An Improved Data Association Rules Mining Algorithm for Intelligent Health Surveillance

Han Yinghua[1,a], Liu Jiaorao[2,b], Miao Yanchun[2,c]

[1]Northeastern University at Qinhuangdao, Qinhuangdao, Hebei, 066004, China

[2]Northeastern University, Liaoning, Shenyang, 110819, China

[a]yhhan723@126.com, [b]1175020381@qq.com, [c]1078114554@qq.com

Keywords: data mining; association rules; Apriori algorithm; Intelligent Health Surveillance.

Abstract. With the growing phenomenon of an aging population, Intelligent Health Surveillance technology has been developing rapidly. Meanwhile, as of things, the development of computer vision and other information technology to make rapid growth of Intelligent Health Surveillance data and diversified characteristics. Therefore, economic significance and the scientific value of the data has been an unprecedented increase. Mining association rules fully business and data, between data become the next hot spot for the Health Surveillance system to promote and applications. Due to the existing Apriori association rules data mining algorithms require to scan the Smart Health Care database many times and generate a large numbers of Health Care candidate sets, which produce giant I/O expense issues, result in low data mining computational efficiency. An improved algorithm based on the Apriori algorithm-the data association rules algorithm for intelligent health surveillance (DAR-IHS) was proposed. Under the premise of scanning database only once, we changed the storage structure of intelligent health monitoring database monitoring data and utilized binary bit operation, which greatly improved the efficiency of the algorithm and supports updating mining.

Introduction

To alleviate the pressure of the population aging and optimize allocation of medical resources, Intelligence Healthy Surveillance has become a topic of general concern nowadays, and attracted the wide attention of domestic and foreign experts and scholars still earlier. At present, researchers have acquired more in-depth results in theory. For example, Least-squares algorithm was proposed, which analysed Health Care data[1]; SPSS software was used to statistical analysis the distribution of the occupational diseases[2]; The statistical methods was adopted for analyzing the occupational health examination data[3];

However, the use of technical analysis association rules Intelligent Health Care data is still in its infancy. Therefore, the study of Intelligent Health Surveillance data for association rules is imperative, and it's also the attention and focus research directions. Many scholars have done a lot of research on this topic and worked for the development of data mining, which have made a great contribution. The improvements of traditional association rule mining are mostly based on Apriori algorithm. The biggest flaw of Apriori algorithm is necessary to repeatedly scan the database, which affects the data mining operating efficiency. Although improved it in many ways later, but the efficiency is still not very high [4,5].P.-G. Cheng et al proposed NFUP algorithm, which joins strong large itemsets into small quantitative of candidate itemsets based on strong large itemsets concept, and adopts early pruning strategy to cut down the times of scanning database [6]. X. Lv et al focus on the issues about large number of candidate itemsets and the time of scanning the database, proposed an efficient algorithm for mining the candidate itemsets to overcome above problems[7]. S.-L. Zhang proposed a new algorithm, which filters out the transactions unconcerned with mining targets by a presupposed filter, greatly improving the whole performance of the algorithm[8]. A. Zeng el at proposed an improved Apriori algorithm based on similarity [9].

This paper proposed an improved association rule data mining algorithm. Under the premise of scanning database only once, we changed the storage structure of Intelligent Health Surveillance

database and utilized binary bit operation, which greatly improved the efficiency of the algorithm and supports updating mining.

Generally, there are two steps in association rules mining: (1) Find out all frequent patterns. (2) Generate all strong association rules. The Apriori algorithm was proposed by R. Agrawal and R. Srikant in 1994, which is the classical algorithm for finding frequent patterns. Apriori adopts a layer-by-layer searching method, where *k*-itemset is used to generate (*k*+1)-itemset.

Intelligent Health Surveillance data processing

This part is to change the structure of the data which is stored, to convert raw transaction data into a temporary database. For elderly people of different genders, with five items: blood pressure, blood oxygen, heart rate, weight, height, the five items we specified them a normal range.

For 60 years or older, we states: normal range of blood pressure is 140/90, the normal range of blood oxygen is greater than 95%, normal heart rate range is 60 to 100.

Man over 60 years old standard weight: standard weight (kg) = height (cm) * 0.65-48.7. Women over 60 years old standard weight: standard weight (kg) = height (cm) * 0.56-33.4. Elderly people weighing more than 10% of the standard weight is obesity. Similarly, Elderly people weight less than 10% of the standard weight is emaciation. Male is set to 1, and the female is set to 0 .

The normal blood pressure is set to binary 0. Abnormal blood pressure is set to 1; Similarly, for oxygen, heart rate, blood sugar, normal are set to 0, abnormal is set to 1 .

This part of the program was written in MATLAB. Table 1 shows the Intelligent Health Surveillance data before processing. Table 2 shows the processed of the Intelligent Health Surveillance data.

Table 1 original Health Care data

High pressure	Low pressure	Oxygen	Heart Rate	Height	Weigh t	Sex
150	92	93	80	75	170	1
130	70	97	90	90	178	1
145	93	98	75	70	168	0

Table 2 processed Health Care data

Hypertension	Hypotension	Oxygen	Heart Rate	Obesity	Thin	Sex
1	0	1	0	1	0	1
0	0	0	0	1	0	1
1	0	0	0	1	0	0

Data Association Rule Algorithm For Intelligent Health Surveillance

Before describing DAR-IHS algorithm, we has given several important definitions as follows:

Definition 1: Collection of all the items in the transaction database D consisting of a database called an itemset I, then $I = \{i_1,i_2,...,i_n\}$, where m is the total number of items contained in D. D is composed of a collection of multiple transactions, and each transaction T is a collection of some of the projects, so each of which specifies a unique transaction identifier (TID).

Definition 2: Collection of D consisting of some of the items referred to item sets, which contain itemset k items is called the k-itemsets.

Definition 3: The support of A to B is the number of A and B in transactions database ratio of the total number of transactions in the database contains A∪ B, which is the probability P (A∪ B).

Definition 4: The confidence of A to B is the number of transactions containing A ratio of the number of A∪ B included in transactions, which is the conditional probability P (B | A).

In an iterative process, Apriori algorithm scan multiple databases and generate a lot of candidates set, which cause low efficiency of the algorithm. To improve the efficiency of the algorithm, we proposed the following improvement ideas. DAR-IHS algorithm describes the process as follows:

Input: A transaction database; minimum supports

Output: The frequent itemsets of D

(1)L_1={frequent_1-itemsets}; //generate the frequent 1- itemsets.

(2) for(k=2;L_{k-1} =;k++) { //get frequent k-itemsets by the frequent (k-1)-itemsets.

(3) begin;
(4) C_k=apriori_gen(L_{k-1}); //generate new candidate k-itemsets by k-1 itemsets.
(5) for all affairs t∈D do{ //for each database transaction t in D.
(6) C_t=subset(C_k,t);
(7) for all candidates c∈C_t do // for each candidate set c in C_t.
(8) c.count++;
(9) end;
(10) L_k={c∈C_k | c.count≥min_sup} //add the candidate itemsets to the frequent item.
(11) end;
(12) return L=∪ $_k L_k$; //itemsets L is output.
The function apriori_gen described as follows:
(1) funtion apriori_gen(L_{k-1},min_sup){
(2) for each itemset l_1∈L_{k-1}; //for each l_1 item set in L_{k-1}.
(3) for each itemset l_2∈L_{k-1}; //for each l_2 item set in L_{k-1}.
(4) if (l_1[1]=l_2[1])∧ .....∧ (l_1[1]=l_2[1]) then{
(5) C=$l_1 \infty l_2$; //after the connection of l_1 and l_2, generating the candidate set C.
(6) if has_infrequent_subset(C,L_{k-1}) then delete C. //candidate set to drop a non-frequent subset.
(8) else add C to C_k . } //adding c to the C_k.
(9) return C_k;}
Function has_infrequent_subset for non-frequent subset test, described as follows:
(1) for each (k-1)-subsets of C
(2) If s □L_{k-1} then
(3) return true;
(4) return false;}

Experimental results and the analysis

Using the DAR-IHS algorithm, we concluded that the largest collection of associated items: {1,5,7}, which means that hypertension, hypotension, female, obesity is the maximum frequent items. The support is set to 0.5, so obese women were more likely suffer from hypertension. In addition, the probability is more than fifty percent. That remind obese women pay more attention to the usual diet, which can prevent them suffering from hypertension.

In order to test executive efficiency of the two algorithms, we use about 400 standard sample simulation data sets, which were provided by Almaden Research Center of IBM. We carried on simulation experiments using the DAR-IHS algorithm and the Apriori algorithm. The results are shown in Fig.1 and Fig.2.

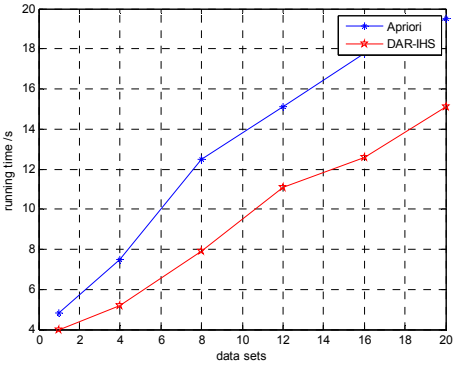

Fig.1 Running time of each algorithm with different data sets

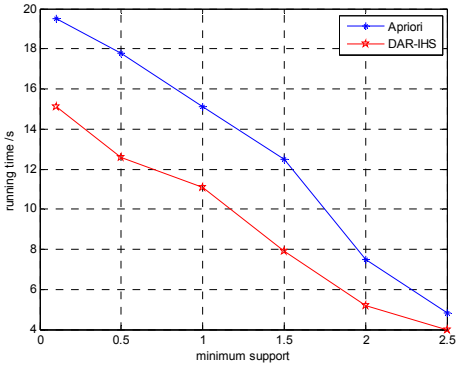

Fig.2 Running time of each algorithm with different support

Fig.1 is the running time of each algorithm with different data sets.As can be seen from the chart, the running time of DAR-IHS algorithm is always less than Apriori algorithm. It is clear that DAR-IHS algorithm is more superior than Apriori algorithm.

Fig.2 is the running time of each algorithm with different support. It's clear that the DAR-IHS algorithm always keeps the optimal running time than Apriori algorithms at different support. Whatmore, with the growth of data sets, the proposed DAR-IHS algorithm has a good scalability and high efficiency performance.

Summary

In this paper, we introduced the DAR-IHS algorithm for discovering the relationships of the Intelligent Health Surveillance data, which is based on Apriori and adopts binary storage structure. Since the traditional Apriori algorithm cause low efficiency, The DAR-IHS algorithm can enhance strong points and avoid weaknesses. As shown in the theoretical analysis and the experimental results, the algorithm can achieve significant improvements in reducing the time overhead. Furthermore, the application of association rules in intelligent health monitoring system achieved a better scientific and intelligent Elderly Health Surveillance. Therefore, we will continue to study the DAR-IHS algorithm and extend it to other areas in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No.61104005 and 61374097, by Natural Science Foundation of Liaoning Province under Grant No.201202073, and by the Central University Fundamental Research Foundation, under Grant.N142304004.

References

[1] Li Qiyi, Zou Yanbiao. Research on analysis method for monitoring data based on time serious modeling. Electrical Engineering Technology, 2006,09:58-59+99+104.

[2] Shao Ahmad, Zhang Jinlong. Health Care Data Analysis in Wuxi City in 2010-2012[J] Chinese Journal of Industrial Medicine, 2014,02: 155.

[3] Gong Jian, Lide Yun . Analysis of occupational hazards among populations under occupational health surveillance in Zhuhai in 2010-2012[J]. Practical Preventive Medicine, 2014,05: 569-571.

[4] K. Taboada, S. Mabu, E. Gonzales, K. Shimada, and K. Hirasawa, Genetic network programming for fuzzy association rule-based classification, in 2009 IEEE Congress on Evolutionary Computation, CEC, Trondheim, Norway, 2009:2387-2394.

[5] D. Becerra, D. Vanegas, G. Cantor, and L. Nino, An association rule based approach for biological sequence feature classification, in 2009 IEEE Congress on Evolutionary Computation, CEC, Trondheim, Norway, 2009: 3111-3118.

[6] P.-G. Cheng, Y. Chen, and X. Yi, Research on an improved association rules data mining algorithm and its application, in 2nd International Conference on Advanced Computer Theory and Engineering, ICACTE, Cairo, Egypt, 2009:1211-1219.

[7] X. Lv, Y. Li, and X. Lu, A web data mining algorithm based on Weighted Association Rules, in 2011 International Conference on Materials, Mechatronics and Automation, ICMMA, Melbourne, VIC,Australia,2011:1386-1391.

[8] S.-L. Zhang, A new mining algorithm of association rules and applications, in 7th International Conference on Intelligent Computing, ICIC, Zhengzhou, China, 2011: 123-128.

[9] A. Zeng, D. Liu, and H. Chen, An improved apriori algorithm based on similarity, in 2012 2nd International Conference on Materials Science and Information Technology, MSIT 2012, Xi'an, Shaan, China, 2012:1825-1829.