ging, 26(3), p.p.405-421.

10. D. Adalsteinsson, J. A. Sethian (1999) The Fast Construction of Extension Velocities in Level Set Methods. *Journal of Computational Physics*, 148, p.p.2-22.

11. Guopu Zhu, Qingshuang Zeng, Changhong Wang (2007) Boundary-based Image Segmentation Using Binary Level Set Method. *Opt. Eng.*, 46(5), p.p.33-39.

An Improved k-Nearest Neighbor Classification Algorithm Using Shared Nearest Neighbor Similarity

Wei Zheng

College of Science, Hebei North University, Zhangjiakou 075000, Hebei, China

HaiDong Wang

Library of Hebei Institute of Architecture and Civil Engineering, Zhangjiakou 075000, Hebei, China

Lin Ma

College of Science, Hebei North University, Zhangjiakou 075000, Hebei, China

RuoYi Wang

School of Information Science and Engineering, Hebei North University, Zhangjiakou 075000, Hebei, China

Abstract
k-Nearest Neighbor (KNN) is one of the most popular algorithms for pattern recognition. Many researchers have found that the KNN classifier may decrease the precision of classification because of the uneven density of t raining samples .In view of the defect, an improved k-nearest neighbor algorithm is presented using shared nearest neighbor similarity which can compute similarity between test samples with nearest neighbor samples. The experiment shows that this method can enhance classification precision compare to the traditional KNN.

Key words: KNN ALGORITHMS, SAMPLE CLASSIFICATION, NEAREST NEIGHBOR

1. Introduction

Pattern recognition is about assigning labels to objects which are described by a set of measurements called also attributes or features. There are two major types of pattern recognition problems: unsupervised and supervised. In the supervised category which is also called supervised learning or classification. Classification of objects is an important area of research and of practical application in variety of fields, including artificial intelligence, statistics and vision analysis. Considered as a pattern recognition problem, there have been numerous techniques investigated for classification[1].

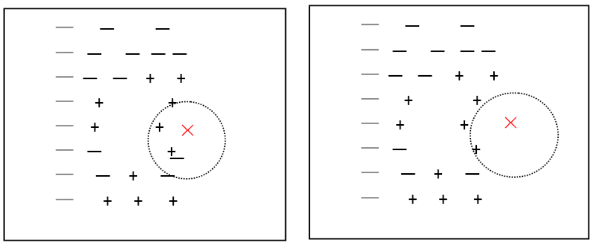
K-Nearest Neighbor (KNN) classification is one of the most fundamental and simple classification methods. When there is little or no prior knowledge about the distribution of the data, the KNN method should be one of the first choices for classification, but it has some disadvantages such as :a)Computation cost is quite high because it needs to compute the distance of each query instance to all training samples; b) Low accuracy rate in multidimensional datasets; c) Need to determine the value of parameter K, which is the number of nearest neighbors; d) Distance based learning is not clear which type of distance to use[2].Researchers have attempted to propose new approaches to improving the performance of KNN method. e.g, Discriminant Adaptive NN [3] (DANN), Adaptive Metric NN [4] (ADAMENN), Weight Adjusted KNN [5] (WAKNN), Large Margin NN [6] (LMNN) and etc. Despite the success and rationale of these methods, most have several constraints in practice.

In this paper, we propose a novel, effective yet easy-to-implement extensions of KNN method named IKNN. Innovation point lies in using the Shared nearest neighbor method, through calculating the Shared nearest neighbor value between test sample with the nearest neighbor samples .Nearest neighbor samples weight value is redistributed based on nearest neighbor value. Through the experiment discovered IKNN method have higher classification accuracy than the KNN method with UCI data set.

The rest of this paper is organized as follows. Section 2 describes the KNN method commonly used. Section 3 studies the nearest neighbor similarity and gives an improved KNN method named IKMM. Section 4 discusses the classifier using in experiment to compare IKNN with other KNN methods, and presents the experiment's results and analysis. In the last, we give the conclusion and future work.

2. KNN algorithm description

K-Nearest Neighbor classifier is an instance based learning method. It computes the similarity between the test instance with the training instances and considering the k top-ranking nearest instances, and finds out the category that is most similar. There are two methods for finding the most similar instance: majority voting and similarity score summing.



(a) 1 - the nearest neighbor (b) 2 - the nearest neighbor

Figure 1. Two instances of nearest neighbor

In majority voting, a category gets one vote for each instance of that category in the set of k Top-ranking nearest neighbors. Then the most similar category is the one who gets the highest amount of votes. In similarity score summing, each category gets a score equal to the sum of the instances of that category in the k top-ranking neighbors. The most similar category is the one with the highest similarity score sum. The similarity value between two instances is the distance between them based on a distance metric. Generally Euclidean distance metric is used. K-Nearest Neighbor classifier algorithm is as follows: For each test sample z=(x', y'), and algorithm will calculate the distance between it and training sample (x, y)∈D, which D is a training sample set, to determine its nearest neighbor lists.

Algorithm 1. K-Nearest Neighbor classifier algorithm

```
algorithm
    1: k is the number of nearest neighbor and D is the training sample set.
    2: for each test sample z=(x' ,y')  do
    3: Calculate the distance d(x',y') between z and each train sample (x, y)
    4: Select a training sample subset Dz whose has k samples, is the nearest to the z
    5: $y' = \arg\max_{v} \sum_{(x_i,y_i)\in Dz} I(v=y_i)$
    6:end for
```

Once you get the nearest neighbor list, most of test sample will be based on nearest neighbor classification[7]:

Majority voting:

$$y' = \arg\max_{v} \sum_{(x_i,y_i)\in Dz} I(v=y_i) .$$

V is the number of categories, and the category of the y_i is a nearest neighbor number. I (.) is the indicator function, if the parameter is true, it returns 1, otherwise it returns 0.

In the majority voting method, because each neighbor's influence on the classification is equal, algorithm is sensitive to the choice of K, as shown in figure 1, Symbols× represents the test sample and Symbols - ,+ represent the test samples .A way to reduce the influence of K is that algorithm distribute weighted to its function according to each of the nearest neighbor distance . Results from the training sample for classification of effects of z than those near the z training sample. Using distance weighted voting scheme, class label can be determined by the following formula. Distance weighted voting that W_i is weight. The algorithm named WKNN.

Distance weighted voting:

$$y' = \arg\max_{v} \sum_{(x_i,y_i)\in Dz} w_i \times I(v=y_i)$$

3. Improved k-Nearest Neighbor Classification Algorithm Using Shared Nearest Neighbor Similarity

3.1. the problems existing in the KNN algorithm

When KNN algorithm is used into data mining, there are several problems .First, because of based on local information to forecast, it very is sensitive to noise data when the K value is very small. Second, In the practical application of classification algorithms. it maybe have multi-peak distribution problems between each category because of the uneven distribution of data. KNN classification algorithm depends on the distribution of training samples, the high density area samples will affect the judgment of category. Figure 2 and figure 3 in graphical form shows the above problems.

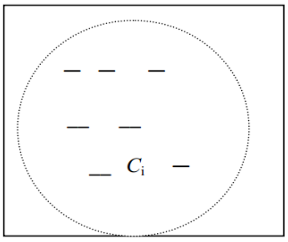


Figure 2. Noise Data exist in category C_i

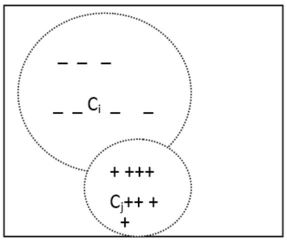


Figure 3. High density data of category C_j

Symbol× is test sample and symbol * is Noise Data exist in category the C_i .Symbol + present High density data of category C_j.Figure 2 shows that the result of KNN classification is easily affected by noise data exist in category C_i . Figure 3shows that the result of KNN classification is easily affected by high density data of category C_j . In order to overcome the KNN algorithm, we proposed an improved algorithm based on shared nearest neighbor similarity.

3.2. Improved k-Nearest Neighbor Classification algorithm

According to paper[8],Levent Ertoz proposed shared nearest neighbor method named SNN to solves the problem of high-dimensional data . The SNN core idea is that the similarity between two objects is defined based on Shared nearest neighbor.It is based on the following principle , If two points are similar to some of the same point, they are also similar, even directly similarity measure can't point out. The computation of SNN similarity between two point can be described as follows and Graphic is shown in figure 4.

Step 1: For each point, find out N nearest points for it and create its nearest neighbor list.

Step 2:If both two points appear in each others nearest neighbor list, there is a link between them. Calculate link strength of every pair of linked points and build a point link graph.

Step 3: The link strength of every pair of linked points is the number of shared neighbor.

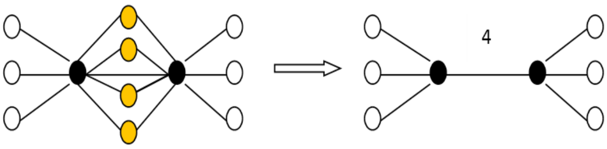


Figure 4. The calculation of SNN similarity between two points

Two black spots there are eight nearest neighbor interaction. These four of the nearest neighbor (orange) is Shared. So the SNN similarity between the two points is 4. Because SNN similarity is good on determining similarity between point and point, according to the SNN similarity ,we proposed an improved KNN classification algorithm.The improved algorithm named IKNN as follows:

Algorithm 2. K-Nearest Neighbor classifier algorithm Using Shared Nearest Neighbor Similarity

```
algorithm                    IKNN algorithm
    1: k is the number of nearest neighbor and D is the training sample set.
    2:for each test sample z=(x’,y’) do
    3:Calculate the distance d(x’,y’) between z and each train sample (x, y)
    4: Select a training sample subset Dz whose has k samples, is the nearest to the z
    5:for each train sample $K_i \in Dz$ do
    6: Calculate the nearest neighbor similarity value Wsnn between Ki with Z
    $y' = \arg\max_{v} \sum_{(x_i, y_i)\in Dz} Wsnn \times I(v = y_i)$
    7:end for
```

4.Experiment and Analysis

4.1. Experimental purposes

We use KNN, WKNN and IKNN algorithm to carry out classification experiment in order to verify the classifying effect. This section compares the IKNN method with original KNN algorithm and t discusses the experimental results.

4.2. Data Collections

The classification algorithm which is KNN, WKNN and IKNN is evaluated on three standard data sets, namely Wine, Iris and Breast Cancer Wisconsin. None of the databases had missing values, as well as they use continuous attributes. These data sets which are obtained from UCI repository [9] are described as Table 3. In these three data sets, the instances are divided into training and test sets by randomly choosing two third and one third of instances per each of them, respectively.Table 3 shows the specific quantity of samples in each category we choose.

Table 1. The data sets used in the experiments

	Data vectors	Attributes	Classes
Iris(IR)	150	5	3
Breast Cancer Wisconsin (BCW)	474	9	2
Wine(WI)	144	13	3

4.3. Performance measure

To evaluate the performance of a classifier, we use accuracy measure put forward by Yuba yavuz (1998) [10]. This measure as follows:

$$\text{accuracy} = \frac{\textit{number of correct predictions}}{\textit{nnumber of total predictions}}$$

4.4. Results and Analysis

When Kvaluefrom the 5 to 25,We carry out classify experiment . The experimental results have shown in Figure 5,Figure 6, Figure 7 .

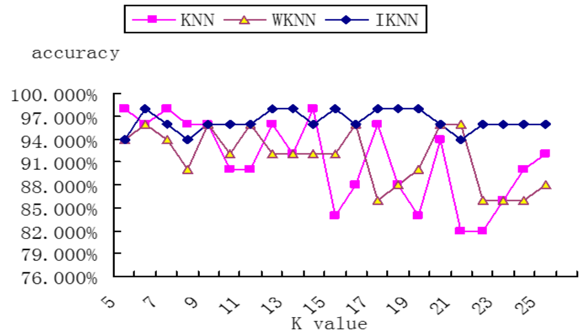


Figure 5. Iris data set

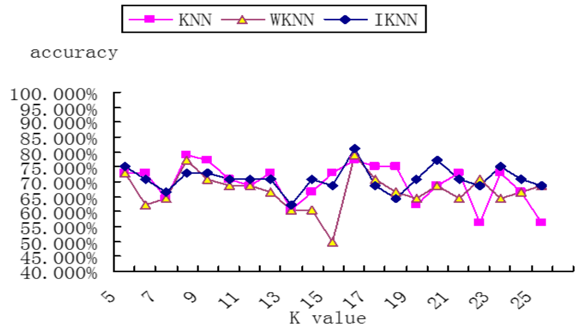


Figure 6. Wine data set

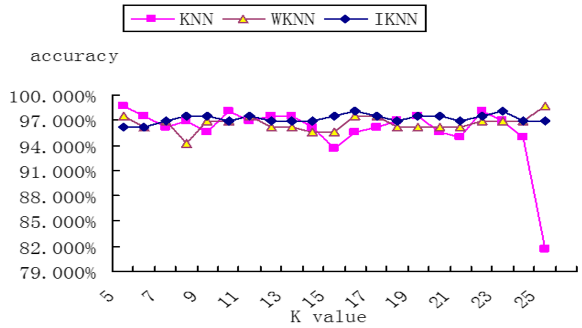


Figure 7. BCW data set

From figure 5, the curve of IKNN classification is clearer ,which classification effect is significantly better than the other two.Through figure 5,we can see that the accuracy for IKNN is greater than KNN and WKNN whose classification curve is smooth. The overall classification affectionfor WKNN was slightly better than KNN, and the KNN classification error rate is the largest.When classification experiments is carried out,using WINE data set, IKNN classification effect is not obvious, as figure 5 shown,with the change of K value, classification curve fluctuations .When k is equal to 16, IKNN classification algorithm get the best classification results that only nine samples was wrongly distinguish in classification experiment. Figure 7 have shown that a experiment was carried out using the BCW data set .

Look from the three classification curve, as the K value is different, KNN method classification effect is not very stable.when K is equal to 25, there are 29 test sample is misjudgment. WKNN method has small fluctuations, and the classification result for IKNN method is more stable,which classification result is better than the other two.

The IKNN Classification algorithm has very good classification effect when it be used to category in the three common data set . Using IKNN method can improve the accuracy of classification, and has a good job stability in the feature extraction.

5. Conclusion

This paper has proposed an improved classification method based on KNN, named IKNN. IKNN use Similarity judgment method based on SNN,and avoid the shortcoming that KNN classification algorithm depends on the distribution of training samples, and the high density area samples will affect the judgment of category. The experiment has shows that IKNN is an effective method to improve the performance of categorization. In the future, we will continue work on the nearest neighbor clustering based on categories.

Acknowledgments

The work was supported the National Natural Science Foundation of China (11404088), and Natural Science Foundation of Hebei Province (F2015405011).

References

1. L.I. Kuncheva (2005) Combining Pattern Classifiers, Methods and Algorithms. Wiley Interscience: New York.
2. R. O. Duda, P. E. Hart and D. G. Stork (2001) Pattern Classification. John Wiley & Sons: New York.
3. Hastie, T., Tibshirani, R. (1996) Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal*, 18(6), p.p.607-616.
4. Domeniconi, C, Peng, J, Gunopulos, D.(2002) Locally adaptive metric nearest-neighbor classification. *IEEE Trans. Pattern* , 24(9), p.p.1281-1285.
5. Han, E-H.S, Karypis, G, Kumar, V.(2001) Text categorization using weight adjusted k – nearest neighbor classification. *Proc. Conf. Knowledge Discovery and Data Mining* , p.p.53-65.
6. Weinberger, K.Q, Blitzer, J, Saul, L.K. (2009) Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 24(10), p.p.207-244.
7. Pang-Ning Tan, Michael Steinbath, Vipin Kumar (2005). Introduction to Data mining. Addison Wesley: New Jersey .
8. L. Ertoz, M. Steinbach, V. Kumar (2003) Finding topics in collections of documents: a shared nearest neighbor approach. *Clustering and Information Retrieval*, 11, p.p.83-103.
9. *UCI Repository of machine learning .databases. http://www.ics.uci.edu/~mlearn/ MLRepository.html.*
10. Yuba, Yavuz (1998)Application of k-nearest neighbor on feature projections classi to text categorization. *Proc. Conf. on on Computer and Information Sciences*, p.p.234-246.