ATLANTIS PRESS

Advances in Engineering Research, volume 149

2nd International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018)

Research and Application of Data Mining in Chronic Diseases

Yuliang Shi [a], Jun Tao [b]

School of Beijing University of Technology, Beijing 100124, China.

[a]shiylidc@126.com, [b]1320177931@qq.com

Keywords: Apriori algorithm, data mining, association rules.

Abstract. In recent years, with the acceleration of people's pace of life, the number of chronic diseases in China is increasing. The attention and investment of the country to the medical industry is increasing year by year. At the same time, with the maturity and perfection of data mining technology, many countries have applied this technology to the research and mining of medical data. In this paper, the Apriori algorithm of data mining technology, and improve the data format of the Apriori algorithm is applied to the prediction of nephropathy, establish the association rules between chronic disease and a number of physical data by the algorithm, and the experimental results proved that Apriori algorithm is effective in the medical data mining.

1. Introduction

Chronic diseases have become an important public problem endangering people's health in the twenty-first Century. Chronic diseases often have a serious impact on the health and life of the patient, and will bring a serious burden to the family and society of the patient. However, with the rapid development of computer, network and information technology, the concept of digitalization and information has entered many fields in all walks of life and people's lives. Our country's hospitals have quietly entered the era of digitization and informatization. While deepening digitization and informatization, a great deal of medical data of patients with chronic diseases are produced, which contain a great deal of valuable information[1]. Data mining technology will be hidden in the massive medical data useful information mining has become the focus of research in data mining technology applications. There are many data mining methods, one of the most widely used method is to find the association rules in the data[2].

Based on the introduction of association rules and Apriori algorithm, Apriori algorithm is used to mine the association rules of chronic diseases. According to the situation of medical data services, the frequent item sets are extracted by using vertical data format to improve the association between Apriori algorithm and chronic disease data association rules mining in the operational efficiency, at the same time prove Apriori algorithm in the medical data mining effectiveness.

2. The Basic Theory of Association Rules

2.1 Overview of Association Rules

Association rules for the data pattern mining can be formally defined as: Let I = {i1, i2, i3, ..., in} is a collection of n different items[3]. For a specific database transaction set D, D each transaction T is a non-empty item set, and satisfy T, I. Each transaction has a unique identifier called TID. Let A be a set of items, satisfying that T contains A if and only if A, T. Association rules can be expressed as the form of the form A, B, where A, I, B, I, A, B, and A, B =, The meaning of association rules is that it is possible to derive additional items from some items that can be in the transaction.

2.2 Association Rules Related Concepts

Association rules have the following basic concepts,

Support s: represents the percentage of transactions in D containing A, B, which is the probability P (A, B). which is

$$\text{Support}(A \Rightarrow B) = P(A \cup B) \quad (1)$$

Confidence c: represents the percentage of D contains the A transaction also contains the B transaction, which is the conditional probability P (B | A). which is

$$\text{Confidence}(A \Rightarrow B) = P(B|A) \quad (2)$$

Since the confidence of rule A, B is easily derived from the support counts of A and A, B, we get the formula

$$\text{Confidence}(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)} \quad (3)$$

Where support_count (A, B) is the number of transactions that contain item sets A, B and support_count (A) is the number of transactions that contain item set A.

3. Apriori Algorithm

3.1 Algorithm Overview

Apriori algorithm was proposed by Agrawal and R.Srikant in 1994. It is an original algorithm for mining frequent item sets of Boolean association rules. Its algorithm idea [4] is an iterative method of layer-by-layer search, using known k-dimensional frequent item sets to generate k+1-dimensional frequent item sets.

3.2 Algorithm Process

Apriori algorithm[5] first by scanning the database, the cumulative count of each item, will meet the minimum support count items to count, find a set of frequent 1 set, denoted as L1. Then use L1 to find the set L2 of frequent 2 sets, find L2 using L2, and so on until you can not find frequent K sets. Finding each Lk requires a full scan of the database.

4. Application of Apriori Algorithm

4.1 Chronic Disease Data Pretreatment

The data selected in this paper is the medical data of patients with nephropathy. To apply Apriori algorithm to this data mining, we must first make some preprocessing of the collected data[6]. Nephropathy patients state needs to be based on the obtained blood pressure, urinary specific gravity and other discrimination, so we need to dig out the conditions from a number of conditions and its relationship to the disease, and blood pressure, urinary weight and other physiological parameters of patients is not able to Direct data mining, and therefore need to be based on the characteristics of medical data, their generalization of the form[7].

The specific approach is: the gender is divided into two categories, the men with M1 said, the woman with M2 said; the age is divided into three categories, 70 <age with A3 said, $50 \leq age \leq 70$ with A2 said 30 , age < 50 with A1 that; the proportion of urine divided into three categories, 1.025 <urine specific gravity with S3 said, 1.010 , urine specific gravity , 1.025 with S2 said, urine specific gravity <1.010 with S1 said; blood pressure is divided into three categories, 100 < blood pressure with P3 said, 80 < blood pressure , 100 with P2 said, 60 , blood pressure , 80 with P1 said. Use these newly split categories as values for the data mining dimension.

The following chart is obtained by transforming the collected patient data. Table 1 shows the original physiological data table. Table 2 shows the transaction data table for data mining (status N indicates no kidney disease and Y indicates kidney disease).

Table 1 original physiological parameters table

Id	Sex	Age	Sg	Bp
1	Man	45	1.02	80
2	Man	37	1.005	90
3	Women	33	1.01	70
4	Man	50	1.015	110

Table 2 Transaction Data Sheet

Sex	Age	Sg	Bp	State
M1	A1	S2	P1	N
M1	A1	S1	P2	Y
M2	A1	S2	P1	N
M1	A2	S2	P3	Y

4.2 Applying Apriori Algorithm to Mining Strong Association Rules

After preprocessing the data, we get 401 transaction data. Due to the large number of transactions, if we conduct mining according to the traditional algorithm, we will generate a large number of frequent item sets[8], resulting in the algorithm running for a long time. Therefore, We use the vertical data format to mine frequent item sets, take each item sets separately and count the TIDs that contain

the item sets, so that we can directly count the number of statistical TIDs as its support count, from which we can get The transaction database vertical data format[9] table, as shown in Table 3.

Table 3 transaction database vertical data format table

Item set	TID-set support count	Item set	TID-set support count
M1	233	S2	103
M2	168	P1	67
A1	75	P2	148
A2	220	P3	186
A3	106	N	42
S1	298	Y	359

Based on the above data, Apriori algorithm can iteratively find 5 items sets, and find the corresponding Confidence, if greater than the set minimum confidence, is strong association rules[10]. In the iteration with Apriori algorithm, we set the minimum support to 100 according to the characteristics of the medical data. Finally, we get the table of vertical data format 5 items set in Table 4 below.

Table 4 vertical data format 5 sets table

Item set	TID-set support count
{M1,A2,S1,P3,Y}	127
{M1,A2,S1,P2,Y}	133
{M2,A2,S2,P3,Y}	109

From the table data, combined with the four sets of data iteration, take {M1, A2, S1, P3} , Y as an example and calculate the confidence level as:

Confidence({M1,A2,S1,P3}⇒Y) = 127/140 = 0.907

4.3 Simulation Algorithm and Analysis

Through experiments, The simulation results of the Apriori algorithm and the traditional Apriori algorithm in the transform data format are shown in Figure 2.

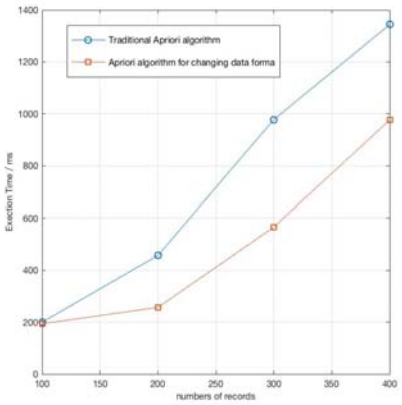

Figure 2 Algorithm run time and the relationship between the number of records

As can be seen from Figure 2, the Apriori algorithm changes the data format consumes much less time than the traditional Apriori algorithm, and as the number of records increases, changing the data format of the Apriori algorithm, advantage has become more and more obvious because of When seeking frequent item sets, the items that need to be counted are less, so the time for scanning the database is greatly reduced, which improves the running efficiency of the algorithm.

The improved Apriori algorithm was used to mine the nephrology data. By setting the minimum support[11] 100 and the minimum confidence[12] 0.9, the experimentally obtained strong association rules are as follows:

(1) 50-70 years old, male, urine specific gravity <1.010, blood pressure <100, Kidney disease detected - 90.7%;

(2) 50-70 years old, male, urine specific gravity <1.010, 80, blood pressure <100, Kidney disease detected - 93.8%;

(3) 50-70 years old, female, 1.010, urinary specific gravity, 1.025, blood pressure <100, kidney disease - 92.6%;

For example, the first association rule, which indicates that men aged 50-70 years, if the urinary specific gravity is less than 1.010, blood pressure less than 100, the risk of suffering from kidney disease is about 90.7%. Therefore, the medical staff can determine the condition of the medical staff according to the acquired association rules through the collected physiological parameters.

It can be seen from the above experiments that the Apriori algorithm which changes the data format has been applied to the excavation of nephropathy data, which provides the basis for doctors to make timely diagnosis.

5. Conclusion

In this paper, we introduce the association rules and Apriori algorithm in data mining, and then solve the problem of large number of frequent items in the process of getting frequent item sets in patients with nephropathy, and change the data format of traditional Apriori algorithm, reducing the number of transactions in the transaction database [13], thus reducing the number of frequent items generated during the iteration of the algorithm and verifying the significant improvement of the algorithm's time efficiency through experiments. At the same time, it is proved that Apriori algorithm can effectively reduce the number of chronic disease data Effectiveness of mining.

References

[1]. Murdoch T B, Detsky A S. The inevitable application of big data to health care [J]. JAMA, 2013, 309 (13): 1351

[2]. Agrawal R I, Nski T, Swam I A. Mining association rules between sets in items in large database [C] // Proc of the ACM SIGMOD international conference on management of data. Washington D C: ACM, 1993: 207-216

[3]. Imielienskin T, Swami A, Agrawal R. Mining association rules between set of items in large databases [J]. Acm Sigmod Record, 1993, 22 (2).

[4]. Wei Wang. Research and Improvement of Apriori Algorithm in Association Rules [D]. Qingdao: Ocean University of China, 2012.

[5]. Inokuchi A, Washio T, Motoda H. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data [J]. 2000, 1910 (1): 13-23.

[6]. Bayardo, Roberto J., Jr, R. Agrawal, and D. Gunopulos. "Constraint-Based Rule Mining in Large, Dense Databases." Data Mining & Knowledge Discovery 4.2-3 (2000): 217-240.

[7]. Ruixin Li, Shuping Zhang. Data preprocessing in the construction of data warehouse [J] .Application of Computers, 2002, 11 (5): 18-21.

[8]. Bingying Long, Xiaohui Chen. Research and Application of Improved Apriori Algorithm in Hospital Guardianship Center [J]. Computer Technology and Development, 2013,23 (8): 137-140.

[9]. Silverstein C, Brin S, Motwani R. Beyond Market Baskets: Generalizing Association Rules to Dependence Rules [J]. Data Mining & Knowledge Discovery, 1998, 2 (1): 39-68.

[10]. Chen Wu, Ronghua Yang. Selective Integration Algorithm Based on Frequent Closed Itemsets in Vertical Data Format [J] .Electronic Design, 2016,24 (19): 69-72.

[11]. Thober M, Pendergrass JA, McDonell C D. Improving coherency of runtime integrity measurement [C] // Proc of the 3rd ACM workshop on scalable trusted computing. Alexandria, USA: ACM, 2008: 51 -60.

[12]. Lin D I, Kedem Z M. Pincer-search: A new algorithm for discovering the maximum frequent set [C] // International Conference on Extending Database Technology: Advances in Database Technology. Springer-Verlag, 1998: 105-119.

[13] . Yuke Yang, Changguo Li. An improved scheme of fuzzy comprehensive evaluation of software quality based on the least confidence and evaluation [J]. Journal of Computer Applications, 2009, 29 (9): 2530-2533.