The Research on Analyzing Risk Factors of Type 2 Diabetes Mellitus Based on Improved Apriori Algorithm

Wang yuzhen[1,a], Jiang donghong[1] ,Wei Zhe[1,2], Ye Guangjian[2, b]

[1]Lanzhou General Hospital, Lanzhou Military Area Command, Gansu, Lanzhou 730050, China

[2] School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Gansu, Lanzhou 730050,China

[a]245891156@qq.com, [b]11445697@qq.com

Keywords: data mining; Apriori Algorithm; association rules; improved algorithm

Abstract. *Purpose:* We do it to improve the efficiency of analyzing risk factors of Type 2 Diabetes. *Method:* We use the patients' data from the information department of one tertiary referral hospital in Lanzhou which include course note of disease and their health record form January 2009 to March 2014.We find out one improved algorithm applies to analyze risk factors of Type 2 Diabetes based on original Apriori Algorithm and it's requirement. And we analyze the efficiency by programming both of the algorithms with C#. *Result:* We can analyze the chart of frequent item and support degree, time and number of records, time and support degree. *Conclusion:* This new improved Apriori Algorithm has a high efficiency in analyzing risk factors of Type 2 Diabetes.

I. Introduction.

Diabetes Mellitus is due to the secretion of insulin and the role of defects caused by chronic high blood sugar with carbohydrates, metabolic disabled of fat and protein chronic disease characterized. Type 2 Diabetes Mellitus, which is called non-insulin-dependent Diabetes Mellitus as well and due to insulin resistance with relatively lack of insulin secretion, holds 90% to 95% of all the patients with Diabetes Mellitus. Diabetes Mellitus has been one of the most common chronic non-communicable diseases with the prevalence showing a rising trend in the whole world recent years. It is predicted that the global total number of adults with Diabetes Mellitus will grow from 171 million in 2000 to 366 million in 2030, with a growth of 1.14 times[1]. So research on Diabetes Mellitus is very important.

We find defect of Apriori Algorithm in research on mining association rules of Type 2 Diabetes Mellitus risk factors. First, Apriori Algorithm has to used to scan the database once when generate a frequent item set each time. And second, when generating *k* candidate item sets from (*k*-1) frequent item sets, it will product many candidate item sets which is unnecessary later and have a long time in data mining of risk factors and a low work efficiency. We propose a modified Apriori Algorithm suitable for risk factors of Type 2 Diabetes Mellitus with a large data and attribute value of risk factors.

II. Structuring the Mining Rules

Apriori Algorithm

Apriori Algorithm is one kind of association rule algorithm with layer-by-layer searching and iterating. For example, we search 3-length frequent item set with 2-length frequent item set, and then iterate the result to search 4-length frequent item set. In one word, we use the *k* item set to search the *(k*+1*)* item set. In detail, we first scan the database to count the number of each item which is larger than the min-support degree and structure it as 1-length frequent item set called L_1. Then we search L_2 in L_1 in the same way which we can also find L_3, L_4 and *Ln* until all the frequent item sets have been found. It is obvious that we have to scan the database once to search for L_k each time, and the two steps can be showed as those [2]:

1. Find all the frequent item sets L.

2. Find the strong association rules from L.

The first step shows the key of the algorithm which decides the quality and efficiency, and most of the improvement of Apriori Algorithm is based on the first step. The main measure has two steps: the connection step and the prune step.

The connection step: we connect the L_{k-1} item set with the candidate item set C_k to find We assume l_1 and l_2 are the item sets of item set L_k , so we can define l_i [j] as the jth item, then we do the connection. The connection demands l_1 and l_2 from L_k can be connected, and if (l_1[1]= l_2[1])^ (l_1[1]= l_2[2])^ ……^(l_1[k-2]= l_2[k-2])^ (l_1[k-1]< l_2[k-1]), when l_1[1], l_1[2] ……l_1[k-1], l_1[k-2] are the connection result item sets.

The prune step: we scan the database to find support degree number of the candidate item set C_k [3], however the candidate C_k is probaly large. We can compress C_k with the theorem that any (k-1) item set from the non-frequent item set can never be the subset of the k item set from the the frequent item set. So if (k-1) item set from candidate k item set is not in L_{k-1}, the candidate item set must be non-frequent and can be deleted from C_k.

Apriori Algorithm has been widely used in data mining field when proposed by R.Agrawal et al in 1994. It has defects like the first the most classic association rule algorithm: we have to scan the database once when generating frequent item set each time and product many unnecessary candidate item sets.

Predecessors have made a lot improvement to the defects of Apriori Algorithm. The 4th literature proposes a theory based on zone which has the main mind of 2-times scan of database. The 5th literature proposes a high efficiency way to generate frequent item set based on Hash Algorithm. We can find in the experiment that the main work of finding the frequent item set is generating the k item set L_k, and the Hash Algorithm will improve the generation of the k frequent item set. The 6th literature proposes a theory based on sampling. It analyzes and combines the data of previous scan in detail, and gets a way to improve the algorithm: we get some rules which are possible to be true from the whole database base on its sampling and test the rules with the rest database.

One Improved Apriori Algorithm to Analyze Risk Factors of Type 2 Diabetes Mellitus

We collect more than 30 thousand course notes of disease and health records of patients with Type 2 Diabetes Mellitus from the information department of one tertiary referral hospital in Lanzhou, and mine the data to find risk factors of Type 2 Diabetes Mellitus. We choose 15 risk factors: gender, age, education level, body mass index (BMI), waist hip ratio (WHR), personality, trauma history, drinking, tea, smoking, sleep, exercise, income level, occupation, meals on time. We preprocess the original data in the way below: taking BMI as an example, BMI focuses on the interval region [23, 31], so we separate the region equally to (0, 23], (23, 25], (25, 27], (27, 29], (29, 31], (31, ∞], and define the 6 interval regions as B1, B2, B3, B4, B5, B6. Totally we transform the 15 risk factors into 44 variables. It is obvious that 44 variables from at least 30 thousand data will have a large operation and cost a lot PC RAM and long time. So we propose a modified algorithm to analyze the risk factors.

Theory Basis

Apriori Algorithm has qualities below:

1. Supersets of non-frequent item sets must be non-frequent.
2. All non-empty subsets of frequent item sets are frequent.
3. We assume a transaction set I includes k frequent item set L_k, if L_k can generate L_{k+1} , it must be sure that the number of L_k is larger than k[7].

Improving the Scan of Database

Improvement of scanning database aims at that we must scan the database once to generate one frequent item set each time and as the classic Apriori Algorithm demands that we will cut the item sets less than min-support degree and save the others. It is obvious that scanning the database repeatedly is unnecessary which should be improved, so that the improvement will reduce most of the times of scanning the database and increase the efficiency of operation and analysis well. The details of the improved theory are expounded below:

We should build a two-dimension array $A[m][n]$ first and then count all the items in the database to rank themselves in a certain way. We scan the whole database again after the rank has already be done, and we define that number 1 equals to items included and 0 equals to opposite. We should save the data as the form of two-dimension array before we count out the number of items included 1 in each line in the two-dimension array. The number is called frequency of including as well which is also called frequency degree of 1 item set. We compare the result with the min-support degree that if it is less than min-support degree, the item set is non-frequent and should be deleted, and if it is larger than min-support degree, the item set is frequent.

Improving the Prune Step

We use the same idea of 3rd quality in theory basis to improve the prune step. The original prune step of classic Apriori Algorithm based on the 1st quality in theory basis: supersets of non-frequent item sets must be non-frequent. However there will be many unnecessary item sets in candidate item sets which increase the workload. So we define new frequent item set after we delete the sets whose elements meet $|L_{k-1}(n)|<k$-1. We get new candidate item set when connecting the new frequent item set with itself.

Improving the Boolean Matrix

We use the same idea of 3rd quality in theory basis to improve the Boolean matrix. We define a candidate *k* item set to be scanned and then we will get *k* frequent item set. We scan the database and compress the matrix in the same time, and it means that when we analyze the support degree, we compare it with the number of 1 included in each line, if the number is less than *k* or equals to *k*, we delete this line so that we need not scan this line again. This theory reduces the time of operation and analysis, and increases the efficiency.

Realization of the Improved Apriori Algorithm

Input: database D, min-support.

Output: frequent item set L

```
(1)   Initalzing Arrary(D,A[m][n+1]);
(2)   L1=find_frequent_1_itemset(A[m][n]);
(3)   for(k=2;Lk-1$\neq \emptyset$;k++)
(4)    { Ck=apriori_gen(Lk-1,min_sup);
         (5)   for each c  Ck;
(6)   for(i=1,i$<=$n;i++)
(7)   if(A[i][C[1]]) $\wedge$(A[i][C[2]]) $\wedge$…(A[i][C[k]]
(8)   c.count++; }
(9)   Lk= { c  Ck|c.count$>=$min_sup } ;
(10)  return L=UkLk
```

III. Conclusions

We use C# to program Apriori Algorithm and improved Apriori Algorithm to test the efficiency and performance, and analyze the risk factors of Type 2 Diabetes Mellitus by data mining with the two models and the preprocessed data. The equipment of the experiment is: Intel i5 CPU, 4G RAM, Win5 system. We compare frequent item sets with support degree, time and number of records, time and support degree, and the 3 figures show the result.

The 1st figure shows the relation between the number of frequent item sets and support degree, and we can get the result that the improved Apriori Algorithm can deleted unnecessary item sets when scanning database. The 2nd figure shows the comparison of time and records. We can directly know that the improved Apriori Algorithm can reduce the time of operation and analysis, and increase the efficiency of algorithm. When the records increase, in another word the database increases, the result becomes obvious. The 3rd figure shows the relation between time and support degree. We can see that, the improved Apriori Algorithm has a better performance and efficiency than classic Apriori Algorithm.

References

[1]Wild S, Roglic G, et al. Global Prevalence of Diabetes-Estimates for the Year 2000 and Projections for 2030[J]. Diabetes Care, 2004, 27(5):1047-1053.

[2]Huang Xiaoxia and Xiao Yunshi. Research and Expectation of the Application of Data Mining[J]. Computer Aided Engineering, 2001, vol.10(4): 23-30.

[3]Shao Fengjing and Yu Zhongqing. Principles and Algorithms of Data Mining[M], Science Press , 2009, 8: 1-2.

[4]Purnami. A New Expert System for Diabetes Disease Diagnosis Using Modified Spline Smoothsupportvector Machine[J].Computational Science and Its Applications, 2010,3:83-92.

[5] Rakesh Agrawa, Jerry Kiernan. Water Marking Relation Databases. Proceeding of the 28th VLDB Conference, Hong Kong, China, 2002.

[6] R. Agrawal, T. Imielinski and A. Swami. Mining Association Rules Between Sets of Items in Large Databases[A]. In Proc. of the ACM SIGMOD Intl Conf. on Mangagement of Data[C]

[7] Shao Fengjing and Yu Zhongqing. Principles and Algorithms of Data Mining[M]. Chinese Water Resources and Hydropower Press, 2003.

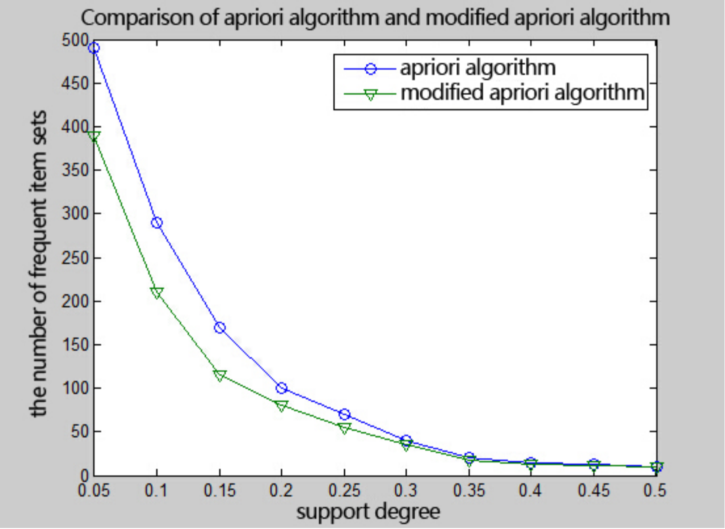

Fig.1 Comparison of the number of frequent item sets and support degree

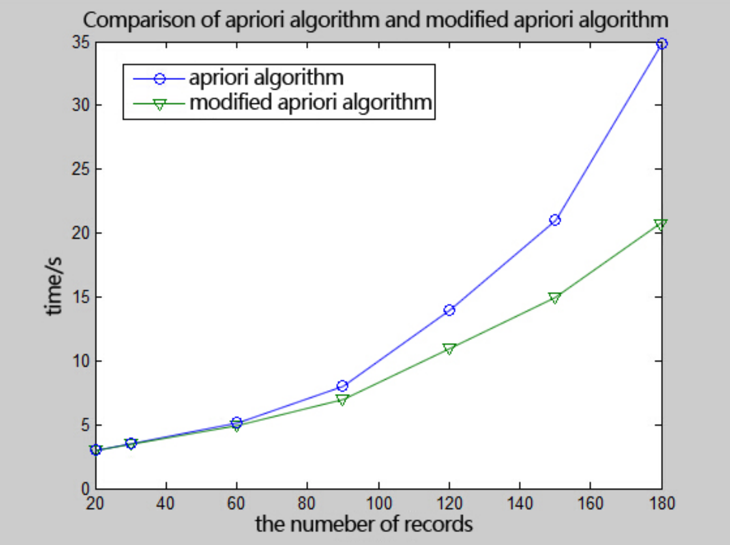

Fig.2 Comparison of time and the number of records

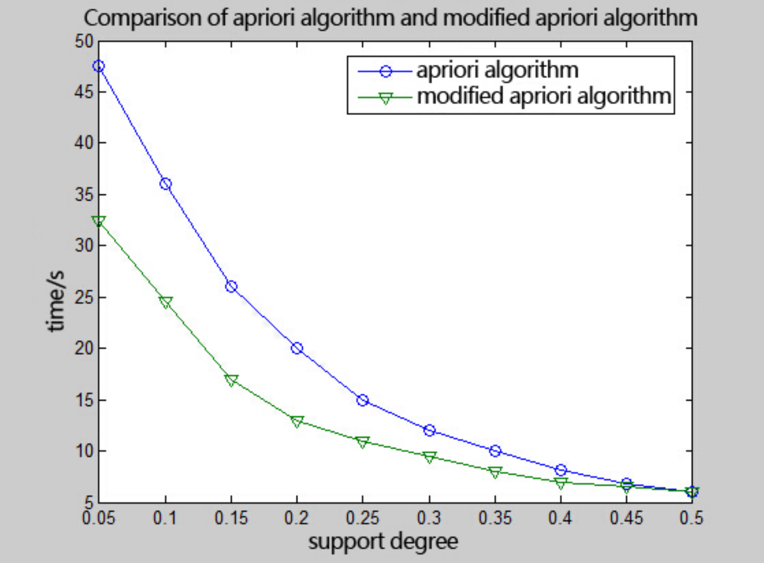

Fig.3 Comparison of time and support degree