An Improved KNN Algorithm in Text Classification

Xiaoni Wang [1, a], Zhenjiang Zhang [2,b], Wei Cao [3,c]

[1]Key Laboratory of Communication & Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing, 100044, China

[2]Key Laboratory of Communication & Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing, 100044, China

[3]China Information Technology Security Evaluation Center

[a]09211167@bjtu.edu.cn, [b]zhangzhenjiang@bjtu.edu.cn, [c]caow@itsec.gov.cn

Correspondence should be addressed to Zhen-Jiang Zhang; zhangzhenjiang@bjtu.edu.cn

Keywords: Data mining; Classification algorithm; KNN algorithm

Abstract. Classification is make irregular tings regular according to different characteristics. Classification algorithm is a very important technology in data mining. The 21st is a century of information technology, people living in the world that has a variety of information. Classification techniques are used everywhere: we buy different products on different floors in the mall, different goods in different areas in supermarkets and so on. Also in industrial, aerospace, Internet and other areas, classification techniques can be seen. This article describes several existing classification algorithms. Their own characteristics were compared. As for traditional KNN algorithm's deficiencies when facing massive database has been improved. We proposed an improved KNN algorithm, and were compared in classification performance with the traditional KNN algorithm.

Introduction

With the development of science and technology, electronics and information technology has become an indispensable part of people's lives. We receive phone newspaper every morning, visit social networking sites when having a rest, search the Internet for information, all this things can't work without electronic information. When facing complex and vast amounts of information, how to efficiently derive useful information? Modern scientists have been committed to research this problem. Classification technology in the efficient use of information is particularly important. This paper discusses three main parts. They are some common classification techniques, KNN classification algorithm deficiencies and improvements, before and after improvement algorithm comparison. In the first part, we introduce three of the most traditional methods: "Rocchio algorithm"," Naive Bayes algorithm "," SVM algorithm ". The second part introduces the KNN algorithm. In the third part, the KNN algorithm has been improved, and the improvement of the algorithm are compared with the regular KNN[1,2].

The Classification Algorithm

2.1 Classification Algorithm Outlined

Classification algorithm is a very wide and important technology in the application areas of data mining. There has been a variety of sophisticated algorithms in the world. Classification is based on the characteristics of the data set to construct a classifier using the classifier to classify a sample of unknown class technique. Classifier construction process is generally divided into two steps: training and testing. In training phase, the computer analyze known classes of characteristics of the

sample set, suggesting classification rules for each class and defining a category satisfying the condition or model. In testing phase, the computer use the inferred model, classify the samples of unknown class. Test the classification accuracy.

2.2 Several Common Classification Algorithms

Currently there are some common types of classification algorithm, such as: decision tree, Rocchio, Naive Bayes, neural networks, support vector machines, linear least squares fit, KNN and so on. This article will introduce several algorithms and make a comparison[3,4].

(1)Rocchio. Rocchio algorithm should be thought of first when people think about the text classification problem and it also the most intuitive solution. Rocchio is an efficient classification algorithm and widely applied to text classification, query expansion and other fields. It is a method by constructing a prototype vector to get the optimal key. The basic idea is to average all samples in a category of documents to get a new vector which is called "centroid". centroid is the representative vectors of this category. When a new document needs to be determined which category it belongs to, compare the distance between the document and the center of mass, you can determine whether the document belongs to this category. Rocchio classifier is very intuitive and simple, easily understood by humans, so it is often used to compare different algorithms.

(2) Naive Bayes. Bayes algorithm focuses on the probability that the document belongs to a category. The probability of document belonging to a category is equal to the probability that each word in the document belonging to the category of integrated expression. To some extent, we can use the number that the word appears in the training documents roughly estimate the probability word of belonging to the category, thus making the whole calculation process to become feasible. Using Bayes algorithm, the main task of the training phase is to estimate these values. Naive Bayes algorithm assumes that each word is independent of other words in an article. The emergence of a word is unrelated with the emergence of another word, but in reality is obviously wrong. Naive Bayes algorithm requires a lot of training samples in order to obtain results. This makes the early artificial classification risen sharply in workload, also have a higher requirements for post-processing storage and resource.

(3) Support Vector Machine. Support Vector Machine is Cortes and Vapnik first proposed in 1995, it exhibit many unique advantages in the small size sample, nonlinear and high dimensional pattern recognition, and be able to promote the application to function fitting and other machine learning problems.

KNN Algorithm Introduction

3.1 KNN Algorithm Principle.

Only according to the class of the nearest one or several samples to determine the class of the samples needed to be categorized. Here the value of “several” is indeed the value of *K*. In normal condition, *K* should be an odd number to avoid collision. This is the main idea of the K-Nearest Neighbor algorithm. In principle, through the algorithm is relay on limit theorem. However we only take a few neighbor samples to determine the categorization. As is illustrated in Fig.1, after categorization, the value of K plays a very important role to determine to class of the circle in the center. Assume the K=3 and we will find there are two red triangle samples and a blue square one in the three samples (closest to the green circle). Hence the uncategorized center circle belongs to the class of the red triangle one. Assume *K*=5, there are 2 red triangle ones and three blue square ones in the five nearest samples. So the uncategorized center circle belongs to the class of the blue square one[5,6,7].

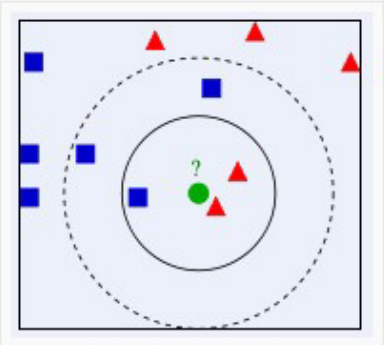

Fig. 1 KNN algorithm

The method to implement the KNN algorithm is as follows:

(1) According to a collection of feature words, describe the feature vectors of the training text.

(2) When the new texts come in, it will be categorized by the feature word to determine the description of the new texts.

(3) Compute the similarity between every text and the new text and the train the text to find out the most suitable *K*.

(4) Use *K* most similar texts to determine the class of the new text.

There are two methods to compute the similarity:

Euclidean distance: Euclidean distance between two standardized texts vectors *a*, *b* is computed by Eq.1:

$$D(a,b)=\sqrt{\sum_{i=1}(a_i-b_i)^2} \quad (1)$$

Cosine distance, the cosine of the angle between two vectors is calculated by equation Eq.2：

$$\cos<\vec{a},\vec{b}>=\frac{\vec{a}\bullet\vec{b}}{|a||b|} \quad (2)$$

The result of the KNN algorithm categorization is illustrated in Fig.2. The categorization result of the black point is the same class with “+” ones.

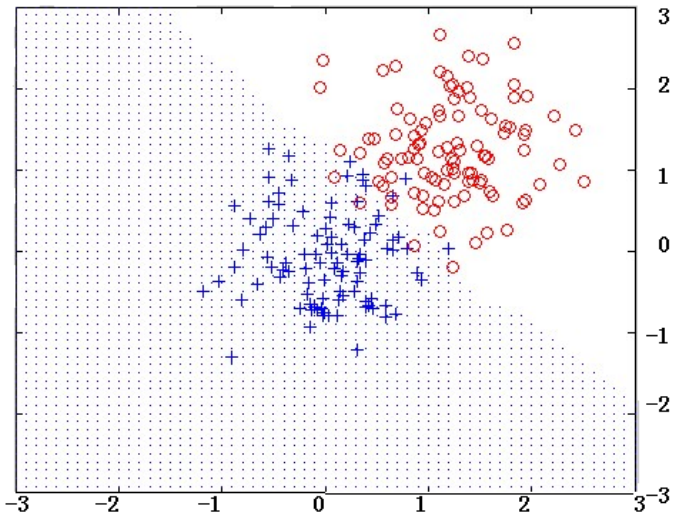

Fig.2 KNN result of the KNN algorithm categorization

3.2 Limitations of KNN Algorithm

KNN algorithm skips the process of training with simple implementation steps and high efficiency. However, the KNN algorithm should store all the samples in the very beginning. When categorization, we take out the samples to do some calculations like temporary classification and reduce dimensionality, which lead to large amount of calculation and low speed when the amount of samples becomes larger and larger. So the worst disadvantage of the KNN is when determining the class of a new document, it should be compared with all the current training documents, which cannot be afford by all the systems. According to this disadvantage, we make some improvements.

An Improved KNN Algorithm

4.1 A Introduction of the Improved KNN Algorithm

Shown in Fig.1, classify the center circle of the image. Blue squares and red triangles are sample files that already confirmed the categories. This example uses the distance represents similarity, the farther, the smaller. Assume the radius of solid line circle of 3. Located the closest triangle to the center is A. When finding the calculating distance of green text and A is less than 3, the sample which distance from A is farther than 5 is ignored. Namely, the samples can be discarded which is outside the circle that A as the center and 5 as a radius.

4.2 The Implementation of Improved KNN

The implementation of improved KNN are as follows:

(1) According to a collection of feature vectors re-training text description vectors.

(2) Calculating the similarity between each sample.

(3) When a new text arrives, segment the new text word according to the characteristics and identify new text vector representation.

(4) Calculate the new text and training text similarity. If the new text X1 and training text Ap (p <n) is greater than the similarity threshold *Sx*, then skip calculate X1 and Aq (q <n) (the smilarity between Ap and Aq is below the threshold value *Sn*, *Sx*> *Sn*.) .

(5) Selected *K* most similar samples in the training.

(6) Comparing the weight of the class, put the text into the largest category.

Fig.3 shows the classification results of the improved KNN algorithm. Where the black point of the classification results and "+" belongs to the same category.

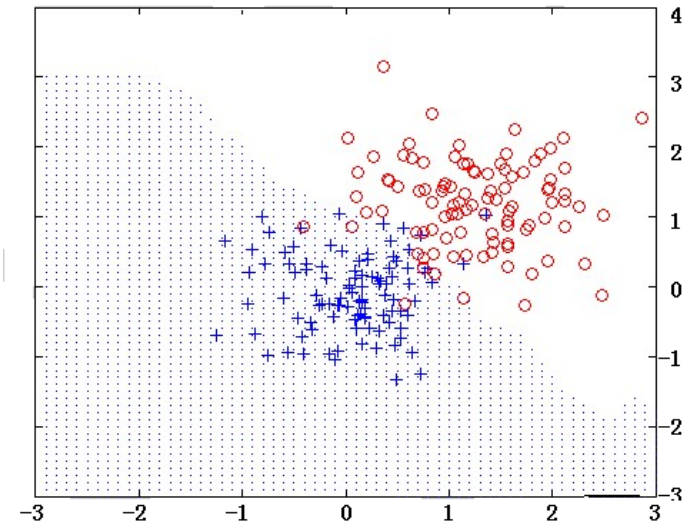

Fig.3 Improved KNN algorithm simulation results

4.3 Comparison

Compare Fig.2 and Fig3, we can find that the improved has little effects on the accuracy of the classification. For massive database, we can ignore the errors.

Assuming uniform distribution of sample corpus, *Sn* is distance corresponding to more than 80% similarity. When the sample anthology is 1000, the computation of improved KNN algorithm and the original KNN algorithm were compared in Fig. 4.

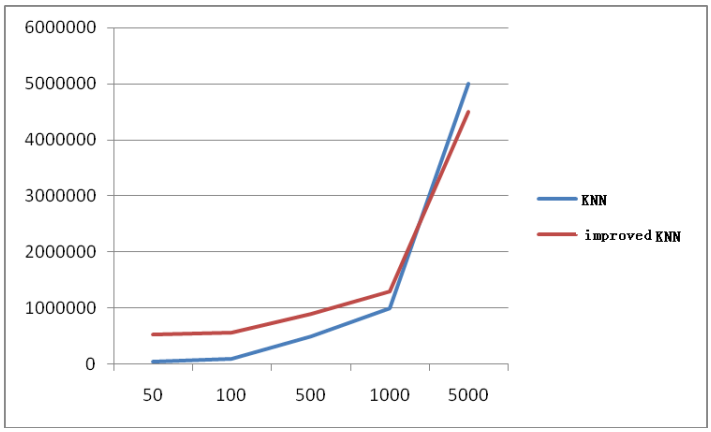

Fig. 4 Comparison when sample set is 1000 and the text to be classified is less

When the anthology to be classified increased, the computation of improved KNN algorithm and the original KNN algorithm were compared in Fig.5.

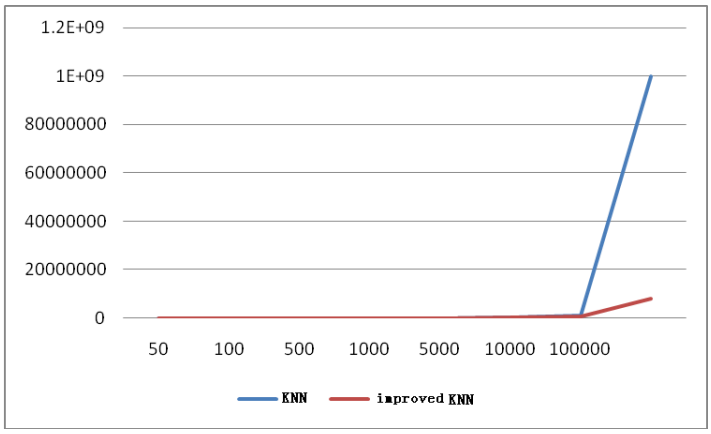

Fig.5 Comparison when sample set is 1000 and the text to be classified is much

When the sample anthology is 100, the computation of improved KNN algorithm and the original KNN algorithm were compared in Fig.6.

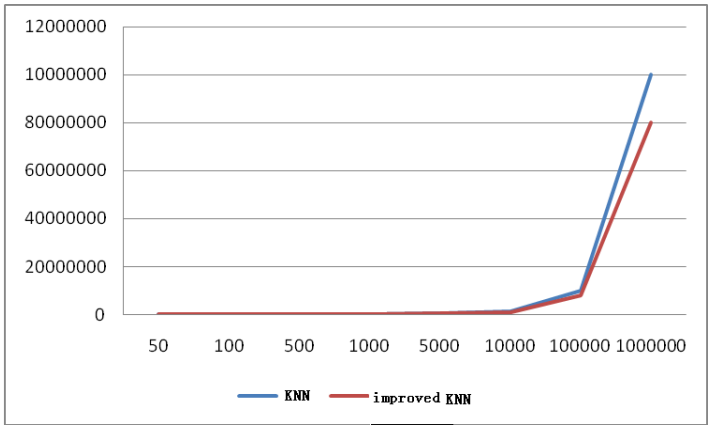

Fig.6 Comparison calculation when sample set is 100

After testing, the algorithm has the following advantages: saving calculation, improve computing speed, better for the huge amount of classification system. With the sample text to be classified increases, the calculation of the amount of the reduced amplitude increases. But there are also some disadvantages, such as: the choice of threshold have a significant impact on the classificationg, reducing the classification accuracy. It is better used for accuracy less demanded systems, such as mobile advertising push, web pages recommend and so on.

Conclusions

Classification algorithm is an important means in data mining process. This report makes some improvement to the current KNN algorithm and mainly focuses on the disadvantage of the calculation and the memory location. As the experimental results shows, through the advanced algorithm is almost the same compared to the original one on the accuracy, but the amount of calculations is reduced obvious. In the era of data explosion, the speed of the algorithm and the amount of memory location plays a very important role. The accuracy of the algorithm is the next step of improvement.

References

[1] Quinlan J R. C4. 5: Programs for Machine Learning [M].San Mateo, California: Morgan Kaufmann, 1993.

[2] Zait M, Messifah,A Comparative Study of Clustering Methds [J],FGCS,1997,13:149

[3] Yan Hu, Huzi Wu, Ge Zhong. Chinese Text Classification Based on POS for feature extraction method [J]. Wuhan University Technology,2007,29(4) :1322135.

[4] Hui Liu. The Chinese Text Classification based on KNN Algorithm [D]. Xinan Jiaotong University, 2010.

[5] Cha G-H, Zhu X, Petkovic D, Chung C-W. An efficient indexing method for nearest neighbor searches in high-dimensional image databases [J]. IEEE Transactions on Multimedia, 2002, 4 (1):76-87.

[6] Information on http://baike.baidu.com/view/1485833.htm

[7] Information on http://blog.sina.com.cn/s/blog_7fc305440101d0vr.html

Acknowledgement

This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant 20100009110002, the Beijing Science and Technology Program under Grant Z121100000312024.