AN INFECTIOUS DISEASE PREDICTION METHOD BASED ON K-NEAREST NEIGHBOR IMPROVED ALGORITHM

Yaming Chen[1], Weiming Meng[2], Fenghua Zhang[3],Xinlu Wang[4] and Qingtao Wu[5]

[1,2&4]Computer Science and Technology, Henan University of Science and Technology, Luo Yang, China
[3]Computer Technology, Henan University of Science and Technology, Luo Yang, China
[5]Professor, Henan University of Science and Technology, Luo Yang, China

ABSTRACT

With the continuous development of medical information construction, the potential value of a large amount of medical information has not been exploited. Excavate a large number of medical records of outpatients, and train to generate disease prediction models to assist doctors in diagnosis and improve work efficiency.This paper proposes a disease prediction method based on k-nearest neighbor improvement algorithm from the perspective of patient similarity analysis. The method draws on the idea of clustering, extracts the samples near the center point generated by the clustering, applies these samples as a new training sample set in the K-nearest neighbor algorithm; based on the maximum entropy The K-nearest neighbor algorithm is improved to overcome the influence of the weight coefficient in the traditional algorithm and improve the accuracy of the algorithm. The real experimental data proves that the proposed k-nearest neighbor improvement algorithm has better accuracy and operational efficiency.

KEYWORDS

Data Mining,KNN, Clustering,Maximum Entropy

1. INTRODUCTION

Data mining, as a new technology for data information reuse, can effectively provide important information for hospital management decisions [1]. It uses the database, artificial intelligence and mathematical statistics as the main technical pillars for technical management and decision-making [2]. In the process of medical information management, application data mining technology can better organize and classify medical and health information to establish a management model, which can effectively summarize data and provide valuable information for the efficient operation of medical work [3].

The prediction method based on medical big data is still in the stage of contending, and can be divided into two categories: one is based on data to train and generate predictive model [4-6], which is based on supervised machine learning for specific purposes. Methods, such as classification, regression, deep neural network, etc., generate training models to predict clinical events in unknown patients; the other is to measure patient distances, establish patient similar groups, and predict target patients by similar group characteristics In the case of this, this is called

patient similarity analysis [7-11]. Specifically, patient similarity analysis refers to the selection of clinical concepts (such as diagnosis, symptoms, examination, family history, past history, exposure environment, drugs, surgery, genes, etc.) as characteristics of patients in a specific medical environment. Quantitative analysis of the distance between concepts in the complex concept semantic space, thereby dynamically measuring the distance between patients, screening out similar groups of patients similar to index patients. In the medical big data scenario, the various features of the patient-like group can theoretically provide multiple predictions, and have better universal characteristics than the training model for specific goals.

An improved weighted KNN algorithm uses a distance from the sample to be identified to each test sample to weight it, that is, a path weighting.That is to say, we give a path weight variable. When our sample to be identified is close to which test sample, we give it a larger weight, and instead give it a small weight [12]. The semi-supervised SVM-KNN classification method makes full use of unmarked data, and through basic experiments, when the known training information is insufficient, the fusion training of a large number of unmarked data can improve the final classification accuracy [13]. The KNN algorithm is parallelized on the Spark platform, which can effectively solve the problem of low efficiency of searching historical database in the neighborhood search process of KNN algorithm and improve the computational efficiency of the algorithm [14]. The above methods only consider efficiency or accuracy and do not fully integrate.

In this paper, based on clustering and maximum entropy, K-nearest neighbor algorithm based on Clustering and Maximum Entropy (CME_KNN) is proposed. In the data preprocessing, clustering is used to extract clusters. The generated samples near the center point are applied to the K-nearest neighbor algorithm as a new training sample set, and the improved K-nearest neighbor algorithm based on the maximum entropy preserves all the uncertainties, closest to the natural state, without The weight is determined, so the formula is not affected by subjective factors, overcoming the significant shortcomings of traditional algorithms. Based on the data training results, design feedback mechanism, continuously optimize the model, and gradually improve the accuracy of the prediction results.

2. RELATED WORK

2.1 DATA MINING

Data mining combines theories and techniques of high-performance computing, machine learning, artificial intelligence, pattern recognition, statistics, data visualization, database technology, and expert systems [15]. The era of big data is both an opportunity and a challenge for data mining. Analyze big data, establish appropriate systems, continually optimize and improve the accuracy of decision-making, so as to be more conducive to mastering and adapting to the multi-terminal changes in the market [16]. In the era of big data, data mining has been recognized as the most commonly used data analysis method in various fields. At present, domestic and foreign scholars mainly study the application of classification, optimization, recognition and prediction in data mining in many fields.

2.1.1 Classification.

Along with the progress of the times and the rapid development of science and technology, as a populous country, China's public data generated in health care and aging society has grown

geometrically, and the value of the data based on big data is urgently needed solve.The structure, scale, scope and complexity of health care data are constantly expanding. Traditional calculation methods cannot fully satisfy the analysis of medical data. Data mining technology can be based on some characteristics of medical data: pattern polymorphism, information Loss (missing values in the data due to personal privacy issues), timing, and redundancy classify health care data to provide accurate decision-making for doctors or patients [17].

At the same time, China is accelerating its entry into an aging society, and the Internet is an important medium for improving an aging society. Big data is an important technical means of assessing an aging society. Qu Fang et al. [18] proposed the “Internet big data” model for the realization of old-age care. The entire old-age service system is based on multi-dimensional heterogeneous information aggregation and data fusion mining. The “Internet Big Data” pension system integrates a variety of information and communication technologies, including communication technology, data mining technology and artificial intelligence technology.

2.1.2 Optimization.

The traffic conditions of the roads are closely related to people's travel. With the rapid development of the city and the improvement of living standards, the scale of motor vehicles has gradually expanded, which has caused problems such as traffic congestion. Data mining technology can effectively solve the optimization problem between traffic roads and logistics networks. Pan et al. [19] proposed a data mining forecasting model, which is used to “predict real-time” short-term traffic conditions and drive to traffic jams. The staff brought great help.
With the development of technology, online shopping has become more and more popular, and at the same time, it has brought problems such as congestion and embarrassment of logistics and transportation. Jingdong, one of China's largest online trading platforms, uses the UAV to detect road condition feedback data in the era of artificial intelligence optimization, and uses data mining technology to accurately calculate the parameters required for logistics network transportation, which can easily and efficiently alleviate logistics. The problem of transporting cockroaches led to the first robotic courier in China to deliver the first item to Renmin University of China. With the increase of the length and complexity of the traffic network in the future, the difficulty of implementing the unmanned automation strategy is also greatly increased. Only through data mining technology can the results be quickly calculated, and the high value generated from the complex road information can be obtained.

2.1.3 Identification.

Since the advent of digital images in the 1950s, digital images have become an indispensable "data" in human society. In computer applications, data mining is becoming more and more popular in image recognition applications. Representative applications are face recognition and fingerprint recognition. Face recognition further analyzes and processes reliable and potential data by data mining of the obtained information base, and fully prepares the data analysis work and future development work. Wright et al. [20] described robust face recognition based on sparse representation, and gave a detailed theoretical analysis and practical summary.

Sha Yaqing et al. [21] proposed an identity authentication scheme based on smart card and fingerprint identification for the insecure use of username and password in the current electronic tax filing system, and combined with fingerprint technology to construct new password parameters, thus making The safety is significantly improved. With the continuous development

of data mining technology, the accuracy of big data recognition of faces and fingerprints will become higher and higher.

2.1.4 Forecast.

Forecasting problems are the most studied issues in various fields, and their purpose is to predict future data values or trends through historical data. Most of the historical data is time series data, which means that they are arranged in chronological order and a series of observations are obtained. Due to the continuous advancement of information technology, time series data has also increased dramatically, such as weather forecasting, oil exploration, and finance. The ultimate goal of time series data mining is to predict the trend of the future and its impact by analyzing historical data of time series.

"Meteorology" is closely related to the ecological balance of the earth and people's normal life. Therefore, accurate forecasting of meteorology is particularly important. Zhou Lei et al [22] summarized the current meteorological monitoring model, based on the drought of remote sensing data, classified the current remote sensing monitoring methods, classified the external environmental conditions (temperature, humidity, etc.) and proposed solving complex problems. new method.

As a non-renewable resource, oil is currently shrinking global reserves, making oil exploration more and more important. In petroleum exploration management, the collected data has the characteristics of large amount of data, large amount of calculation, single source of acquisition and complex data processing flow [23], and high-performance parallel computing of large data sets collected by data mining technology. Analysis can ensure the validity and accuracy of the results.

In the era of big data, daily business such as banks, securities companies, and insurance companies will generate massive amounts of data. Using the current database system, data entry, query, and statistics can be efficiently implemented. Currently, from simple queries to It is especially important to use data mining technology to mine knowledge and provide decision support. The application of data mining technology in the financial industry is feasible. The application of the theoretical basis to relevant examples includes forecasting stock indices, discovering implicit patterns in financial time series, credit risk management and exchange rate forecasting.

2.2 KNN ALGORITHM

The K-nearest neighbor (KNN) is also called the nearest neighbor method, which selects a certain distance function and then calculates the distance between the test sample and the training sample in the multidimensional space, and selects the nearest distance from it[24]. The K training sample points are finally judged according to the category of the training sample points [25].The concept of tree structure is introduced into the KNN algorithm to form a K-nearest neighbor on tree (TKNN).

The KNN algorithm is different from positive classification algorithms such as neural networks and support vector machines. It is a passive classification method. That is to say, the classification algorithm does not need to learn the rule information according to the training sample data in advance, and only performs the calculation classification when the classification task appears.

The rule of the KNN algorithm is to train the data sample itself. Therefore, the KNN algorithm takes almost no time in the training phase, and its time consumption is used in the classification phase. Moreover, as a non-parametric classification method, the KNN algorithm has the advantages of simple and intuitive, easy to implement, high classification accuracy and robustness for unknown and non-normally distributed data. Although the KNN method relies on the limit theorem in principle, it is only related to a very small number of adjacent samples in class decision making. Since the KNN method relies mainly on surrounding neighboring samples rather than relying on the discriminant domain method to determine the class of the test sample, the KNN method is more suitable than other methods for the sample set to be cross-over or overlapped.

2.3 TF-IDF ALGORITHM

The TF-IDF algorithm is a statistically based method to measure the keyness of a word or phrase in textual information. Its main principle: a word is used more frequently in the target text, but it is used less frequently in the corpus, then it can have good text distinguishing ability [26]. The TF value of a word in the target text refers to how often the word appears in the text. When calculating this frequency, you need to normalize it to prevent it from being biased towards long text with a large number of words. The formula for calculating the TF value is as follows:

$$TF_{\mathrm{i},j} = \frac{f_{i,j}}{\sum_{i=1}^{n} f_{i,j}} \tag{1}$$

Equation 1 is the calculation of the word frequency TF_{i} of the word i in the text j, where the molecular content is the number of times the word i appears in the text, and the denominator content is the total number of all words in the text j. Through the calculation method of the word frequency, the bias of the word frequency TF_{i} on the longer text is effectively prevented. In addition, another word frequency normalization method can be used, such as formula 2. Where the denominator represents the maximum word frequency value that appears in the text j [27], ie

$$TF_{\mathrm{i},j} = \frac{f_{i,j}}{f_{\max,j}} \tag{2}$$

The critical measure of a word can be expressed in terms of IDF values. Assuming that the number of texts containing the word i in a text corpus collection is smaller, the IDF_i value of the word should be larger, so that the better the distinguishing ability of the word i, the more likely it is to be a keyword. The specific formula is as follows:

$$IDF_{\mathrm{i}} = \log \frac{N}{n_i + 1} \tag{3}$$

Where N is the total number of texts in the text set and n_i contains the number of texts in word i. The +1 in the denominator is to deal with the case where the denominator is 0 in the formula.

The calculation by the above method can be summarized into the formula 4. Suppose a word i appears in a given text j with a higher frequency and less text in the entire text set containing the word, then its TF-IDF value is higher, that is, the word i is easier to distinguish the document j, can be used as a keyword.

$$TF\text{-}IDF_i = TF_{i,j} \times IDF_i \quad (4)$$

For the patient's text medical data information, the ik tokenizer can be used for word segmentation. For the special treatment of the word segmentation of medical data, the proprietary medical terminology can be included, put into the custom word segmentation dictionary, and then the TF-IDF algorithm is used to extract the feature of the text content. The past history, current medical history, complaints and physical examination data of the outpatient medical records were extracted according to the above methods, and digitally represented.

3. TEST METHODS

3.1 SYSTEM ARCHITECTURE

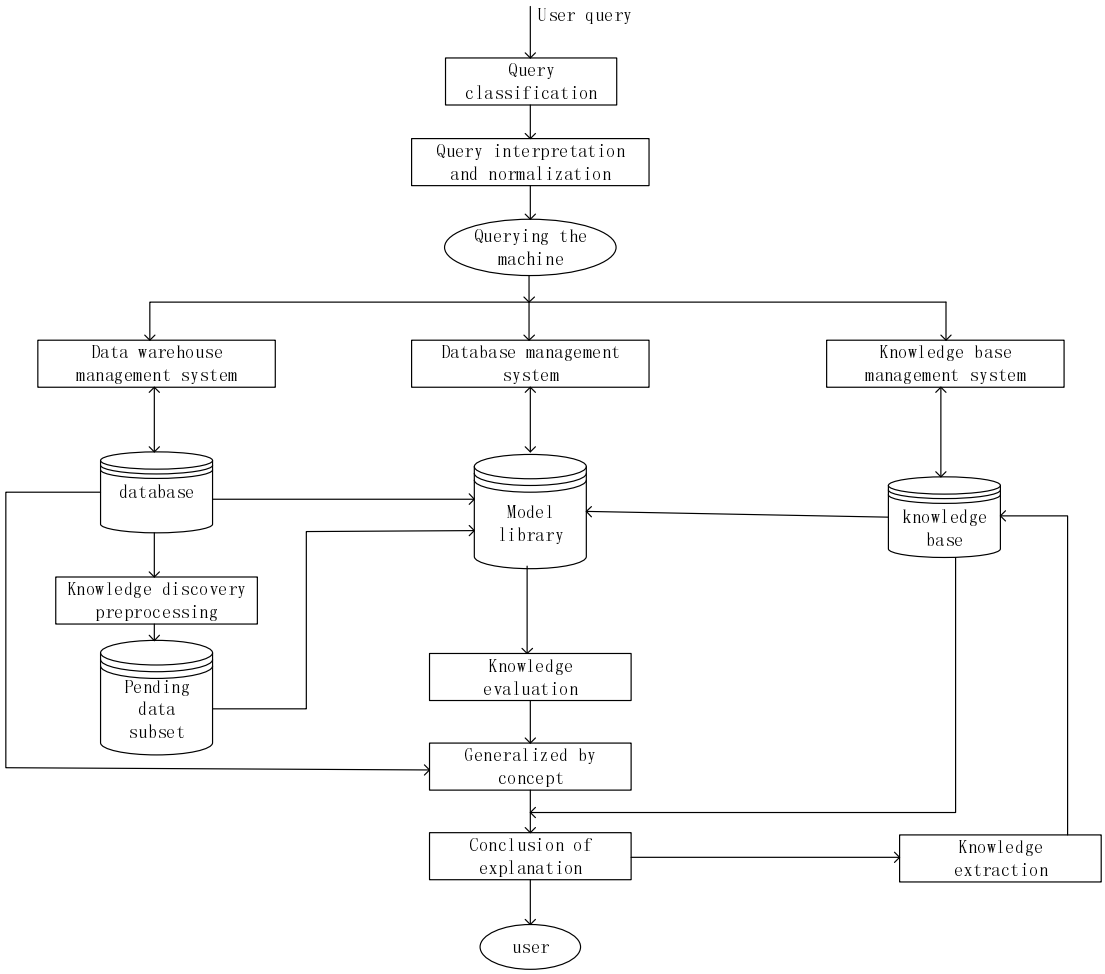


Figure 1. Structure of a general data mining system based on data warehouse

A structural framework of a general data mining system based on data warehouse is used to handle various types of data mining tasks. The model is mainly composed of user query interface, query collaborator, data warehouse management system, knowledge base management system, model library management system, knowledge base discovery preprocessing module, knowledge evaluation module and conclusion expression module.

The functions of each part are briefly described as follows:

Query interface: The knowledge discovery request submitted by the user is interpreted into a normalized query language through certain steps, and is processed by the query synergy machine. The work of the query interface can be divided into two parts: query classification, query interpretation and normalization.

Query coordinator: collaborative data warehouse management system, model library management system and knowledge base management system, together to process query requests submitted by the query interface. Querying the collaboration machine is an important part of the system.

Data warehouse management system: directly responsible for the management of the data warehouse, and complete the extraction of data in various heterogeneous distributed data sources, to shield the impact of various heterogeneous data sources on the system to the maximum extent.
Knowledge Base Management System: Manage and control the knowledge base, including the addition, deletion, update and query of knowledge. On the one hand, accepting the knowledge base query request generated by the querying machine and performing the query, and submitting the result to the knowledge discovery module; on the other hand, accepting the knowledge pattern obtained by the knowledge extraction module from the final comprehensive conclusion, depositing Knowledge base to enrich the content of the knowledge base.

Model library management system: Manage the model library. An important part of the model library is the knowledge discovery module, which includes various data mining tools, such as: association rule discovery sub-module; classification rule discovery sub-module; cluster analysis discovery sub-module; feature rule discovery sub-module, model library at the same time It can also manage various types of analysis tools. This unified management of data mining tools and other analysis tools enables data warehouse-based data mining systems to be extended to decision support systems.

Knowledge discovery preprocessing module: Under the cooperation of the data warehouse management system, according to the metadata and dimension tables, the data stored in the entire data warehouse is processed to generate the requirements that meet the requirements of the user query and meet the requirements of the knowledge discovery tool set. Process data subsets.
Knowledge Evaluation Module: Evaluate the patterns found during the data mining phase. At this time, if there is a redundant or irrelevant mode, it needs to be culled; if the mode does not meet the user's requirements, then the entire discovery process needs to be returned to the discovery stage, the data is re-selected, and a new data transformation method is used to set a new one. Data mining parameter values, even replacing data mining algorithms.

Conclusion Expression Module: Generalize the conclusions obtained according to the semantic hierarchy, and draw conclusions at each semantic level: then explain the conclusions obtained,

and present the discovered patterns to the users in a visual way, or Transform the results into a natural language form that is easy for users to understand.

3.2 BUILDING KNN BASED ON CLUSTERING AND MAXIMUM ENTROPY

3.2.1 CLUSTERING TO BUILD A STREAMLINED SAMPLE SET

Clustering technology is a very effective data analysis method. Clustering algorithm refers to the clustering analysis of data without any data prior information. This algorithm is also called unsupervised learning method. Clustering learning is one of the earliest methods used in pattern recognition and data mining tasks, and is used to study large databases in various applications, so clustering algorithms for big data are receiving more and more attention [28] . In many practical problems, traditional clustering algorithms sometimes fail to obtain effective clustering results because there is no prior analysis of the data itself. In some practical problems, a priori knowledge of a small amount of data is obtained, including class labeling and data point partitioning constraints (such as pairwise constraint information), how to use these only a small amount of prior knowledge to a large number of no priors Clustering analysis of knowledge data has become a very important and urgent problem to be solved.

In the data preprocessing, the clustering method is used to extract the samples near the middle point generated by the cluster, and these samples are applied as a new training sample set in the K-nearest neighbor algorithm.

Firstly, each type of training sample set is clustered separately, and then the sample near the center point of the cluster is used as a subclass of the category to represent all the samples of the category, so that the classification accuracy can be ensured. The number of original training samples is greatly reduced, and the calculation speed of the K-near algorithm is greatly improved. Because CURE [29] (Clustering Using Representatives) clustering algorithm solves the problem of preference sphere and similar size, it also has excellent performance in eliminating isolated points (noise). This paper uses CURE clustering algorithm. However, the premise of using the CURE algorithm is that the number of clusters N needs to be provided, and the number of clusters clustered in practical applications is often unpredictable, and a range will be given to N in the algorithm, such as [1, 8]. The clustering calculation is performed for each N in the range, and then the classification accuracy of the clustered model is tested one by one, and finally the N value with the highest classification accuracy is selected.

3.2.2 CONSTRUCTING A DISTANCE METRIC FUNCTION BASED ON MAXIMUM ENTROPY

Most of the improved KNN algorithms for the distance metric function are based on the improvement of the Euclidean distance or the cosine of the vector angle, ie the weight adjustment factor is increased. The determination of feature weights may be determined based on the effect of each feature on the classification, or may be set by the role of the feature in the training sample set. Therefore, the introduction of the weight adjustment coefficient has certain subjectivity, which naturally affects the accuracy of the classification.

Entropy is originally a concept used to describe the degree of molecular chaos in thermodynamics. The more chaotic the molecule, the larger the entropy. With the advent of information theory, entropy has been introduced into information science [30]. In 1948, Shannon proposed the concept of information entropy, and believed that people's understanding of things

is uncertain. This degree of uncertainty is called information entropy. The information entropy of a random event is defined as: the random variable η, η has n possible cases ($\eta_1, \eta_2, \mathrm{L}\ , \eta_n$), and the probability of each case is $p_1, p_2, \mathrm{L}\ , p_n$, then it is not Determine the degree, that is, the information entropy is:

$$H(\eta) = -\sum_{i=1}^{n} p_i \log p_i \tag{5}$$

Similarly, the continuous variable η, the density function is $p_{(\eta)}$ then the information entropy is defined as:

$$H(\eta) = -\int_{\Omega} \mathrm{p}(\eta) \log p(\eta) d\eta \tag{6}$$

The principle of Maximum entropy (ME) was proposed by Jaynes in 1957. He pointed out that when only partial information of unknown distribution is grasped, the distribution obtained when entropy is maximized is the most objective and closest to the actual distribution. The most uncertain is the most unbiased, the principle of maximum entropy emphasizes that the determined probability distribution should be consistent with the known partial information, that is, the measured data or sample, without any subjective assumptions or estimates for the unknown information. [31]. It can be seen that the probability distribution obtained according to the principle of maximum entropy is the closest to the actual distribution.

In recent years, the concept of maximum entropy has been widely used in many fields, such as information theory, pattern recognition, and transportation. It represents a measure of the uncertainty of things. When the entropy is maximum, it means that the most uncertain things, that is, the closest to the actual state, which is the main advantage of the maximum entropy method. Maximum entropy is a general method for solving the inverse problem of a solution that cannot be determined due to insufficient data. It is often used for linear and nonlinear model estimation. It turns out that the maximum entropy essence is a kind of similarity measure, that is, the distance measure between the observed value and the actual value. This kind of measure is more effective when the sample data are positive. Therefore, we introduce the maximum entropy instead of the Euclidean distance as the distance measure function of KNN. Experiments show that the method can effectively classify the data.

The most striking feature of the maximum entropy technique is that many different features can be integrated into a probabilistic model, and there is no independent requirement for these features, and the maximum entropy model has the characteristics of training time period and small classification complexity, so we use The maximum entropy is used as the distance function yj between the training sample xi and the test sample. The specific formula is as follows:

$$d\_ME\left(x_i, y_j\right) = \sum_{k=1}^{l} \mathrm{y}_j^k \log \frac{y_j^k}{x_i^k} \tag{7}$$

Let 0log0=0. Use the formula instead of the Euclidean distance in the TR_KNN algorithm. Since the maximum entropy retains all the uncertainties and is closest to the natural state, the weight is

not determined. Therefore, the maximum entropy metric formula is not affected by subjective factors, which removes the shortcomings of the subjectivity of traditional algorithms.

3.2.3 IMPROVED KNN ALGORITHM

An improved KNN algorithm combining clustering algorithm and maximum entropy. The improved algorithm steps are as follows:

Step 1:

Use the CURE clustering algorithm to streamline the training sample set.

Step 2:

Construct a training sample set and a test sample set. The training sample set is represented as Ω, $\Omega=\{(x_i,c_i)|i=1,2,\mathrm{L}\ ,n\}$,where $x_i=(x_i^1,x_i^2,\mathrm{L}\ ,x_i^l)$ is a l-dimensional vector, that is, the feature dimension is 1, x_i^l representing the j-th feature component value of the i-th training sample. c_i represents the corresponding category of the i-th sample, and c_i belongs to the label set C, $C=\{1,2,\mathrm{L}\ ,t\}$, and t is the number of categories. The test sample set is expressed as Φ, $\Phi=\{y_j \mid j=1,2,\mathrm{L}\ ,m\}$,where y_j^l is the i-th eigen component value of the j-th test sample.

Step3:

Determine the initial value of K;

Step 4:

Apply the maximum entropy formula to calculate the distance between each test sample and the training sample, and sort the distance, and select the K samples closest to the test sample as the K neighbors of the test sample;

Step 5:

Find the main categories: set K neighbors $x_1',x_2',\mathrm{L}\ ,x_k'$, the corresponding category labels are $c_1',c_2',\mathrm{L}\ ,c_t'$, these category labels belong to the label set C. The queried test samples are classified according to the categories of K neighbors and applying the maximum probability. The probability used refers to the proportion of each category appearing in K neighbors, calculated by dividing the number of samples in each of the K neighbors by K. Then the category with the highest probability is recorded as the main category.Let $S=\{s_1,s_2,s_3,\mathrm{L}\ ,s_t\}$ be a collection of sample sizes for each of the K neighbors. Then $\tau^*=\arg\max_{r\in C}(s_\tau/k)$, the sample to be tested y_i is classified into τ^* classes.

Step6:

Evaluation. If the classification effect is not good, return to Step3 and continue the steps from Step3 to Step6, otherwise the algorithm ends.

The maximum entropy is used as its distance metric function, and no introduction of weight coefficients is involved, and the principle of maximum entropy does not introduce subjective estimates or assumptions. Therefore, it is completely objective and the classification result is more accurate.

4 RESULTS AND CONCLUSIONS

4.1 EXPERIMENTAL SYSTEM IMPLEMENTATION

This research developed a smart medical information system to provide a disease prediction method for doctors based on specific conditions. According to the patient's past history, current medical history, chief complaint, physical examination, advanced training data model, according to the CME_KNN algorithm, the patient's diagnosis is inferred, and the doctor can provide a reference for the doctor, and the doctor can also judge the result of the estimated patient diagnosis. Feedback information, the existing data model is gradually modified until it is more complete. Improve the efficiency of doctors and contribute to the automation of medical care. The operating environment of this system is: processor Inter Core i5-4460 3.20GHz, memory 8GB, hard disk 931.41GB, operating system is Windows 10 x64, programming language is C#.

4.2 EXPERIMENTAL RESULTS

4.2.1 RELEVANT DEFINITIONS

In order to evaluate the performance of the CME_KNN algorithm and the TR_KNN algorithm, four standard UCI (http://archive.ics.uci.edu/ml/datasets/Adult) data sets are selected in this paper. Table 1 gives the four data sets. specific information. The Abalone data set is incomplete. For ease of use, only the first 100 sets of data for each class are selected from the three categories of the standard Abalone data set. The performance of the two algorithms is analyzed from multiple dimensions by comparing the two algorithms in various cases.

Table 1. UCI data sets used in the experiment

Data set	Attribute type	Number of samples	Number of categories	Number of features
Iris	Real	150	3	4
Wine	Integer,Real	178	3	13
Abalone	Real	300	3	8
Balance	Categorical	910	5	4

The classification effect evaluation indicators commonly used are the recall rate (r), precision ratio (p), Fi test value, macro average and micro-average, etc. The macro-average index is the evaluation of the average situation of the class, and the micro-average index is the sample. The evaluation of the average situation, the classification performance evaluation indicators used in this paper have macro average recall rate ($\mathrm{ma}cro\text{-}\bar{r}$), macro average precision rate ($\mathrm{ma}cro\text{-}\bar{p}$) and macro F1 measurement value ($\mathrm{ma}cro\text{-}\mathrm{F}_1\ measure$), the calculation formula of each evaluation index is as follows:

$$\text{ma}cro\text{-}\bar{r} = \frac{\sum_{k=1}^{n} r_k}{n} \tag{8}$$

$$\text{ma}cro\text{-}\bar{p} = \frac{\sum_{k=1}^{n} p_k}{n} \tag{9}$$

$$\text{ma}cro\text{-}\mathrm{F_1} = \frac{\sum_{k=1}^{n} \mathrm{F_{1k}}}{n} \tag{10}$$

In Equation 8, r_k represents the recall rate, $r_k = a_k / b_k$, where a_k represents the number of correctly tested in the Kth test sample, b_k represents the number of Kth samples; t is the number of categories of the test sample In Equation 9, p_k is the precision, $p_k = a_k / d_k$, where d_k represents the number of test samples tested as Class K samples. In Formula 10, $\mathrm{F_{1K}}$ is the F1 measurement, $F_{1K} = 2r_k p_k / (r_k + p_k)$, which combines the recall rate $\mathrm{r_k}$ and the precision $\mathrm{p_k}$ into one measurement. The macro F1 measurement is the average of all individual category F1 values.

In addition, classification accuracy is often used as an evaluation index for classification performance. Classification accuracy refers to the proportion of samples correctly classified in the entire test sample set. The formula is

$$ACC = \frac{\text{Sort the correct number of samples}}{\text{Total number of all samples}} \tag{11}$$

4.2.2 EXPERIMENTAL RESULTS BASED ON DATA SETS

In the experiment comparing CME_KNN algorithm with TR_KNN algorithm, we set the percentage of training samples in the whole data set to 2/3, k value is 10, the experimental results are shown in the table.

Table 2. Comparison of classification results of CME_KNN and TR_KNN based on four data sets

Data set	**r_{TR}/%**	**r_{ME}/%**	**p_{TR}/%**	**p_{ME}/%**	**$F1_{TR}$**	**$F1_{ME}$**
Iris	95.83	97.92	97.92	98.04	96.86	97.98
Wine	96.66	96.80	94.71	97.22	95.67	97.00
Abalone	86.87	89.74	78.04	78.89	82.22	83.97
Balance	63.15	75.80	69.30	77.50	66.08	76.64

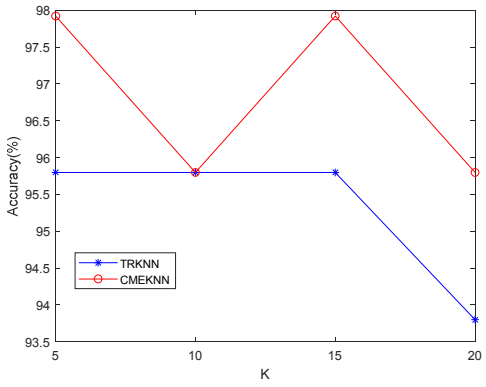
In the table, compared with the TR_KNN algorithm, the CME_KNN algorithm has improved in all the evaluation indexes of all data sets. On the wine dataset, the macro average recall rate is the smallest from TR_KNN to CME_KNN, but the value still reaches 0.14%. However, all the indicator values of the Balance data set are significantly improved, and the classification effect is

obvious. On the Balance data set, ma*cro*-$\bar{r}$, ma*cro*-$\bar{p}$, and ma*cro*-F_1 *measure* are increased by 12%, 8%, and 0.1, respectively. The average growth rate of ma*cro*-$\bar{r}$, ma*cro*-$\bar{p}$, and ma*cro*-F_1 *measure* on these four data sets was 4.4375%, 2.92%, and 3.69%. Therefore, the CME_KNN algorithm has better classification performance than the TR_KNN algorithm.

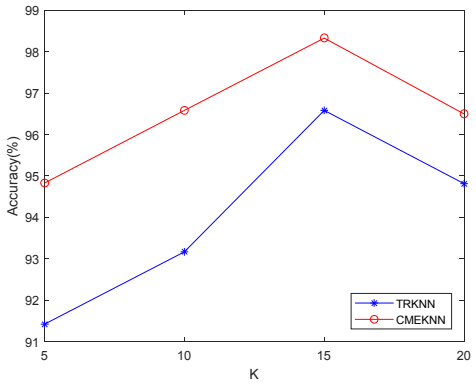
4.2.3 Influence Of Different K Values On Classification Accuracy

k is one of the key parameters affecting the classification accuracy of KNN. In general, if the value of K is very small, the less classification information obtained from the training sample set, the classification accuracy of the KNN algorithm is also relatively low. Within a certain range, the classification accuracy of the KNN algorithm gradually increases as the K value increases. However, when K exceeds a certain atmosphere, the classification accuracy of the KNN algorithm begins to decrease, because too many neighbor samples are selected at the time of classification, and a large amount of noise is generated for the classification information obtained from the training sample set.

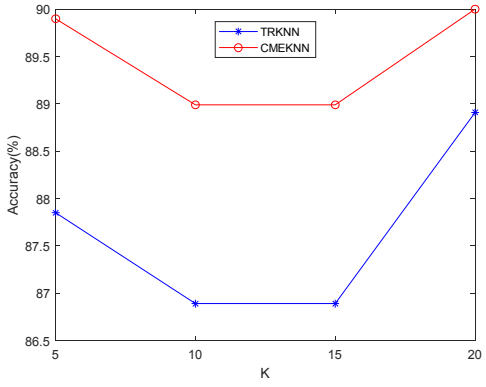
In order to illustrate the influence of K value on classification accuracy, four data sets were selected to verify the improved algorithm and the traditional algorithm. The experiment gave K, 5, 10, 15, and 20 values respectively, and set the training samples throughout. The percentage of the data set is 2/3, and the experimental results are shown in the figure.



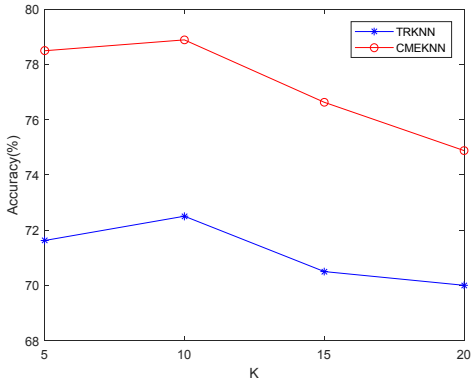
a) Iris



b) Wine



c) Abalone



d) Balance

Figure 2. The accuracy of CME_KNN and TR_KNN in different data sets varies with k value
In the figure, on the Iris dataset, the accuracy of TR_KNN and CME_KNN reaches the maximum at k=5 and 15. On the wine dataset, the accuracy of the improvement before and after k=15 reaches the maximum. For the Abalone dataset, When k=20, the classification accuracy is the highest. On the Balance dataset, the classification accuracy reaches a maximum when k=10. In these four images, the trend of the curves varies, and the k value obtained when the classification effect is the best is different. Since the value of the experiment k is too small, and k is worth the jitter, the graph a , b, d can partially show the trend of classification accuracy with k worth, but the c map can not reflect the influence of k value on classification accuracy. Therefore, it is concluded that the value of k has no definite method, and it can only be adjusted continuously until the classification accuracy is optimal, and the optimal k value for the different data sets is not necessarily the same, although There is no definite method for determining the k value, but the classification accuracy of the same k-value CME_KNN algorithm is significantly better than the TR_KNN algorithm.

4.2.4 Effect Of The Percentage Of Training Samples In The Sample Set On Classification Accuracy

The proportion of training samples in the entire sample set during the classification process also affects the classification results. In order to verify the influence of the proportion of training samples on the classification results, four sample sets were selected for the verification experiments of CME_KNN and TR_KNN algorithms. The specific proportions of the selected training samples were 1/3, 1/2, 2/3, respectively. 5 Experimental results are shown in the table.

Table 3. Effect of the proportion of training sample sets on classification accuracy (application CME_KNN and TR_KNN)

Classification accuracy	Ptrs	Iris	Wine	Abalone	Balance
ACCTR/%	1/3	92.93	89.83	85.33	68.26
	1/2	93.33	**88.64**	86.36	71.81
	2/3	95.83	91.38	87.88	71.52
ACCME/%	1/3	94.95	91.53	88.89	74.38
	1/2	95.67	94.32	88.89	80.18.
	2/3	97.92	94.83	89.90	78.48

As the proportion of training samples in the whole data set increases, the classification accuracy of the CME_KNN algorithm and the TR_KNN algorithm gradually increases. The bold data in the table is an outlier, that is, it does not conform to the increasing trend, but overall, for the same data set. In the case of the same algorithm, when the proportion of training samples is 1/3, 1/2, and 2/3, the classification accuracy of the knn algorithm increases sequentially. The classification accuracy of the CME_KNN algorithm is higher than that of the TR_KNN algorithm.

5. Conclusions

The calculation amount of KNN algorithm is exponentially increased with the increase of sample set data. From the perspective of reducing the amount of data, it is proposed to simplify the training sample set by CURE clustering algorithm. Most improved algorithms based on KNN use the method of adding attribute weights to calculate the distance between two samples. This

method has certain subjectivity and affects the accuracy of classification.An improved KNN algorithm based on maximum entropy is proposed.

The experimental results show that the improved algorithm proposed in this paper is more effective than the traditional KNN algorithm in terms of recall rate and precision. However, the distance metric function has a logarithmic term and therefore only applies if the feature is positive. The improved KNN algorithm proposed in this paper still has some shortcomings, which requires further research and improvement.

ACKNOWLEDGEMENTS

Participating in the information construction project of Yongcheng Central Hospital made me grow up a lot and master some necessary skills. Thank the laboratory for providing such a good environment. Thank the teachers and students for helping me. Thank you for helping me to grow up.

REFERENCES

[1] Ian H. Witten, Eibe Frank, & Mark A. Hall.(2005).Data Mining: Practical Machine Learning Tools and Techniques (Third Edition).

[2] Burges, C. J. C. . (1998). A tutorial on support vector machines for pattern recognition. Data Mining & Knowledge Discovery, 2(2), 121-167.

[3] Rjeily, C. B., Badr, G., Hassani, A. H. E., & Andres, E. (2019). Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field. Machine Learning Paradigms.

[4] Stiglic, G. , Brzan, P. P. , Fijacko, N. , Wang, F. , Delibasic, B. , & Kalousis, A. , et al. (2015). Comprehensible predictive modeling using regularized logistic regression and comorbidity based features. PLOS ONE, 10(12), e0144439.

[5] Nguyen, P. , Tran, T. , Wickramasinghe, N. , & Venkatesh, S. . (2016). Deepr: a convolutional net for medical records.

[6] Choi, E. , Bahadori, M. T. , Kulas, J. A. , Schuetz, A. , Stewart, W. F. , & Sun, J. . (2016). Retain: an interpretable predictive model for healthcare using reverse time attention mechanism.

[7] Hoogendoorn, M. , El Hassouni, A. , Mok, K. , Ghassemi, M. , & Szolovits, P. . (2016). Prediction using patient comparison vs. modeling: a case study for mortality prediction. Conf Proc IEEE Eng Med Biol Soc, 2016, 2464-2467.

[8] Sharafoddini, A. , Dubin, J. A. , & Lee, J. . (2017). Patient similarity in prediction models based on health data: a scoping review. Jmir Med Inform, 5(1), e7.

[9] Zhang, P. , Wang, F. , Hu, J. , & Sorrentino, R. . (2014). Towards personalized medicine: leveraging patient similarity and drug similarity analytics. Amia Jt Summits Transl Sci Proc, 2014, 132-136.

[10] Ng, K. , Sun, J. , Hu, J. , & Wang, F. . (2015). Personalized predictive modeling and risk factor identification using patient similarity. Amia Jt Summits Transl Sci Proc, 2015, 132-136.

[11] Sherry-Ann, B. . (2016). Patient similarity: emerging concepts in systems and precision medicine. Frontiers in Physiology, 7.

[12] Jiang, L. , Cai, Z. , Wang, D. , & Zhang, H. . (2014). Bayesian citation-knn with distance weighting. International Journal of Machine Learning & Cybernetics, 5(2), 193-199.

[13] Islam, M. M., Iqbal, H., Haque, M. R., & Hasan, M. K. (2018). Prediction of breast cancer using support vector machine and K-Nearest neighbors. IEEE Region 10 Humanitarian Technology Conference.

[14] Maillo, J. , Ramírez, Sergio, Triguero, I. , & Herrera, F. . (2016). Knn-is: an iterative spark-based design of the k-nearest neighbors classifier for big data. Knowledge-Based Systems, S0950705116301757.

[15] Han, J. . (2005). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc.

[16] Zhang, X. , Huang, X. , & Wang, F. . (2017). The construction of undergraduate data mining course in the big data age. International Conference on Computer Science & Education. IEEE.

[17] Holzinger, A., & Jurisica, I. (2014). Knowledge discovery and data mining in biomedical informatics: the future is in integrative, interactive machine learning solutions.

[18] Qu Fang,& Guo Hua.(2017). "Internet + big data" pension path to achieve. Science & Technology Review, , 35(16): 84-90.

[19] Pan, T. L. , Sumalee, A. , Zhong, R. X. , & Indra-Payoong, N. . (2013). Short-term traffic state prediction based on temporal–spatial correlation. IEEE Transactions on Intelligent Transportation Systems, 14(3), 1242-1254.

[20] Wright, J. , Yang, A. Y. , Ganesh, A. , Sastry, S. S. , & Ma, Y. . (2009). Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2), 210-227.

[21] Sha Yaqing, Sun Hongwei, & Gu Ming. (2006). Electronic tax certification system based on smart card and fingerprint identification.Computer Engineering, 32.(14):133-135.

[22] Zhou Lei, Wu Jianjun, & Zhang Jie. (2015). Research progress of remote sensing based drought monitoring methods. Geography Science, 35(5): 630-636.

[23] Xie Wei, Liu Bin,&Liu Xin. . (2017). Petroleum seismic exploration system and software platform in big data era. Science & Technology Review, 35(29): 172-174.

[24] Noora Abdulrahman, & Wala Abedalkhader. (2017). KNN Classifier and Naive Bayse Classifier for Crime Prediction in San Francisco Contextdoi. International Journal of Database Management Systems,9(4) .

[25] Juan, L. . (2012). TKNN: An Improved KNN Algorithm Based on Tree Structure. Seventh International Conference on Computational Intelligence & Security. IEEE.

[26] Patil, S., & Kulkarni, S. (2018). Mining social media data for understanding students' learning experiences using memetic algorithm. Materials Today Proceedings, 5(1), 693-699.

[27] Yan, Z., Yun, Q., & Li, C. (2018). Improved KNN text classification algorithm with MapReduce implementation. International Conference on Systems & Informatics.

[28] He, Q. , Li, N. , Luo, W. J. , & Shi, Z. Z. . (2014). A survey of machine learning algorithms for big data. Moshi Shibie yu Rengong Zhineng/Pattern Recognition and Artificial Intelligence, 27(4), 327-336.

[29] Metz, T. (2018). Medicine without cure?: a cluster analysis of the nature of medicine. The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine, 43(3), 306-312.

[30] Stonier, T. (2018). Towards a general theory of information ii: information and entropy. Electronic Notes in Discrete Mathematics, 21(9), 181-184.

[31] Chen, L., Gao, S., Zhang, H., Sun, Y., Ma, Z., & Vedal, S., et al. (2018). Spatiotemporal modeling of pm2.5 concentrations at the national scale combining land use regression and bayesian maximum entropy in china. Environment International, 116, 300-307

AUTHORS

Yanming Chen is a student at Henan University of Science and Technology, master's degree.

Weiming Mengis a student at Henan University of Science and Technology, master's degree.

Fenghua Zhang is a student at Henan University of Science and Technology, master's degree.