods of cutting the samples based on density[4,5]. The above method of cutting samples can solve the problem of sample imbalance and computational complexity to a certain extent, the screening of samples can lead to classification error easily.In addition, how to determine the sample to be cut is also a question worth to continue to explore. In[6], a kNN algorithm based on the improved average distance of class is given and authors proposed a weighted kNN of the sequence ,it took k-nearest samples and the samples that have been classified into consideration to get the classification result.But it is difficult to ensure that the classification of the classified samples is completely correct ,which maybe make the error rate increasing[7].In [8], the inverse proportion weight kNN was given to release the unbalance of the sample data in a way by using the relationship between the number of one class samples and the total number of all samples.The kNN

2157-1481/18/$31.00 ©2018 IEEE
DOI 10.1109/ICMTMA.2018.00083

algorithm based on correlation distance of attribute value proposed in [9] has improved kNN from the above two aspects.But the first step of the distance measurement is not rigorous, so that the process will have an impact on the back process.Current improvements of the kNN algorithm include some ways aiming at its own defections and ways combining with other ways such as rough sets, SVM, and so on.

III. WEIGHTED KNN BASED ON IPROVED CLASS CONTRIBUTION

A.Definitions

Definition 1 (Ability to distinguish class of feature)

The category distinguishing ability of the i-dimensional feature is defined as follows:

$$Disc_i = 1-(pre_i - pre_t) \tag{4}$$

In this paper, the pre_t is the sum average accuracy of the traditional kNN algorithm in the original dataset when the value of k is 3,5 and 7 on the condition of 5-folder cross validation, and the pre_i is the similar sum of the traditional kNN algorithm on the data lacking the i-th feature.

From the formula (4) we can learn that the accuracy of classification after removing the i-th dimension is decreased when $pre_i - pre_t < 0$ that is $Disc_i > 1$,that is to say, to a certain extent, this feature is conducive to raising the correctness of the final classification results, so the corresponding weight of the feature is increased on the basis of original value,which is reasonable.On the contrary, when $pre_i - pre_t > 0$ that is $Disc_i < 1$,which represents that the lack of the i-th dimension feature leads to the increase of accuarcy,that is to say ,the feature is less important.It is obvious that the degree of importance depends on the the gap of the accuracy, the greater the gap is the more unimportant the feature is.Therefore, the definition of (4) can measure the importance of each feature on the accuracy of classification effectively, that is, the ability of determining classification of features.Finally, the corresponding weight w_i is obtained by normalizing the i-dimensional feature, as shown in the following equation, where n is the number of dimension of features:

$$Disc_i = \frac{Disc_i}{\sum_{i=1}^{n} Disc_i} \tag{5}$$

Definition 2 (Weighted Euclidean distance)

From the definition 1 we derive the weight of each feature, on the basis of which we improve the Euclidean distance as shown in the following formula:

$$dist(X1, X2) = \sqrt{\sum_{i=1}^{n} w_i (x_{1i} - x_{2i})^2} \tag{6}$$

From the weighted Euclidean distance we know that the distinction the distance between the two samples will be more accurate if we treating two features that have different separating capacity differently.What we do above will lay a basis for the next one step that determining the label of the unseen instance.

Definition 3 (Class Contribution)

The concept of class contribution(CT)takes the sample number and the distance of the sample in the k neighborhoods of the tested sample that are important into account , defined as follows:

$$CT_j = \frac{K}{N_j} + \frac{1}{N_j} \sum d(X, Y_j) \tag{7}$$

In the above formula: k is the value of k in the experiment, N_j is the number of samples of the j-th class in the k-nearest neighbors, and $\sum (X, Y_j)$ is the sum of the distance between all the samples in the k-nearest neighbors.The first item on the right side of the equation is the reciprocal of the proportion of the j-th class in the k-nearest neighbor.When the k value is determined, the smaller the item is, the more j-th class points in the nearest points of x. The second item on the right is the average distance of the x-class of the j-th sample in the k-nearest neighbors. The smaller the item is, the closer the j-th sample is to x, so we choose the index of the smallest value of CT_j as the final category of the sample to be sorted, and the discriminant is shown in equation (8).In addition, if the k-nearest neighbors does not contain any sample of some category, that is $N_j = 0$, in order to avoid the infinite operation of the situation, we specify $CT_j = 0$,indexof()represents index.

$$C_X = indexof(\min(CT_j)), i = 1:l \tag{8}$$

Table 1 DCT-kNN Algorithm

Improved Algorithm DCT-kNN
Input: n-dimensional dataset D,l categories and k Output: Category of tested sample x 1.Classifying D by traditional kNN and kNN that delete one feature to get the value of pre_t 、 pre_i ,and then calculate the weight of the feature by (4)(5). 2. Obtain the Euclidean distance between the training sample and x by the formula (6) and the k-nearest neighborsare. 3.Get the number in the k-nearest neighbors and the average distance of every category. 4. Put the value we get into (7), the minimum value can be obtained by the formula (8) which is the label of the test sample. If more than two CT_j have the same value,we regard the test data belongs to the nearest one.

IV. EXPERIMENTAL RESULTS

In order to verify the validity of the algorithm, the three datasets on the UCI public dataset are experimented under matlab2016a.Experimental environment of the computer configuration: CPU for the Core i3, memory 2G, the operating system is Windows7, the experimental data are obtained in the experimental environment.

Table 2 Experimental dataset

Dataset	feature number	instance number	class number	proportion
Iris	4	150	3	50:50:50
Liver	6	340	2	195:145
Blood	4	740	2	562:178

The detailed informations of datasets taken from the UCI are given in Table 2 above, where the original numbers of instances of Liver and Blood are 345 and 748, respectively.5 and 8 instances were randomly cut to make sure the 5-fold and 10-fold cross validation is performed smoothly.

A. Experimental Evaluation

5-fold and 10-fold cross validation were conducted and we make contrast experiments between DCT-kNN and kNN algorithm, [6] and [10].The final accurates were shown in the following figure1 to figure3.

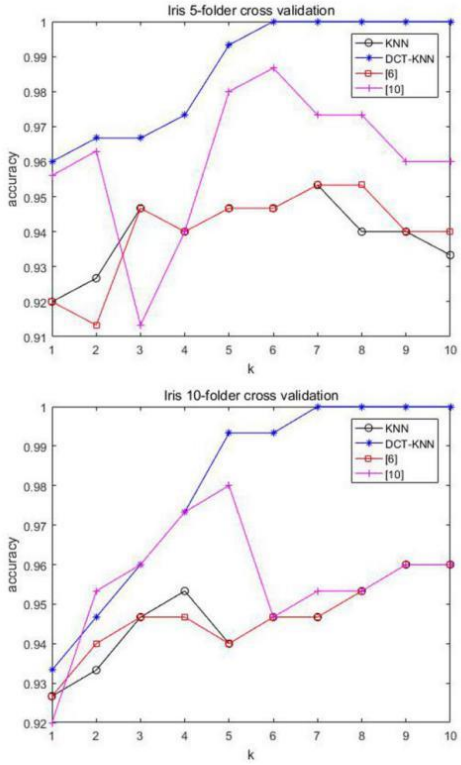

Figure 1. 5-fold and 10-fold cross validation results of Iris

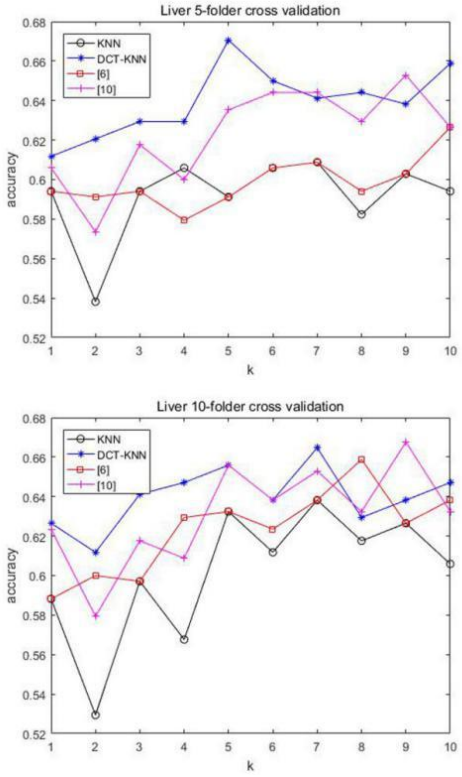

Figure 2. 5-fold and 10-fold cross validation results of Liver

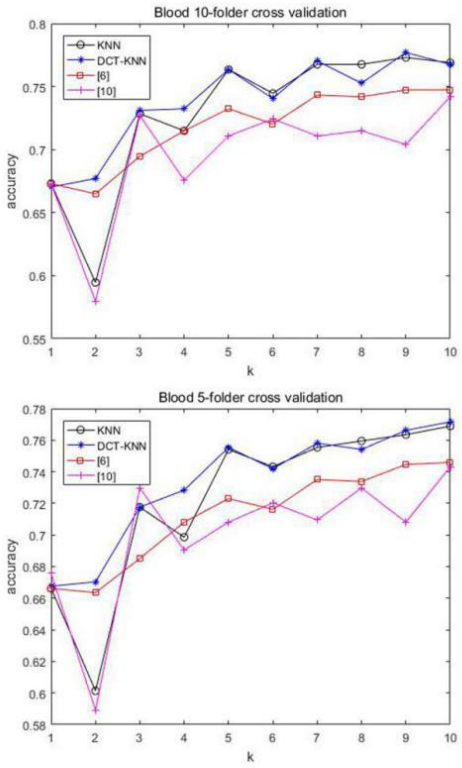

Figure 3. 5-fold and 10-fold cross validation results of Blood

From the above three figures, we can see that in figure 1, the clasification accuarcy of [6],[10] and DCT-kNN have increased compared with the original kNN in the 10-fold experiment .Our method have an obvious improve on the overall accuarcy and the effect of [10] is slightly worse .In the 10-fold experiment , the accuracy of our method is significantly higher than the other two methods.In general, the accuracy rate increases with the increase of the value of k.Figure 2 shows that in the 5-foler experiment of Liver , the accuracy of the method [6] is relatively improved compared with the original method, but it is still less than the method of [10],while our method is significantly improved compared with [10].In the 10-folder experiment,the DCT-kNN,[6] and [10] improved the accuracy,of which DCT-kNN is the best.In these cases, the accuracy rate also increases with the increase of the k.Finally, in the figure 3 ,the method of literature [6] and [10] obviously worse than the other two methods. Compared with the traditional kNN,our method is overall improved in the classification accuracy,however,the 5-folder cross experiment is more obvious than 10-folder cross validation.It can be seen that the weighting kNN based on class contribution can improve the accuarcy of classfication effectively.

V. CONCLUSIONS

DCT-kNN makes full use of the relationship between the feature and the category.From the shortcomings of the traditional kNN algorithm, we consider the two important factors,such as the number of samples in the k neighborhood and the distance of the k-nearest neighbors to the sample x,and then we improve kNN direct at the two factors.Thus,we improve the accuracy of the traditional kNN method in distance measurement and category judgment, which helps improving the accuracy of classification. How to ensure the efficiency based on the accuarcy when dealing with datasets with multi-dimensional characteristics is another issue for future.

ACKNOWLEDGEMENTS

This work was financially supported by the National Natural Science Foundation Project (No.61373148,No.61502151),Ministry of Education Humanities and Social Science Foundation Project(No.14YJC860042),the Natural Science Foundation of Shandong Province (No.ZR2014FL010),Shandong Province Outstanding Young Scientists Award Fund funded projects(No.BS2013DX033),Shandong Province Higher Education Science and Technology Program (No.J15LN02, No.J15LN22),Shandong Province Social Science Planning Project(No.16CXWJ01,No.16CFXJ05).

REFERENCES

[1] LIN Shi-min, TIAN Feng-zhou, LU Yu-chang.Study on Bayesian Classifier for Data Mining [J] .Computer Science, 2000, 27 (10): 73-76.

[2] CHEN Jing, XIAO Ding.Application of decision tree algorithm in data mining [J]. Software Journal, 2008 (3): 98-99.

[3] Han Hong, Yang Jingyu. Combination of neural network classifier [J] .Journal of Computer Research and Development, 2000, 37 (12): 1488-1492.

[4] Li Ronglu, Hu Yunfa. Based on the density of kNN text classifier training sample cutting method [J]. Computer Research and Development, 2004, 41 (4): 539-545.

[5] Luo Xianfeng, Zhu Shenglin, Chen Zejian, et al.Improved kNN text categorization algorithm based on K-Medoids clustering [J]. Computer Engineering and Design, 2014, 35 (11): 3864-3867.

[6] Yan Xiaoming. Weighted kNN classification algorithm based on average distance of class [J]. Journal of Computer Applications, 2014, 23 (2): 128-132.

[7] ZHU Ming-dry, LUO Da-yong, YI Li-qun. A Weighted kNN Classification Method for Sequences[J] .Journal of Electronics, 2009, 37 (11): 2584-2588.

[8] Li Weiping, Yang Jie, Wang Gang. Proportional inverse weight kNN algorithm and its stream processing application [J]. Computer Engineering and Design, 2015 (12): 3355-3358.

[9] Feng Guohe, Wu Jingxue. KNN classification algorithm to improve research progress [J]. Library and Information Work, 2012, 56 (21): 97-100.

[10] XIAO Hui-hui, DUAN Yan-ming.Improvement of kNN Algorithm Based on Attribute Value Correlation Distance[J] .Computer Science, 2013, 40 (11a): 157-159.