Information Engineering and Applied Computing (2018)

ORIGINAL RESEARCH ARTICLE

A Survey of kNN Algorithm

Jingwen Sun,Weixing Du,Niancai Shi

Information Engineering College, Panzhihua University of Technology, Sichuan, China

ABSTRACT

The kNN algorithm is a well-known pattern recognition method, which is one of the best text classification algorithms. It is one of the simplest machine learning algorithms in machine learning classification algorithm. In this paper, we summarize the kNN algorithm and related literature, introduce the idea, principle, implementation steps and implementation code of kNN algorithm in detail, and analyze the advantages and disadvantages of the algorithm and its various improvement schemes. This paper also introduces the development of kNN algorithm, the important published papers. At the end of this paper, the application of kNN algorithm is introduced, and its implementation in text classification is emphasized.

KEYWORDS: kNN algorithm; k nearest neighbor algorithm; Machine learning; Text classification

1. Introduction

Classification is the core and basic technology in data mining. It has a wide range of applications in business, decision making, management, scientific research and other fields. At present, the main classification techniques include decision tree, Bayesian classification, kNN classification, artificial neural network and so on. In these methods, kNN classification is a simple, effective, nonparametric method, has been widely used in text classification, pattern recognition, image and spatial classification and other fields. This paper summarizes the kNN algorithm from all angles.

The structure of this paper is as follows:

In the second part, we mainly introduce the basic principles, ideas, implementation steps, Java implementation code and development course and classical thesis of kNN algorithm.

The third part is the discussion of many shortcomings of kNN algorithm, and gives some improved schemes.

The fourth part describes how the kNN algorithm handles multi-tag data.

The fifth part introduces the main application field of kNN algorithm, and highlights its excellent performance in text classification.

2. Introduction to 2kNN algorithm

2.1. Algorithm introduction

KNN algorithm is a simple learning algorithm inside the machine. The overall idea is relatively simple: calculate the distance between a point A and all other points, remove the k points with the nearest point, and then count the k points which belong to the classification The largest proportion, then point A belongs to the classification. Here's an example to illustrate:

Movie's name	The number of fights	Number of kisses	Movie type
California Man	3	104	Romance
He's Not Really into Dudes	2	100	Romance
Beautiful Woman	1	81	Romance
Kevin Longblade	101	10	Action
Robo Slayer 3000	99	5	Action
Amped II	98	2	Action

Simply talk about the meaning of this data: here with the number of fights and the number of kissing to define the type of film, as above, kissing more than the type of Romance and fighting more is the action movie. There is a name unknown (where the name is unknown to prevent the name from the name to guess the type of film), the number of fighting for 18 times, the number of kissing 90 times the film, which in the end it belongs to what type of film?

KNN algorithm to do is to use the number of times to fight and the number of kisses as the coordinates of the film, and then calculate the other six films and the distance between the unknown movie, made the former K distance of the nearest film, and then count the nearest k movie, Which belongs to the type of film up, such as Action up, then the unknown that the film belongs to the action movie type.

In actual use, there are several issues that are worth noting: K value of the selection, choose how much appropriate? Calculate the distance between the two, with which distance will be better? How much do you calculate? Assuming the sample, the type distribution is very uneven, such as Action movies have 200, but Romance film only 20. Calculated, even if not the Action movie, but also because the Action too many samples, resulting in k nearest neighbors There are a lot of Action movies, so how to do it?

There is no omnipotent algorithm, only in the optimal use of the environment in the algorithm.

2.2. Algorithm guiding ideology

KNN algorithm is the guiding ideology of 'near the red, near ink black', by your neighbors to infer your category.

The distance between the sample to be sorted and the training sample of the known category is calculated, and the k neighbors closest to the sample data to be sorted are found. The categories of the sample data to be classified are determined according to the category to which the neighbors belong.

2.3. Algorithm calculation steps

1. Calculate the distance: Given the test object, calculate the distance from each object in the training set;

2. Find neighbors: delineation of the nearest k training objects, as the test object of the neighbors;

3. Classification: According to the k kordia attribution of the main categories, to test the object classification.

2.4. Categories

Voting decision: a small number of subordinate to the majority of neighbors in which the category of points up to the class.

Weighted voting method: according to the distance of the distance, the neighboring vote to weight, the closer the distance the greater the weight (the weight of the reciprocal of the distance from the square)

2.5. Advantages and disadvantages

2.5.1 Advantages

1. Simple, easy to understand, easy to implement, no need to estimate parameters, no training;

2. Suitable for the classification of rare events;

3. Particularly suitable for multi-classification issues (multi-modal, the object has multiple categories of labels), kNN better than the performance of SVM.

2.5.2 Disadvantages

1. Lazy algorithm, the calculation of the classification of the test sample is large, memory overhead, the score is slow;

2. When the sample is not balanced, if the sample size of a class is large, and other sample size is very small, it may lead to the input of a new sample, the sample of K neighbors in the large-capacity class of the majority of samples;

3. Can be interpreted poorly, cannot give the decision tree as the rules.

2.6. Frequently Asked Questions

2.6.1k value setting

K value selection is too small, the number of neighbors is too small, will reduce the classification accuracy, but also to amplify the noise data interference; and if the value of k is too large, and the classification of the sample belongs to the training set contains fewer data classes, Then in the choice of k neighbors, the fact is not similar to the data is also included, resulting in increased noise caused by the reduction of classification effect.

How to choose the appropriate K value has become the research hotspot of KNN. The k value is usually determined by cross-checking (based on k = 1).

Experience rule: k is generally lower than the square root of the training sample number.

2.6.2 Determination of categories

The voting method does not take into account the distance of the nearest neighbor, and the closer neighbors may even decide the final classification, so the weighted voting method is more appropriate.

2.6.3 Distance measurement mode selection

The effect of high dimension on distance measurement: It is well known that the greater the number of variables, the worse the distinction of Euclidean distance.

The effect of the variable range on the distance: The larger the range of variables will often dominate the distance calculation, so the variables should be standardized first.

2.6.4 Guidelines for training samples

Scholars have studied the selection of training samples in order to achieve the purpose of reducing the calculation; these algorithms can be divided into two categories. The first class reduces the size of the training set. The KNN algorithm stores sample data that contains a large amount of redundant data that increases the cost of storage and computes the cost. The method of narrowing the training sample is to remove some of the sample samples that are not relevant to the classification in the original sample, to use the remaining sample as a new training sample, or to select some representative samples as a new training in the original training sample Samples, or clustering, the center of the cluster generated as a new training samples.

In the training set, some samples may be more worthy of dependency. You can apply different weights to different samples, enhance the weight of dependent samples, and reduce the impact of untrusted samples.

2.6.5 Performance issues

KNN is a lazy algorithm, and the lazy consequences: the constructor model is simple, but the overhead of the classification of the test sample is large because the training samples are scanned and the distance is calculated.

There have been some ways to improve the efficiency of calculations, such as compression training samples.

2.7. Algorithm flow

1. Prepare the data and preprocess the data

2. Use the appropriate data structure to store training data and test tuples

3. Set parameters such as k

4. Maintain a size of k by the distance from large to small priority queue, used to store the nearest neighbor training tuple. Randomly selecting the k tuples from the training tuples as the initial nearest neighbor tuples, calculating the distance from the test tuple to the k tuples, and storing the training tuple labels and distances into the priority queue

5. Traverse the training tuple set, calculate the distance between the current training tuple and the test tuple, and the distance L between the obtained distance L and the maximum distance Lmax in the priority queue

6. Compare. If L> = Lmax, the tuple is discarded and the next tuple is traversed. If L <Lmax, the element of the maximum distance in the priority queue is deleted

7. Group, the current training tuple into the priority queue.

8. After traversing, calculate the majority of the k tuples in the priority queue and use it as the category for the test tuple.

9. Test the tuple set after the completion of the calculation of error rate, continue to set different k value to re-training, and finally take the smallest error rate k value.

The Java implementation code for the 2.9kNN algorithm

2.8 Classical literature

The kNN algorithm is an improvement to the nearest neighbor algorithm of NN (nearest neighbor) algorithm. The initial neighbor algorithm is proposed by TM Cover in its article 'Rates of Convergence for Nearest Neighbor Procedures,' which is based on all training samples as punctuation, The distance between the test sample and all the samples is calculated and the nearest neighbor's class is used as the decision, and the scholar has made various improvements to the neighbor algorithm. One of the directions is the KNN algorithm, the original KNN algorithm is made by whom I now have two doubts, one is Trevor Hastie proposed KNN algorithm, I found his article Discriminant Adaptive Nearest Neighbor Classification, but in the article Finally, the author quotes R. Short \u0026 d K. Fukanaga's 'A'New Nearest Neighbor Distance Measure,' and claims that our use of our metric with W = I and E = 0, with B established by in a neighborhood of size R. Short has already proposed the concept of KNN, pending further reading of the relevant literature.

T.M. Cover	Nearest neighbor pattern classification, Proc. IEEE Trans. Information Theory, 1967 Rates of Convergence for Nearest Neighbor Procedures,Proc. Hawaii Int'l Conf. Systems Sciences,1968
C. J. Stone	Consistent nonparametric regression, Ann. Stat., vol.3, no. 4, pp. 595-645, 1977.
W Cleveland	Robust Locally-Weighted Regression and Smoothing Scatterplots, J. Am. Statistical Soc., vol. 74, pp. 829-836, 1979.
T. A. Brown & J. Koplowitz,	The weighted nearest neighbor rule for class dependent sample sizes, IEEE Trans. Inform. Theory, vol. IT-25, pp. 617-619, Sept. 1979.
R. Short &K. Fukanaga,	A New Nearest Neighbor Distance Measure, Proc. Fifth IEEE Int'l Conf. Pattern Recognition, pp. 81-86, 1980. The Optimal Distance Measure for Nearest Neighbor Classification," IEEE Trans. Information Theory 1981
J.P. Myles and D.J. Hand,	The Multi-Class Metric Problem in Nearest Neighbor Discrimination Rules, Pattern Recognition,1990
N. S. Altman	An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression,1992
Min-Ling Zhang & Zhi-Hua Zhou	Ml-kNN: A Lazy Learning Approach to Multi-Label Learning,2007
Peter Hall, Byeong U. Park & Richard J. Samworth	Choice Of Neighbor Order In Nearest-Neighbor Classification, 2008
Jia Pan & Dinesh Manocha	Bi-level Locality Sensitive Hashing for K-Nearest Neighbor Computation,2012

3. Algorithm improvements

Because of the earlier time of the proposed algorithm, the many shortcomings of kNN algorithm are gradually revealed; so many improved algorithms of kNN algorithm are emerge.

In view of the above algorithm, the improvement direction of the algorithm is divided into two aspects: classification efficiency and classification effect.

Classification efficiency: prior to the reduction of the sample attributes, delete the classification results of the smaller impact of the property, quickly get the category to be classified samples. The algorithm is more suitable for the classification of the relatively large category of sample size, and those with smaller sample size are more likely to produce errors.

Classification effect: the use of the weight of the method (and the distance from the small neighbors close) to improve, Han and others in 2002 to try to use the greedy method, for the file classification can be adjusted to adjust the weight of the nearest neighbor law WAkNN (Adjusted k nearest neighbor) to promote the classification effect; and Li et al. Proposed in 2004 due to the different classification of the file itself has a number of differences, it should also be in accordance with the training set of various categories of documents, select a different number of recent Neighbors, to participate in the classification.

The following describes the main direction of improvement, and then simply give an example of improved algorithm.

3.1. Main improvement direction

3.1.1 Improve the efficiency of the algorithm by reducing the computational complexity

The KNN algorithm stores all the sample data for the training set, which results in significant storage overhead and computational cost. There are a lot of literature proposed to reduce the calculation of the algorithm, these algorithms can be divided into two categories. The first class reduces the size of the training set. The KNN algorithm stores sample data that contains a large amount of redundant data that increases the cost of storage and computes the cost. The method of narrowing the training sample is to remove some of the sample samples that are not relevant to the classification in the original sample, to use the remaining sample as a new training sample, or to select some representative samples as a new training in the original training sample Samples, or clustering, the center of the cluster generated as a new training samples. The main method of the literature [25-26]. These methods filter the appropriate new training samples, and for the large training sample set, this workload is very great. The second category uses a fast algorithm to quickly search for K nearest neighbors. KNN algorithm to find K nearest neighbor point, you have to calculate the test point to the distance of all training samples, and then find out where K distance has the smallest data points, when the training sample is very large, KNN algorithm is impractical, In order to speed up the KNN search process, the main method, one of the methods is part of the distance calculation, in [27] proposed a KNN search algorithm based on the wavelet domain partial distance calculation, [28] proposed a fast algorithm (KWENNS). Another way is to introduce efficient indexing methods, efficient indexing method can greatly reduce the K nearest neighbor computing overhead, especially in high dimensional space is more obvious, the literature [29] proposed a new index Although some algorithms can effectively reduce the K nearest neighbor computing overhead and improve the classification speed of KNN, they cannot guarantee the global optimal search.

3.1.2 Optimization of similarity measures

The basic KNN algorithm calculates the similarity based on the Euclidean distance, which makes the KNN algorithm very sensitive to the noise characteristics. In order to change the flaws in the traditional KNN algorithm, we can assign different weights to the feature in the distance formula of measure similarity. The weight of the feature is generally set according to the function of each feature in classification. Can be weighted according to the classification of the feature in the entire training sample library, or can be weighted according to its classification in the local sample of the training sample (near the sample set of the sample to be tested). People have studied the methods of adjusting the weight of learning, thus improving the performance of KNN classifier.

3.1.3 Optimization decision strategy

One of the obvious drawbacks of the traditional KNN decision rule is that when the sample distribution density is not uniform, it is only a matter of miscarriage of justice that affects the performance of the classification according to the order of the former K neighbors regardless of their distance. Moreover, in the case of practical design of the classifier, because some categories are easier to obtain than other types of training samples, it tends to result in an imbalance between the various categories of training samples, that is, the number of training samples in each class is basically close The size of the regional differences, will also cause the distribution of training samples are not uniform. The improved method has the uniformity of sample distribution density. In [30], the decision rule of kNN is improved, which solves the problem of kNN classifier classification performance when all kinds of data distribution are not uniform. The use of a large number of neighbor sets instead of a single set of KNNs, and the relative support values are obtained by accumulating the support of the neighboring data sets for different categories, thus improving the adjacency rules.

3.1.4 Select the appropriate K value

Since almost all of the calculations in the KNN algorithm occur at the classification stage, and the classification effect is largely dependent on the selection of the k value, the choice of the k value is important. K value selection is too small, the number of neighbors is too small, will reduce the classification accuracy, but also to amplify the noise data

interference; and if the value of k is too large, and the classification of the sample belongs to the training set contains fewer data classes, Then in the choice of k neighbors, the fact is not similar to the data is also included, resulting in increased noise caused by the reduction of classification effect. How to choose the appropriate K value has become the research hotspot of KNN.

3.1.5 Multiple algorithm integration

In addition to the above methods, researchers have integrated KNN classification methods and other algorithms to improve the classification performance of KNN classifiers. The KNN, Grouping and LSA are integrated by integrating SVM and KNN. The genetic algorithm and fuzzy KNN are integrated. The Bayesian classifier and KNN classifier are integrated, and the P-tree and KNN are combined. Other algorithms are integrated to improve the classification performance of the KNN classifier.

3.2. An Improved KNN Algorithm for Optimizing Distance of Class 3.2 Correlation

The relationship between the entropy of the sample feature parameter and the probability of the sample distribution is taken as the correlation of the characteristic parameters for the classification, and the degree of influence of the characteristic parameters on the classification is calculated according to the correlation value. The distance between the samples is calculated to solve the KNN When the selection of large categories, high-density sample of the situation.

3.2.1 Optimization algorithm of K nearest neighbor for optimization of feature class correlation degree

Aiming at the weakness of KNN algorithm classification accuracy and efficiency, a new data preprocessing mechanism is proposed to improve the algorithm, and the concept of feature parameter class is introduced into KNN classification. A large number of studies have shown that the type of training data, if not subdivided, the number of categories is not a lot of features of the repetition rate is higher, especially for some commonly used data is so. If you use the dimension reduction method to calculate the similarity of the data, it will always lose too much and even important characteristic information because of the low dimension of the data vector or the unsatisfactory dimension, which will affect the classification effect. Based on this factor, the improved algorithm still uses the traditional KNN algorithm to set all the characteristic attribute values of the training set sample and the candidate sample as the similarity calculation parameter. The optimal distance mechanism is used to guarantee the accuracy of KNN classification and efficiency.

The realization of KNN improved algorithm based on class correlation optimization distance:

Input: The training set and the test set are expressed as,…

Output: The category label of the test set.

(1) According to the formula (1), we calculate the correlation degree of each feature parameter of each sample and the sample to be sorted in the training set, and quantize the data feature set from the Diff value, and carry out the correlation calculation of the sample parameters. Sample Feature Extraction Based on Class Correlation.

For (int i = 1; i <m; i ++)

// Calculate the correlation of the q characteristic parameters t contained by X;

For (j = 1; g;, ++)

If does not exist

Retrieve, collect the set of parameters, calculated by the formula (1, and store the value;

Storing a new matrix,

(2) Use formula (2) to calculate the distance between the sample to be sorted and the training set.

For (int i = 1; i <m; i ++)

// Calculate the distance between Ge and X;

For (j = 1; g; J ++)

By the formula (2 cumulative plus the average distance value.

(3) To determine the category of the sample attribution.

Sort by the distance and get the nearest k samples, based on the category of k samples.

3.3. KNN algorithm improvement based on clustering

The training set text is clustered, the training set text is divided into several clusters, and then the KNN algorithm is used to test the test documents. Finally, the KNN algorithm is used for several training sets in the nearest n clusters. Test the text for classification. Because of its reduced computational effort, the efficiency of execution is improved.

3.3.1 Clustering the training set

Step 1: If the text object P is not classified as a cluster or marked as noise, check its specified radius neighborhood r. If the number of objects contained in the specified radius neighborhood is greater than or equal to the given value m, Cluster C, add all points in the specified radius field r of p to cluster C;

Step 2: Check the specified radius neighborhood of all objects in C that have not yet been processed (classified as a cluster or marked as noise). If the number of objects contained in the neighborhood is greater than or equal to the given value m, The object in the neighborhood that is not classified as any cluster is added to C;

Step 3: Repeat the second step, continue to check the C was not processed object, until no new object to join the current cluster C:

Step 4: Repeat the above steps until all objects are processed.

Where the key parameter is the radius represented by the distance calculated as the density, the number of other points that are the minimum of the points that are contained within the specified radius. Through these two parameters we can calculate the density value around any point.

In this way, the training focused text is clustered into several classes. The category of each cluster depends on the majority of the text categories in the cluster.

3.3.2 Classification by kNN algorithm

Combined with the KNN algorithm, the distance between the test set text and the training group text cluster is calculated, which can reduce the influence of the calculation amount and the individual isolated points on the test set text.

Step 1: For any given test set text, calculate the distance from each cluster in the training set, using (2) for the test set text score

Where x is a test set text, c is the category of the training set, and t is one of the k text clusters closest to x. Is the similarity between the text x and the text t cluster, which refers to the distance. Is to say whether the t cluster belongs to class c, if it belongs to class c is 1, otherwise 0.

Step 2: Sort the results according to the score and select the first k clusters.

Step 3: Select the n closest text from the test set text from these clusters. According to (1), we judge the text of the test set and return to the traditional KNN algorithm.

In the improved algorithm, there is a radius r in the algorithm, the minimum number of m in the specified neighborhood, the number k of clusters, and the number of n texts with the smallest distance from the k cluster.

4. The 4kNN algorithm is used to classify multi - tag data

Data classification is an important branch of the field of data mining, which can be expressed as a function whose task is to give an ordered pair of Boolean values, where D is the definition field of the data instance, representing the predefined set of topic classes, also known as the label set). If the data instance 4 has a label, otherwise it will be strong. In practice, there are two types of data classification: the case of a single label is called single-label Categorization; a case with 0 to a label is called Multi-Label Categorization, Is the subject class owned by the data instance. The data classifier is fully automatic and semi-automatic. The automatic classifier makes a clear decision about T or F, which is called 'hard' categorization; and semi-automatic classifiers are given only T or F according to the likelihood of possession The Confidence Ranking, the final definite decision is made by the expert according to the result of the reliability ranking, called Ranking Categorization.

There are a large number of multi-tag data classification problems in Internet information resources such as text data, image data and music data. Sebastiani [15] and Tsoumakas [16] systematically reviewed the progress of text classification and multi-tag data classification in 2002 and 2007. According to the different ideas of the solution, the multi-label classification algorithm can be divided into two types: problem conversion and algorithm adaptation. The former transforms the multi-label classification problem into multiple single-label classification problems. The latter

is adopted by a single label such as C4.5 and AdaBoost The classification algorithm is extended to form a multi-tag classification algorithm. Recently, there have been some new multi-label classification algorithms [17] - [21], which have different characteristics of practical problems. Among them, ML and KNN of Zhang and Zhou [20] are a sorting algorithm of KNN (K-Nearest Neighbor) and Bayesian law, which has the advantages of simple thinking, nonparametricization and superior performance. Is that the calculation is large, the classification efficiency is low, so it is not suitable for high real-time requirements of the occasion.

5. Application of 5kNN Algorithm

KNN algorithm as one of the most classic machine learning classification algorithm, it must have its very wide range of applications. Here are just a few common applications, and focus on the following kNN algorithm in the text classification applications.

5.1. The main application areas of 5.1kNN algorithm

1) Pattern recognition, especially optical character recognition;

2) Statistical classification;

3) Computer vision;

4) Databases, such as content-based image retrieval;

5) Coding theory (maximum likelihood coding);

6) Data compression (MPEG-2 standard);

7) Wizard system;

8) Network marketing;

9) DNA sequencing;

10) spell check, suggest correct spelling;

11) Plagiarism investigation;

12) Similarity score algorithm, used to infer the athlete's professional performance.

5.2. The 5.2kNN algorithm deals with text classification

5.2.1 Introduction to text classification

The automatic text categorization is initially due to the requirements of the information retrieval (IR) system. With the popularity of the global Internet, the automatic classification of text for the meaning of information processing has become more important. On the Internet, electronic document information is increasing dramatically every day, through the network, people can easily share huge information resources. However, the rapid expansion of network information, information resources cannot be effectively used. In the face of the massive information on the Internet, the traditional approach is to classify online information and organize and organize them to provide a relatively effective means of information acquisition. However, this practice of artificial classification there are many drawbacks: First, spend a lot of manpower, material and energy; secondly, the classification results are not high consistency. Even if the classification of people's language quality is higher, for different people to classify, the classification results are still different. Even the same person, at different times to do classification may also have different results. The proliferation of network information increases the urgent need for fast, automatic text categorization. On the other hand, we have prepared sufficient resources for the text classification method based on machine learning. The automatic classification of electronic information processing technology is increasingly showing its superiority, text automatic classification and related technology research is increasingly becoming a research hotspot.

Text classification is mainly used in information retrieval, machine translation, automatic digest, information filtering, mail classification and other tasks. Text classification in the search engine also has a lot of use, web page classification / hierarchical technology is a key search system technology, search engines need to study how to classify the page, hierarchical, different types of Web pages using differentiated storage And processing to ensure that under limited hardware resources, to provide users with an efficient retrieval system, while providing users with relevant, rich search results. In the search engine, the text classification mainly has these uses: relevance sorting will be based on different types of pages to do the appropriate sorting rules; according to the page is the index page or information page, download scheduling will do a different scheduling strategy; When you take the search, you will be based on the results

of the page classification to do a different extraction strategy; in the search intention to identify the time, according to the user click on the url belongs to the category to infer the category of the search string.

5.2.2 Text classification process

Taking the text in the Internet as an example, the semi-formatted Web pages, documents, and the main forms of Internet information are stored in HTML format. Text knowledge mining is to find the implicit rules, in order to achieve the intelligent Internet data mining, leaving the text knowledge mining, intelligent cannot be achieved. The most commonly used method of text knowledge mining is based on the Characteristic Vector Space Model (CVSM).

1 Document model is established

1) Preprocessing process. The second is to use the feature dictionary set (including universal set and professional set) for word segmentation, if there is no word in the word set, then it as a whole as a whole Words and records for artificial word segmentation.

2) Conceptual mapping and conceptual disambiguation. Some words are different but have the same concept. They are required to map them to the same concept according to the concept set. For unregistered words, the word with the highest coexistence rate is chosen as the concept. For the word with multiple concept labels, The most frequent occurrence of its marked.

3) General feature extraction and name date and other features extracted, the results stored in the document vector library.

4) Feature set reduction. The number of feature sets obtained by the above method is huge, so it must be reduced. The algorithm is to construct an evaluation function, evaluate each eigenvector, and then select a subset of the eigenvectors of a certain number or over the threshold according to the size of the evaluation value. The result of the feature set reduction is stored in the document vector library.

2 Knowledge discoveries

1) Text summary. The basic idea is that the sentence is closely related to the theme of the sentence selected; such sentences are often located in a special part or contain more features, the general weight of the sentence as the evaluation criteria.

2) Text classification. Text classification is the main purpose of text knowledge mining. The basic idea is to compare training set, vector set with document vector set. There are Naive Bayesian classification algorithm and K-nearest neighbor classification algorithm.

3 Model evaluation

The text evaluation model is more; usually the data set is divided into training set and test set. Learning - test cycle repeated implementation, and finally with an indicator to measure the quality of the model. The specific evaluation index of the model has classification accuracy rate, precision rate, recall rate, precision rate, recall rate of the average, and information valuation.

5.2.3 kNN algorithm to achieve text classification

One of the key questions in text automatic classification is how to construct a classification function (classifier) and use this classification function to classify the text to be classified into the corresponding category space. Training methods and classification algorithms are the core of the classification system. Here we introduce KNN classification algorithm to classify text knowledge.

The basic idea of the kNN algorithm is to consider the K text of the recent (most similar) distance of the new text in the training text, and to determine the category of the new text according to the category to which the K text belongs. The algorithm steps are as follows:

1) Re-describe the training text vector according to the feature set;

2) After the arrival of the new text, according to the new words of the word, to determine the new text vector representation;

3) Select the K text that is most similar to the new text in the training text set.

4) In the new text of the K neighbors, in turn calculate the weight of each class.

5) Compare the weight of the class, the text assigned to the weight of the largest category.

References

1. Cover T, Hart T P. Nearest neighbor pattern classification [J]. IEEE, 1967 (1): 21 - 27.
2. Cover T. Rates of convergence for nearest neighbor procedures [J]. Systems Sciences, 1968.
3. Stone C J. Consistent Nonparametric Regression [J]. Institute of Mathematical Statistics, 1977 (7), 5 (4): 595-620.
4. Cleveland W S. Robust locally weighted regression and smoothing scatterplots [J]. Journal of the American Statistical Association, 1979, 74: 829-836.
5. Brown, T., Koplowitz, Jack. The weighted nearest neighbor rule for class dependent sample sizes (Corresp.) [J]. IEEE, 1979 (9) .IT-25: 617 - 619.
6. Short R, Fukunaga K. A new nearest neighbor distance measure [J]. IEEE, 1980: 81-86.
7. Robert D .; Fukunaga, K. The optimal distance measure for nearest neighbor classification [J]. IEEE, 1981 (9), 27 (5): 622 - 627.
8. Myles J, Hand D. The multi-class metric problem in nearest neighbor discrimination rules [J]. Pattern Recognition, 1990, 23 (11): 1291-1297.
9. Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression [J]. 1992, 46 (3): 175-185.
10. Zhang M, Zhou Z. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007 (7), 40 (7): 2038-2048.
11. Hall P, Samworth B. Choice of neighbor order in nearest-neighbor classification [J]. The Annals of Statistics, 2008 (10), 36 (5): 2135-2152.
12. Pan J, Manocha, D. Bi-level locality sensitive hashing for k-nearest neighbor computation [J]. IEEE, 2012 (4): 378-338.
13. Michel M Deza, Elena Deza. Encyclopedia of Distances. Springer, 2009
14. Zhou J, Liu J. A KNN algorithm for optimizing distance using class correlation. Journal of Computer Applications, 2010 (11), 31 (11): 7-12.
15. Sebastiani F. Machine learning in automated text categorization [J] .ACM Computing Surveys, 2002, 34 (1): 1-47.
16. Zhao Jidong, Lu Ke, Wu Yue. A web image search method based on spectral theory [J]. Computer Applications Research, 2008 (5): 12-13.
17. Zhang Hua. Research on image semantic information extraction method [J]; Journal of Shandong Normal University;
18. Wen Xiaobin. Research and implementation of interact image search engine [D]. Haikou: Hainan University, 2006.
19. Cai Dang, He Xiaofei, Li Zhiwei, et al. Hieraehical clustering of WWW image search results using visual Textual and link information [C]. Proceedings of the ACM International Conference on Multimedia, New York, USA, 2004: 952-959.
20. Cheng En, Jing Feng, Zhang Chao, et al. Search result clustering based relevance feedback for web image retrieval [C]. Interactional Conference on Acoustics, Speech, and Signal Processing, Hawaii, 2007: 961-964.
21. Xie Tong. Based on the text of the Web image search engine research and implementation [D]. Chengdu: University of Electronic Science and Technology, 2007.
22. Cai D, Yu S, Wen J R, et al. VIPS: a vision-based page segmentation algorithm, MSR-TR-2003-79 [R] .Microsoft Research, 2003.
23. Kang Shiyong, Liu Yan. On the semantic components and semantic sentences of Chinese verb predicate sentences [J] .Journal of Tang University, 1998,14 (1): 89-93.
24. Xu Bin. Semantic sentence recognition based on PCFG-HDSM model [D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2008.
25. P E Har. The condensed nearest neighbor rule. IEEE Trans on Information Theory, 1968, IT-14 (3): 515-516.
26. Li R, Hu Y. KNN text classifier training sample crop method based on density. Journal of Computer Research and Development, 2004, 41 (4): 539-546.
27. W J Hwang, K W Wen. Fast KNN classification algorithm based on partial distance search [J]. Electron lett, 1998, 34 (21): 2062_2063.
28. J S Pan, Y L Qiao, S H Sun. Neighbors classification algorithm [J]. IEICE Trans Fundamentals, 2004, E87-A (4): 961-961.
29. Hou Shijiang, Liu Chehua, Yu Jing, Chu Bingyi. K nearest neighbor query algorithm in spatial network database. Computer Science, 2006Vol.33No.8.
30. Sun Qiuyue. KNN algorithm based on SNN similarity degree. Master's Thesis, Yunnan University, 2008.
31. H. Wang. Nearest Neighbors without k: A Classification Formalism based on Probability, technical report, Faculty of Informatics, University of Ulster, N, Ireland, UK, 2002.