Data Mining of Medical Record Front Page Based on Improved Apriori Algorithm

Huizi Sun, Xiaoming Dong* and Yu Liu

The Second Affiliated Hospital of Qiqihar Medical University, Qiqihaer, Heilongjiang 161006, China

*corresponding author

Keywords: Apriori Algorithm; Medical Record Home Page; Data Mining; Causal Analysis

Abstract: At present, data mining technology has been very good development, and has been widely used in many industries and fields, which is very convenient for our life. Data mining technology is also widely used in medicine. On this basis, we improve the traditional Apriori algorithm and construct a data mining method for the front page of medical records, and verify the effectiveness of this method through practical application. On the other hand, we compare the proposed method with the traditional Apriori algorithm, and analyze the efficiency and scalability of the proposed method, so as to highlight the advantages of this method. The results show that when the data volume is 458,916 and 1374, the running time of the data mining method based on the improved Apriori algorithm is 10s, 28s and 48s respectively, while that of the traditional Apriori algorithm is 20s, 42s and 86s respectively. Therefore, the expansibility of this method is better. This study provides an effective reference value for medical record management.

1. Introduction

Data mining is a kind of computer-aided decision analysis, its main task is to predict and describe [1-2]. In brief, the process of data mining is to first obtain data and standardize the acquired data. Then, through various ways and methods, we can find the relationship between data and the law of the influence of data or data groups on the results. Finally, the law is expressed in a way that users can understand [3-4]. As we all know, at present, data mining technology has been widely used. Although the application of data mining technology in the medical field is not very extensive and not very mature, the prospect is optimistic and the development potential is great [5-6].

With the large-scale application of electronic medical records in the medical industry, as well as the trend of digital management of medical equipment and instruments, the information capacity of hospital database is constantly expanding, which brings convenience to people's medical treatment, but also because of the huge amount of data information, it increases the difficulty of hospital information processing. Therefore, it is necessary to use medical data mining technology for processing [7-8]. Before data extraction, it is necessary to screen and filter the information to ensure the consistency and accuracy of data, and then convert it into a form suitable for extraction [9-10]. Therefore, it is of great significance to seek an efficient medical data mining method.

The related theory is the basis of this article. Therefore, in this paper, we first summarize the data mining technology and Apriori algorithm. On the basis of these theories, we improve the Apriori algorithm, and construct a data mining method for the front page of medical records based on the improved Apriori algorithm, and verify the effectiveness of the improved algorithm by practical application. In addition, we also analyze the performance of the proposed algorithm. The results show that compared with the traditional Apriori algorithm, the proposed method is not only efficient, but also has good scalability.

2. Data Mining and Apriori Algorithm

2.1 Data Mining

We first define the meaning of data mining. Data mining, as the name implies, is the process of extracting and mining useful information from huge data. Data mining has its fixed steps, usually the following steps:

(1) Identify data mining objects

Generally speaking, before data mining, we must first determine the object of data mining. Only by understanding the real core purpose can we make clear the goal of data mining.

(2) Data preparation

We know that in order to ensure the smooth and successful data extraction, data preparation is essential. Therefore, in the process of data mining, data preparation is also very important.

(3) Data mining

In addition, in the process of data mining, it is very important to understand the types and characteristics of data. Only when we understand the types and characteristics of data, can we choose the appropriate algorithm to extract data and realize data mining.

(4) Result analysis

Interpret and evaluate the extracted results and translate them into knowledge that users can finally understand.

(5) Application of knowledge

After the end of data mining, we combine the extracted knowledge into the actual operating system to realize the application of knowledge.

2.2 Apriori Algorithm

We know that in association rules, Apriori algorithm can be said to be a very classic algorithm, is a mining frequent itemsets. Generally speaking, according to the given minimum support and confidence, the algorithm can obtain the required frequent patterns from numerous data, so that the association between each data can be found according to the obtained frequent patterns.

In association rules, if we set X as the item set and t as the thing set, then the support degree is an indicator of the frequency of item set X in the event set T. the formula is defined as follows:

$$\sup port(X) = \frac{\left|\{t \in T; X \in t\}\right|}{|T|} \tag{1}$$

The confidence degree is an indicator of the frequency of rules appearing in the set of things, which is defined as:

$$confidence(X \Rightarrow Y) = \frac{\sup port(X \cup Y)}{\sup port(X)} \tag{2}$$

The degree of promotion is an indicator of the strength of the relationship between X and Y:

$$Lifi(X \Rightarrow Y) = \frac{P(X \cup Y)}{P(Y)} = \frac{confidence(X \Rightarrow Y)}{P(Y)} \tag{3}$$

There are fixed steps in the implementation of Apriori algorithm, which is usually divided into two steps: first, constructing frequent items, and then combining frequent itemsets into rules.

3. Practical Application and Algorithm Performance Analysis

3.1 Practical Application

(1) Data sources

This paper selects the data of the inpatient department of our hospital, extracts the cases with hypertension in the main diagnosis. The total number of cases selected this time is 4590. As we all know, there are many kinds of diseases, and so are the names of diseases. In this paper, we only

extract a part of the types of diseases for research, taking this as an example to illustrate.

(2) Experimental environment and purpose

In the process of experiment, it is inevitable to have some random codes and wrong records caused by improper operation of data input and export. In order to make up for the consequences of these improper operations and eliminate noise data, we preprocess the data, and the data preprocessing is carried out in IBM SPSSs statistics. After data preprocessing, we put the data into the data analysis software R studio for modeling, so as to analyze whether there is a causal relationship between hypertension and collaborative diseases.

3.2 Algorithm Performance Analysis

Is the improved algorithm superior? We need to verify this. In order to compare, we choose the traditional Apriori algorithm as a reference, and compare and analyze the efficiency and scalability of the two algorithms.

4. Analysis of Data Mining Method of Medical Record Front Page Based on Improved Apriori Algorithm

4.1 Improvement of Priori Algorithm

Apriori algorithm can be said to be the most classic algorithm for generating frequent itemsets. This algorithm is not only simple, but also very easy to master and apply. However, after a long and large number of studies, Apriori algorithm has its own shortcomings

(1) This algorithm sometimes consumes a lot of time. For example, when dealing with candidate sets with large scale, this is a big defect.

(2) Before matching the candidate set pattern, it is necessary to scan the transaction database many times.

(3) Generally speaking, the adaptability of this algorithm is not strong, and it is very suitable for processing data of single dimension, single layer and boolean type, but it is not applicable if it is processing data with large number and high dimension. However, in real life, it is mostly such data type. Therefore, it is necessary to improve the algorithm.

After the above description, we have a certain understanding of the defects of Apriori algorithm, then, how can we improve the algorithm to make up for the defects of the algorithm? Generally speaking, in view of the defects of the algorithm, the improved method has the following entry points:

(1) First of all, we consider the efficiency of the improved algorithm. Then, we all know that the algorithm supports the calculation of each candidate set by over scanning the transaction database. If the scanning times can be reduced, we can shorten the scanning time and improve the efficiency.

(2) Secondly, we know that this algorithm needs a lot of storage space in the process of operation. This is because the algorithm will get a large number of candidate sets in each operation, and the storage space of these candidate sets is very large. Therefore, we can achieve the purpose of improvement by reducing the storage space.

(3) On the other hand, this algorithm uses the subset of frequent itemsets to prune the candidate sets. If we can reduce the number of candidate sets, we can also improve the efficiency of the algorithm, and then improve the algorithm.

(4) Finally, after each candidate dataset is created, we need to scan the candidate datasets to determine whether they are frequent datasets. However, in the process of scanning the database, there are often some irrelevant data or transactions. If we can filter out these irrelevant data in advance, we can greatly improve the effectiveness of the algorithm.

There are several methods to improve Apriori algorithm

(1) Sampling based method

This method has a major drawback, that is, it is not accurate enough. In this method, we propose partial sampling in the transaction database, and then analyze the sampling, so as to obtain certain rules, because these rules may be applicable to the whole transaction database, so it will lead to

inaccurate results.

(2) Partition based method

In this approach, we usually use specific data structures to maintain a list of identified transactions for each common dataset. Therefore, when the candidate $k-$itemsets are used to generate frequent itemsets, we do not need to scan the database to calculate the support. In this way, the whole process will be simpler.

(3) How to reduce the number of transactions

The purpose of this method is to reduce the number of transactions. However, in practical application, some transactions of frequent itemsets are redundant. If we can delete these redundant transactions, we don't need to scan so many transactions, which can effectively improve the efficiency of the algorithm.

This paper improves the Apriori algorithm by referring to the above methods. The basic idea of Apriori algorithm improvement in this paper is to add a database D_K to each generation of frequent itemsets $L_{K-1}(k=1,2,\ldots,k-1)$. The database D_K is used to store frequent itemsets and their transaction sets. The transaction set of each frequent itemset is marked as E_i, $E_i=(t_1,t_2,t_3,\ldots t_m,\ldots t_p)$.Therefore, when a $k-$ candidate set C_k is generated by L_{K-1} self connection, it is no longer necessary to scan the original database d, but directly scan the transaction E_i of C_k subset in D_K of frequent itemset database, and then calculate the intersection of transaction E_i of C_k. The number of transactions in the intersection set is the support number of the candidate itemset, Finally, delete the candidate itemsets which are less than the support, and the rest is $k-$ frequent itemsets.

4.2 Algorithm Application Analysis

We have applied this method in practice, and the results are shown in Table 1.

Table 1. Analysis of algorithm application results

Name of disease	Number	Is there a causal relationship	Partial correlation coefficient
Type 2 diabetes	3256	0	3.50869
Atherosclerosis	2562	1	8.25686
Cerebral infarction	2132	1	9.23756
Atherosclerosis of muscle	1256	1	6.78394
Hyperlipidemia	1105	0	1.12658
Arteriosclerosis	982	0	0.86125
Hyperuricemia	973	0	1.01569
Chronic bronchitis	856	0	2.35486
Sequelae of cerebral infarction	845	0	1.58963
Cerebral hemorrhage	832	0	2.13568
Benign prostatic hyperplasia	812	0	2.01563
Unstable angina pectoris	786	1	19.3896
Acute myocardial infarction	568	1	13.8965

In Table 1, 0 indicates the pre causality, and 1 indicates that there is a causal relationship between the predicted variable and the target variable. The larger the partial correlation coefficient, the stronger the partial correlation between the two diseases. According to Table 1, hypertension, cerebral infarction, cervical porridge, pain and acute infarction have causal relationship. The results show that the research of data mining in this paper is consistent with the medical practice, and the effectiveness of the algorithm is verified.

4.3 Algorithm Performance Analysis

(1) Algorithm efficiency

In the research, we test the efficiency of the algorithm. In the process of testing, we select a

dataset to test, and the selected dataset contains 458 transactions. In order to highlight the advantages of this method, we choose the traditional Apriori algorithm as a reference, and mine association rules with this method. In the test, we set min_ Sup values are 10%, 5%, 4%, 3%, 2%, 1%, and then the execution time of the two algorithms is recorded for each value. The results are shown in Figure 1.

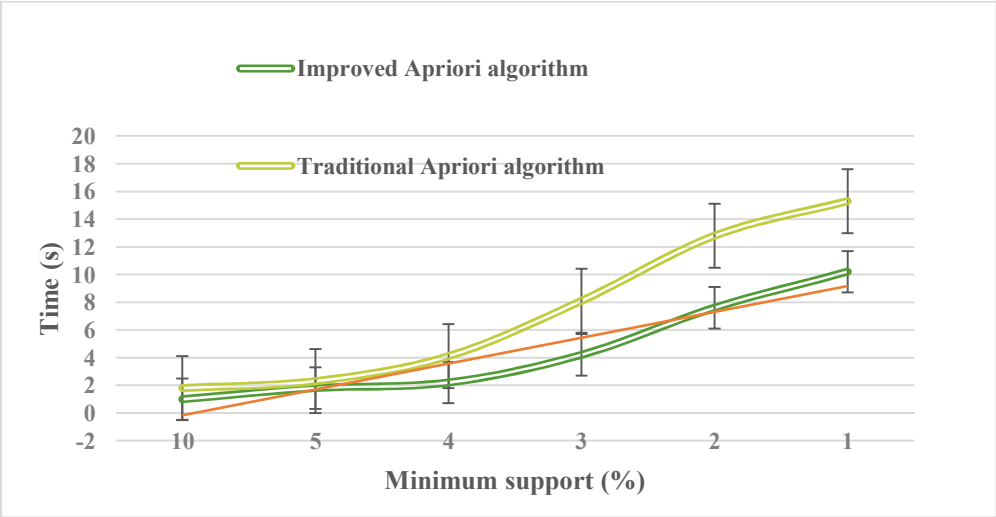

Figure 1. Efficiency comparison of the two algorithms

It can be seen from Figure 1 that when min_sup value is 10%, the execution time of data mining method based on improved Apriori algorithm is 1s, and that of traditional Apriori algorithm is 1.8s. When min_sup value is 5%, the execution time of data mining method based on improved Apriori algorithm is 1.8s, and that of traditional Apriori algorithm is 2.3s. When min_sup value is 4%, the execution time of data mining method based on improved Apriori algorithm is 2.2s, and that of traditional Apriori algorithm is 4.1s. When min_sup value is 3%, the execution time of data mining method based on improved Apriori algorithm is 4.2s, while that of traditional Apriori algorithm is 8.1s. When min_sup value is 2%, the execution time of data mining method based on improved Apriori algorithm is 7.6 s, while that of traditional Apriori algorithm is 12.8 s. When min_sup is 1%, the execution time of data mining method based on improved Apriori algorithm is 10.2s, and that of traditional Apriori algorithm is 15.3s. Therefore, the efficiency of the data mining method based on the improved Apriori algorithm is higher than that of the traditional Apriori algorithm.

(2) Algorithm scalability

In the research, we test the scalability of the algorithm. In this test, min_ the value of sup is determined to be 1%. In order to highlight the advantages of the method in this paper, we choose the traditional Apriori algorithm as a reference to mine association rules with the method in this paper, and record the execution time of each data set, and the results are shown in Figure 2.

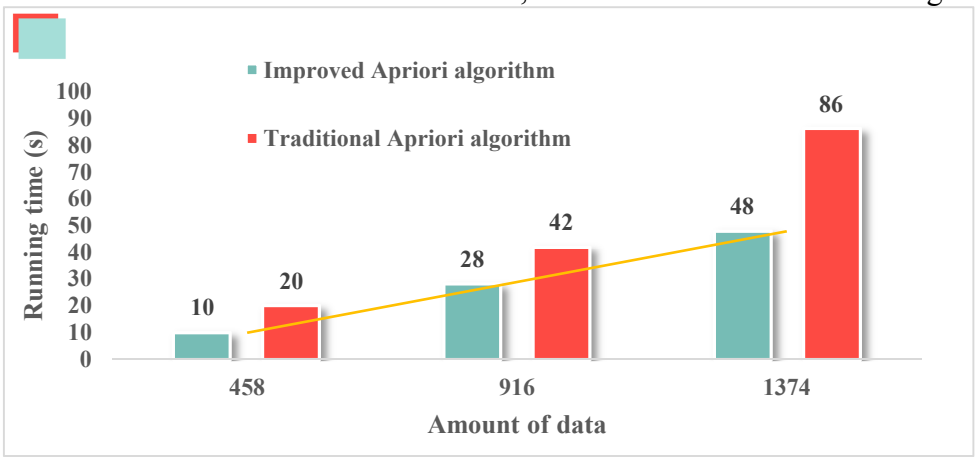

Figure 2. Scalability analysis of the two algorithms

As can be seen from Figure 2, when the amount of data is 458,916 and 1374, the running time of the data mining method based on improved Apriori algorithm is 10s, 28s and 48s respectively, while that of the traditional Apriori algorithm is 20s, 42s and 86s respectively. Therefore, the expansibility of this method is more superior.

5. Conclusions

Nowadays, the development of data mining technology is very rapid, and has been well applied in many industries and fields. In this way, it brings a lot of convenience to our life. At present, data mining technology has been widely used in medicine, which brings great convenience for medical diagnosis and treatment. In order to better carry out the research, we first reviewed some theoretical knowledge, and through the improvement of Apriori algorithm, constructed the improved medical record front page data mining method based on Apriori algorithm. In addition, we also apply the method in this paper, and test the performance of the algorithm based on the traditional Apriori algorithm. The results show that the method in this paper has certain feasibility, and can provide certain reference value for the application of data mining technology in medicine.

Acknowledgement

This work was supported by Research on data mining and statistical analysis of the first page of inpatient medical records to serve hospital management (No.: KJCX5423)

References

[1] Atta-ur-Rahman, Dash S. Data Mining for Student's Trends Analysis Using Apriori Algorithm. International Journal of Control Theory and Applications, 2017, 10(18):107-115.

[2] Guo Y, Wang M, Li X. Application of an improved Apriori algorithm in a mobile e-commerce recommendation system. Industrial Management & Data Systems, 2017, 117(2):287-303.

[3] Xiyu L, Yuzhen Z, Minghe S. An Improved Apriori Algorithm Based on an Evolution-Communication Tissue-Like P System with Promoters and Inhibitors. Discrete Dynamics in Nature and Society, 2017, (2017-02-19), 2017, 2017:1-11.

[4] Jia K, Li H, Yuan Y. Application of Data Mining in Mobile Health System Based on Apriori Algorithm. Journal of Bjing University of Technology, 2017, 43(3):394-401.

[5] Buczak A, Guven E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials, 2017, 18(2):1153-1176.

[6] Chaurasia V, Pal S. A Novel Approach for Breast Cancer Detection using Data Mining Techniques. Social ence Electronic Publishing, 2017, 3297(1):2320-9801.

[7] Helma C, Cramer T, Kramer S, et al. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. J Chem Inf Comput, 2018, 35(4):1402-1411.

[8] Chen C T, Chang K Y. A Study on the Rare Factors Exploration of Learning Effectiveness by Using Fuzzy Data Mining. Eurasia Journal of Mathematics ence & Technology Education, 2017, 13(6):2235-2253.

[9] Shin S, Hwang I. Data-Mining-Based Computer Vision Analytics for Automated Helicopter Flight State Inference. Journal of Aerospace Information Systems, 2017, 14(12):652-662.

[10]Fernandes E , Holanda M , Victorino M , et al. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 2019, 94(JAN.):335-343.