Association Rule Mining Based on Estimation of Distribution Algorithm for Blood Indices

Xinyu Zhang
College of Information Science and Engineering
Northeastern University
Shenyang, China
E-mail: zhangxinyu1995@126.com

Guanghu Sui
College of Information Science and Engineering
Northeastern University
Shenyang, China
E-mail: suiguanghu@foxmail.com

Botu Xue
College of Information Science and Engineering
Northeastern University
Shenyang, China
E-mail: xuebotu1994@163.com

Jianjiang Cui*
Inst. Intelligent Systems
Northeastern University, Shenyang, China
E-mail: cuijianjiang@ise.neu.edu.cn
*Corresponding Author

***Abstract*—To come over the limitations of Apriori algorithm and association rule mining algorithm based on Genetic Algorithm (GA), this paper proposed a new association rule mining algorithm based on the population-based incremental algorithm (PBIL), which is a kind of distribution estimation algorithms. The proposed association rule-mining algorithm keeps the advantages of GA mining association rules in coding and the fitness function. Through using probability vector possessing learning properties to update the population, the algorithm increases the convergence speed and enhances the searching ability, compared to GA. In the experiment of mining association, rules in blood indices data, PBIL algorithm performs better not only in running time, convergence speed, but also achieve better searching results. Meanwhile, this paper proposed a parallel algorithm for association rule mining based on PBIL and designed a system architecture based on cloud computing for blood indices analysis, providing a good example to apply the new algorithm to cloud computing.**

Keywords-Distribution estimation algorithm; Probability vector; Blood indices; Parallel algorithm; Cloud computing

I. INTRODUCTION

Association rule mining is an important branch of data mining. By collecting many records of items in the database then analyzing them, the valuable relationships between huge amounts of data can be found [1]. The significance of the association rules analysis is greater in medical data than in other areas. By mining the medical data, the potential relationships between various diseases and various health indicators can be found, serving for medical research and disease diagnosis [2]. Apriori algorithm is the most typical algorithm for association rule mining. Traditional Apriori algorithm needs to scan the database for many times to generate vast candidate sets, leading to poor extensibility of Apriori algorithm. To overcome the weakness above, some scholars put forward a theory using intelligent optimization algorithm to mine the association rules. In 2004, Li Ying [3] put forward the application of generalized genetic algorithm in Apriori algorithm improving. At first, it uses Apriori to search partial association rules, then it uses genetic algorithm to search global association rules. In this way, the times of traversing the database can be reduced. In 2012, Shiwei Chen [4] put forward a method of association rule mining based on interest measure and genetic algorithm, improving the quality of association rule mining. In 2016, Donghao Xu [5] put forward a method of association rule mining based on improved particle swarm optimization algorithm, verifying the advantage of particle swarm optimization on association rule mining compared with genetic algorithm.

With the rapid development of computer technology, cloud computing has become a direction for the future development of distributed computing. MapReduce programming frame put forward by Google is a representative technology of cloud computing. It is suitable for distributed processing of large-scale datasets and has very high computational efficiency [6]. Therefore, some scholars put forward a method of association rule mining based on Hadoop and other cloud computing technology. They also put forward some parallel algorithms for association rule mining. In 2011, Zhang Sheng [7] put forward an Apriori algorithm based on cloud computing. It deploys MRM-Apriori algorithm in MapReduce frame and has good effect in speed.

Based on the research, an association rule mining algorithm based on distribution estimation (PBIL) is proposed in this paper. Compared with Apriori and GA association rule mining algorithm in experiment, the new algorithm is proved to be effective. Meanwhile, an parallel algorithm based on this algorithm is designed, which is suitable for MapReduce frame. In addition, a realization plan for blood indices analysis system based on cloud computing is introduced in this article.

II. The Classical Association Rule Mining Algorithm

Apriori algorithm is one of the most typical algorithms in the data mining field. It uses a method called iteration of layer by layer to produce high dimension frequent item sets from low dimension frequent item sets. Then the association rules can be produced from frequent item sets [8]. The specific mathematical model is in literature 8.

Support degree in Apriori algorithm is defined as follows: the number of transactions of the entire transaction set is m, and there are n transactions containing the item set, then the support degree of the item set is n/m. If the item set A exists, the support degree of the item set is supp(A).

The candidate set in Apriori algorithm is generated layer by layer, and only after fully scanning the database will the frequent item set of this layer be produced. Therefore, if the database is very large, this work will cost a lot of memory resources, reducing the efficiency of the algorithm.

III. Association Rule Mining Based on Genetic Algorithm (GA)

A. Association Rules Model Based on GA

Based on the analysis of the Apriori algorithm, association rule mining is mainly divided into two parts. First, find frequent item sets in the transaction database. The second is to generate association rules based on the frequent item sets which are found [9]. And the workload of the former is the larger one, which is the direct cause of low efficiency of Apriori. Therefore, genetic algorithm (GA) can be used to realize the global search for frequent item sets.

Genetic algorithm is an effective global optimization algorithm. With its binary coding mode, using genetic operators to evolve population, it is able to keep extracting frequent item sets from the transaction database at a rapid speed, avoiding the operations like join and prune which need to frequently scan the database, improving the computing speed and mining precision. The concrete implementation steps are shown in fig. 2.

B. The Weakness of Association Rule Mining Based on GA

Genetic algorithm will save the individuals, which meet the requirements of frequent item sets in each generation to the next generation when it is mining association rules. And the individuals which meet the requirements will be removed from the previous generation. The main purposes of this population selection method are to reserve the excellent individuals and to keep the population diversity. But these two requirements are contradictory. If the excellent individuals reserved are excessive in each generation, the population diversity will decrease, prone to lead to a prematurity phenomenon. As a result, in association rule mining, the search of frequent item sets will be incomplete, and the extracted association rules will be partial. If the excellent individuals reserved are not enough in each generation, the convergence speed of the algorithm will be reduced and the computing time will be longer. That will affect the advantage of association rule mining based on genetic algorithm in speed.

IV. Association Rule Mining Algorithm based on Estimation of Distribution Algorithm (EDA)

A. Description of EDA

To come over the disadvantages of genetic algorithm mining association rules, EDA can be used. EDA is a kind of evolutionary algorithms developed by the genetic algorithm. It first selects samples from the optimal population and extracts information. Then it uses the information to build proper probability module. At last, it updates the population to increment individuals with more fitness until the end condition. In the way it can maximum the individuals' quantity and keep the population diversity [10]. At the same time, EDA can select new solutions by probability distribution to obtain the optimal solutions with less iteration times. It can effectively prevent the local optimization and precocity in GA when dealing with higher order or long-distance tectonic block problems [11].

B. PBIL Probability Module

When handling the problem, which owns mutual independent variables, PBIL algorithm, a typical form of EDA, can be used. PBIL algorithm is mainly applied to binary-code optimization problem. PBIL collects the data recording the values of variables, whose value is ruled as 0 or 1, to build the probability vector. Then it uses the probability vector to estimate the one-dimensional edge distribution. Assuming that a binary gene population with N gene positions (mutual independent variables) and M individuals is existing and the population can evolve continuously, the gene population on the tth generation can be expressed as:

$$g_i^{X_t^j} = \begin{cases} 0 \\ 1 \end{cases} \quad (i = 1,...,M,\ j = 1,...,N) \tag{1}$$

Where t represents the evolutional generation number of the gene population, $X_t^j \ (j = 1,...,N)$ represents the jth gene position (independent variable) at the tth generation, $g_i^{X_t^j}$ represents the code of X_t^j th gene position (variable) of the ith individual at the tth generation.

Then we use the code condition of every variable at the tth generation to generate the probability vector:

$$P_t = \left(P_t\left(X_t^1\right)\ \ P_t\left(X_t^2\right)\ \ P_t\left(X_t^3\right)\ \ ,...,\ \ P_t\left(X_t^N\right)\right) \tag{2}$$

Then we count the amount of the individuals whose designated gene position (variable) are of code 1 and calculate the percent in total M individuals at current gene population. The percent obtained is equal to the distribution probability of the designated variable as follows:

$$P_t\left(X_t^j\right) = \frac{\sum_{i=1}^{M} g_i^{X_t^j}}{M}\left(g_i^{X_t^j} == 1\right) \tag{3}$$

When using the EDA, it first generates a random original population and figures out the fitness value of every individuals of the population. Then it ranks all individuals in order of corresponding fitness value. Individuals with greater fitness values will be seen more advanced. Then truncation is adopted to select advanced individuals of certain amount. The rate of truncation is named as selerate. So the amount of advanced individuals m can be expressed as:

$$m = selerate \cdot M \tag{4}$$

Therefore, the advanced population is made up of the first ranked m individuals of the original population.

Through equation (3) PBIL algorithm gains the probability vector by those advanced individuals. Then it takes samples basing on the probability vector and the next-generation population is obtained. At the same time to make probability vector describe the probability distribution of the advanced individuals with faster speed and more accurate quality, PBIL algorithm adopts the Heb rules from machine learning theory to update the probability vector, which means that the probability distribution of each variable is adjusted linearly at a certain learning speed[12] as equation (5) shows:

$$P_{t+1}\left(X_t^j\right) = (1-\alpha)P_t\left(X_t^j\right) + \alpha\frac{1}{m}\sum_{i=1}^{m} g_i^{X_t^j} \tag{5}$$

Where m represents the amount of advanced individuals selected at the tth generation.

The process of sampling from the probability vector can be described as generating a random number ranging from 0 to 1. If the number is greater than the probability vector corresponding to a certain individual gene position, which means, the binary value of the gene position is 1, otherwise 0.

C. Codes of Individuals and Item Set

When dealing with the problem of mining association rules, EDA adopts binary code also. That is, when a patient owns an abnormal blood index, the value of the index is set as 1. If the index is normal, the value is set as 0 instead.

D. Selection of Fitness Function

Fitness function is designed to reflect the frequency of the item set and recognize the frequent item set. Since the criteria of being frequent item set is that the support level of the item set is greater than the minimum support level, the fitness function can be defined as:

$$fitness\left(g_i^{X_t^j}\right) = \frac{Supp\left(g_i^{X_t^j}\right)}{MinSupp} \tag{6}$$

Where $g_i^{X_t^j}$ represents the ith individual at the ith generation. The name of the fitness function is fitness. $Supp\left(g_i^{X_t^j}\right)$ represents the support level of the item set corresponding to a certain individual. $MinSupp$ represents the minimum support level which is given by user.

If an individual owns fitness value greater than 1, it illustrates the item set corresponding to the individual has its support level greater than the minimum support level. The item set is frequent item set. The individual will be reserved as a member of advanced population with updating the probability vector. If the fitness value is less than 1, it means that the item set corresponding to the individual is not frequent. Then the individual is eliminated directly.

E. Procedures of Mining Association Rules by PBIL

Based on the analysis above, specific steps of association rule mining based on PBIL algorithm is shown in Fig .1.

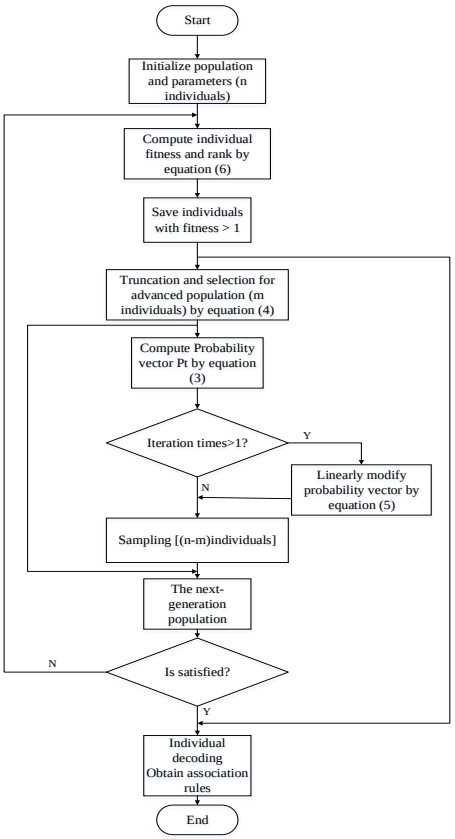

Figure 1. Flow chart of PBIL mining association rules

V. Analysis of Experimental Results

A. Construction of the Transaction Database

All the data in this experiment are from anonymous blood routine laboratory sheets provided by a second grade hospital. Blood routine laboratory sheets provide 9 test indices [13], including white blood cell count (WBC), neutrophil (NE), lymphocyte (LY), monocyte (MO), eosinophil (EOS), basophil (BASO), red blood cell count (RBC), hemoglobin (Hb) and hematocrit (HCT). 255 laboratory sheets were randomly selected to build transaction database. The data in each laboratory sheet is regarded as one item set of transaction database. And each item of the item set is encoded referring to the encoding rules in 4.3 and the test result in the laboratory sheet: Normal index encoding is 0, and abnormal index encoding is 1.

B. The Experiments and Results Analysis

In order to verify the advantages of PBIL algorithm for mining association rules, classical Apriori algorithm and association rule mining algorithm based on genetic algorithm (GA) were compared with PBIL algorithm in the experiments. The three algorithms were compared with each other in three aspects: effect of mining association rules, the time of extracting frequent item sets and the convergence of PBIL.

Referring to the analysis of 3.1, association rule mining model based on GA can be designed. The specific steps are shown in Fig .2.

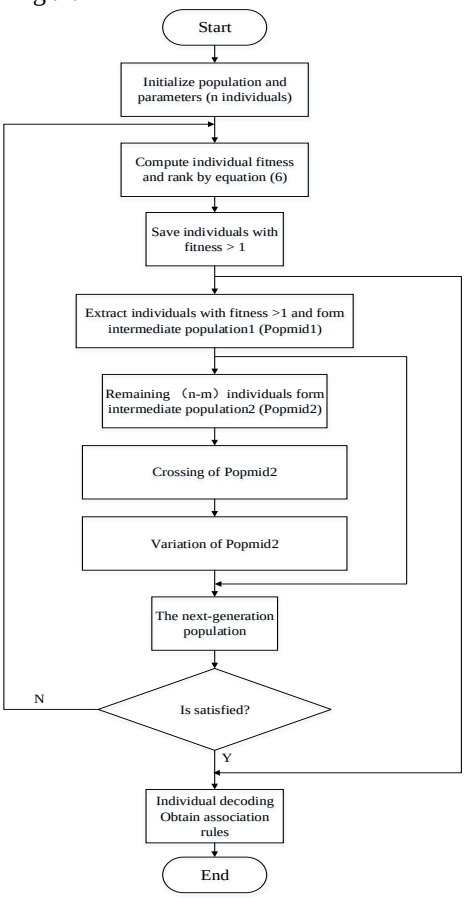

Figure 2. Flow chart of GA mining association rules

Referring to the blood indices data in 5.1, based on Win10 system and Intel Core i5 processor, the algorithm above can be translated to programming in MATLAB 2015a.

C. Analysis of association rules

The transaction database (255 transactions, 9 data items) in 5.1 is calculated by using association rule mining algorithm based on PBIL. The set points are as follows: Minimum support is 0.12 (30/255); Minimum confidence is 0.7; Population size (Popsize) is 500; Iteration times (Iteration) are 100; Truncation selectivity (selrate) is 0.4; Learning rate (learnrate) is 0.1. Merging the similar rules of operation result, the final result is shown in Table I.

TABLE I. FOUND RULES

Rule number	*Rule premise*	*Rule result*	*Support level*	*Confidence level*
1	WBC,NE	LY	0.1294	0.8250
2	WBC,BASO	LY	0.1294	0.7174
3	WBC,Hb	HCT	0.1804	0.7302
4	NE,BASO	WBC	0.1216	0.8185
5	NE,RBC	Hb	0.1608	0.7885
6	NE,Hb	WBC	0.1412	0.8571
7	BASO,Hb	ESO	0.1216	0.7949
8	HCT,RBC	Hb	0.1294	0.8049

The result in Table I can be obtained in classical Apriori algorithm as well. Table I shows the incidence relations between each blood index. For instance, in Association Rule 1, patients with white blood cells, neutrophils and lymphocytes abnormal at the same time are the most common. Therefore, according to the association rule, patients with white blood cells abnormal can be told to prevent or treat diseases caused by abnormal lymphocytes, and vice versa.

D. Mining time comparison between algorithms

Based on the analysis in 3.1, association rule mining can be divided into two stages. The first stage is to find frequent item sets in the transaction database, which costs the main computing time. The second stage is to generate association rules based on the frequent item sets which are found. Parts of the three algorithms in the second stage are the same. Therefore, it is just enough to compare the three algorithms' time of searching for frequent item sets.

E. The relationship between the mining time and the number of transactions

In this experiment, the minimum support is 2/255, and the number of data items (indices) is 9. The three algorithms' computing time can be compared under the premise of searching for the same number of frequent item sets, by keeping changing the number of transactions. The setting parameters of PBIL and GA are as follows: Population size is 500; Iteration times are 100; PBIL truncation selectivity is 0.4; Learning rate is 0.1; GA crossover probability is 0.8; Mutation probability is 0.01. The change of the three algorithms' computing time is shown in Fig .3, and the number of frequent item sets which are found is shown in Table II.

In Fig .3, the changing number of transactions in the transaction set is used as abscissa. The computing time of algorithms is used as ordinate. The gray curve, orange curve and blue curve separately represent the trends of Apriori, GA and PBIL on computing time. In the condition of the same number of data items, the computing time of the three algorithms increases with increment of the number of transactions. The computing time of Apriori is the longest, surpassing PBIL and GA. The PBIL is based on the probability model to evolve population, and it has learning properties. Therefore, convergence of this algorithm is directional. GA is based on rules of crossover and mutation in nature to evolve population. So it has large randomness and doesn't have learning properties. As a result, PBIL has

faster convergence speed than GA, which becomes more obvious with the increasing of data volume.

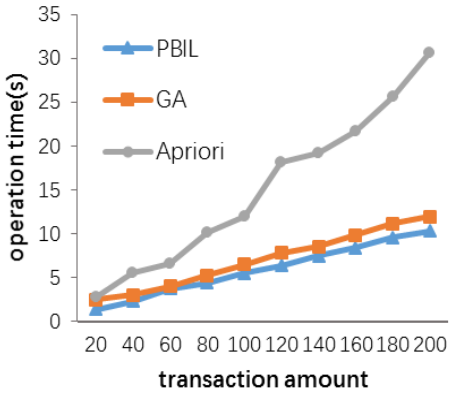

Figure 3. Operation time of algorithms at different transaction amount

TABLE II. FREQUENT ITEM SET'S AMOUNT FOUND IN DIFFERENT TRANSACTIONS' AMOUNT

Order	*Transactions' amount*	*Frequent item sets' amount*
1	20	95
2	40	197
3	60	205
4	80	250
5	100	253
6	120	331
7	140	332
8	160	336
9	180	356
10	200	373

F. The relationship between the mining time and the number of data items

In this experiment, the minimum support is 50/255, and the number of transactions is 255. The three algorithms' computing time can be compared under the premise of searching for the same number of frequent item sets, by keeping changing the number of data items (indices). The change of the three algorithms' computing time is shown in Fig .4, parameter setting and the number of frequent item sets which are found are shown in Table. III.

In Fig .4, in the condition of the same number of transactions, the computing time of the three algorithms increases with increment of the number of data items. The computing time of Apriori increases the most fast with the increment of data dimension. PBIL and GA obtain frequent item sets by searching for them, so the two algorithms are less influenced by data dimension. In addition, PBIL has much faster speed than GA. The reason is as follows: PBIL uses probability vector to evolve population, building a corresponding probability model for each variable of the individual. If an additional dimension is added to the data, a corresponding probability vector will be built. Each of the variables is mutually independent, evolving with its own probability vector, greatly reducing the affect caused by the increment of dimension. GA will strengthen the affect caused by the increment of dimension when additional dimensions are added to the data. It will constantly add high-dimensional data into the population to evolve because of its crossover and mutation. Therefore, PBIL algorithm is better than the former two algorithms.

TABLE III. FREQUENT ITEM SETS' AMOUNT FOUND IN DIFFERENT DATA SETS' AMOUNT

Order	*Data sets' amount*	*Frequent item sets' amount*	*Scale of population*	*Iteration times*
1	3	6	20	5
2	4	7	25	10
3	5	8	35	10
4	6	11	65	20
5	7	16	100	40
6	8	23	150	40
7	9	24	200	50

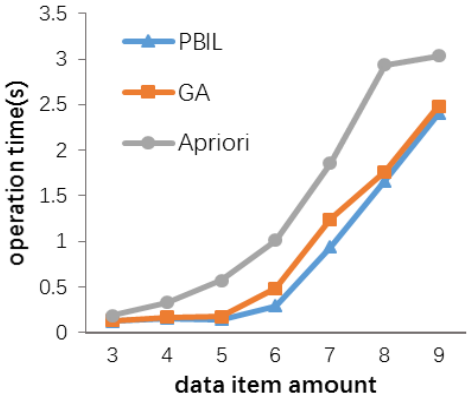

Figure 4. Operation time of algorithms at different data item amount

G. Convergence of PBIL

In this experiment, PBIL will be verified to have better convergence, compared with GA algorithm. The data with 9 transactions and 5 data items is used to experiment. The minimum support is 2/9. Population size is 15. Iteration times are 100. The other parameters are ditto. The best fitness value of each generation can be obtained by operating 20 times. The curve whose convergence speed is the fastest among the 20 operations of GA is used to compare with the convergence curve of PBIL. The result is shown in Fig .5. The maximum support of frequent item sets is 4.5. The two curves represent the two algorithms' ability of searching for frequent item sets with the maximum support. The figure shows that the best fitness value of each generation of PBIL algorithm reaches 4.5 first. GA is later than PBIL for at least 50 iteration periods. What's more, PBIL has found 8 frequent item sets, and GA has found 4. PBIL costs less time as well. Therefore, PBIL algorithm is better at convergence and searching ability.

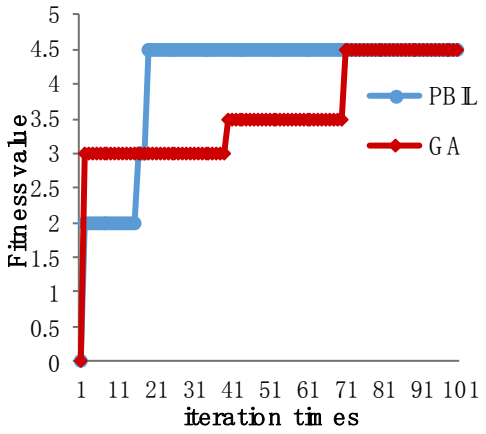

Figure 5. Astringency curve of PBIL vs. GA

VI. Mobile Analysis System of Blood Indices Based on Cloud Computing

A. *Parallel Algorithm of Association Rule Mining Based on PBIL*

The operation speed of the association rule mining algorithm based on PBIL depends on the population scale and iteration times. Therefore, we can consider decomposing the database, declining the scale of the transaction database and gather the results after parallel mining with PBIL algorithm. Basing on the thought above, the currently popular cloud-computing framework Hadoop [14] is used. Then we design the algorithm under the MapReduce framework inside Hadoop and use map function to execute data decomposition and mining. At last we use reduce function to gather the mining results. The framework of the algorithm is shown in Fig .6. Since each map function can achieve parallel computing, which means that searching all frequent item set costs approximately same time as searching a small database, the efficiency of the algorithm is greatly improved.

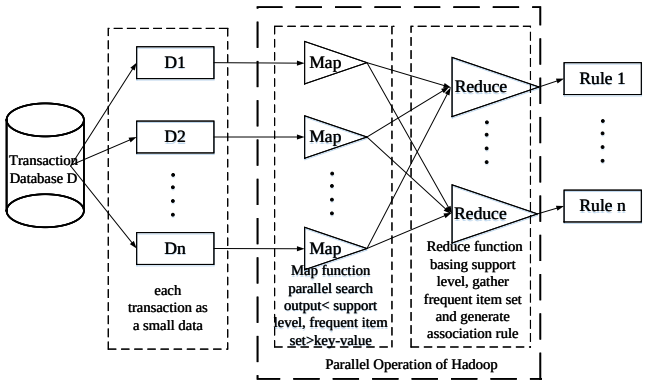

Figure 6. Framework of parallel algorithm mining association rules

B. *Blood-Health Indices Analysis System Based on Cloud Computing*

The framework of the blood-health indices analysis system based on cloud computing is shown in Fig .7. Firstly, the system implements the algorithm in 6.2 by configuring the parallel-computing server cluster basing on Hadoop framework. The newly built database is connected with the hospital's database to update the data dynamically. To make it easier to use the system, an Android mobile application is developed to help analyze the indices of blood. Clients can upload abnormal indices to the server then the server will search for the indices, which form association rules with those indices uploaded according to the computing results. After that, it will read the referred illness symptoms and cure method. Eventually the server will send the information to the mobile clients to accomplish the online analysis of illness.

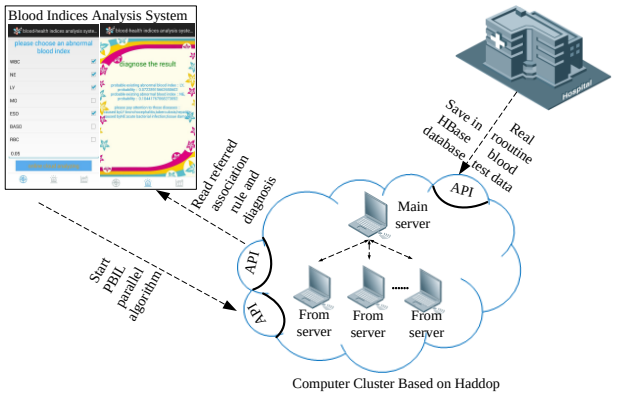

Figure 7. Architecture of blood indices analysis system based on cloud computing

VII. Conclusion

After analyzing the disadvantages of the traditional Apriori algorithm and the module of association rule mining based on GA in operation time and search quality, this paper puts forward a new module of association rule mining based on PBIL and proves that the applied algorithm performed better at searching the frequent item set comparing to Apriori and GA algorithms. Moreover, this paper designs a new parallel algorithm of association rule mining based on PBIL and architecture of the blood-health indices analysis system based on cloud computing which makes good example of the practical application of the algorithm.

References

[1] Jiawei Han. Data Mining: concept and technology [M]. Beijing: China Machine Press, 2004:137-147.J.

[2] Xiaomin Di. Research on Mining Common Risk Factors of Multi-diseases and Predicting Disease [D]. Taiyuan University of Technology, 2013.

[3] Yin Li, Changxiu Cao, Jianghong Ren, etc. Application of General Algorithm (GGA) in the Improvement of Apriori Algorithm [J]. Computer and Modernization, 2004(11):1-3.

[4] Shiwei Chen. Research on Association Rule Mining Based on Interest and Genetic Algorithm [D]. Zhejiang University, 2012.

[5] Donghao Chen, Hongwei Li, Tieying Zhang, etc. Application of Improved PSO Algorithm in Spatial Association Rule Mining [J]. Science of Surveying and Mapping, 2016, 41(2):168-172.

[6] Lämmel R. Google's MapReduce programming model — Revisited [J]. Science of Computer Programming, 2008, 70(1):1-30.

[7] Sheng Zhang. An Apriori—based Algorithm of Association Rules based on Cloud Computing [J]. Communications Technology, 2011, 44(6):141-143.

[8] Zhengchan Rao, Nianbo Fan. A review of associative rule mining Apriori algorithm[J]. Computer Era, 2012(9):11-13.

[9] Guoyan Xu, Yuqing Shi. Application of Genetic Algorithm in Association Rule Mining[j] Computer engineer, 2002, 28(7):122-124.

[10] Zhang Q. On Stability of Fixed Points of Limit Models of Univariate Marginal Distribution Algorithm and Factorized Distribution Algorithm [J]. IEEE Transactions on Evolutionary Computation,2004,8(1):80-93.

[11] Shude Zhou, Zenqi Sun. A Survey on Estimation of Distribution Algorithm [J]. Acta Automatica Sinica, 2007, 33(2):113-124.

[12] H. Muhlenbein, T. Mahnig. Convergence theory and application of the factorized distribution algorithm [J]. Comput. Inf. Technol. 1999,7(1):19–32.

[13] Qin Y J, Sun J S, Wang B Y. The differences of the blood routine indices in patients with fatty liver and non-fatty liver[J]. Journal of ClinicalHepatology,2010.

[14] Qiang Xu, Zhenjiang Wang. Practice of Cloud-computing Application Developing. Beijing: China Machine Press, 2012:64-67.