Presenting a Novel Method for Mining Association Rules Using Binary Genetic Algorithm

Salah Karimi Haji pamagh[1], Dr. Mehdi Afzali[2] and Dr. Amir Sheikh Ahmadi[3]

[1] Department of Computer Engineering, Science and Reseach Branch Kurdistan, Islamic Azad University, Sanandaj, Iran

[2] Department of Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran

[3] Department of Computer Engineering, Islamic Azad University, Sanandaj Branch, Sanandaj, Iran

Abstract

Today, mining association rule is one of the important data mining algorithms which enable managers to make correct decisions based on the knowledge obtained from the detected patterns by databases. Traditional algorithms of discovering association rules such as Apriori and FP-growth may extract millions of rules from databases, many of which are useless, and this issue causes managers to face difficulty to make correct decision. One of the main challenges of rule discovery is presenting a method which can extract useful and approach optimal rules. In this paper, attempts were made to present a new method for useful and optimal mining of association rules in database using binary genetic optimization algorithm. The presented method was implemented by MATLAB R2010b programming language and SQL SERVER 2008 database. Obtained results indicated that the presented algorithm had a high capability in mining optimal association rules and one of the fortes of this method combing with the previous ones was its ability to discover rare rules in large databases.

Keywords: Association rules, Discovering association rules, Genetic algorithm, Support, Confidence.

1. Introduction

Progress its technology in information field, all types of databases have been made for storing information. Therefore, it is increasingly important to analyze these databases for discovering hidden rules [1]. Data mining techniques [1-3] have provided important tools for users in the recent decade. Task of data mining is to extract meaningful information and useful patterns from databases. The knowledge which is obtained from a high volume of data is useful and important [2] and should be accurate, legible and understandable [3]. According to MIT University, novel data mining knowledge is one of the ten developing knowledge areas which will confront the upcoming decade with a technological revolution. Today, this technology is widely used in different fields so that there is no restriction for this knowledge application and working fields of this knowledge range from particles in the floor of oceans to heart of the space. This technique has been successfully applied in commercial, scientific, medical and other fields [1]. Stages of knowledge discovery using data mining are as follows [4]:

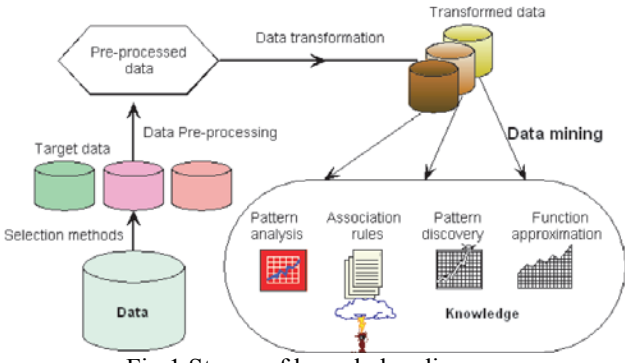

Fig.1 Stages of knowledge discovery

An important data mining algorithms is discovering association rules. This method can discover hidden rules and dependent characteristics and play an important role in decision making [1]. Discovery method of association rules in a database can produce many rules. Thus, it is very important to find efficient and optimal dependent rules in a database [3, 5]. Apriori algorithm is the most well-known algorithm for discovering association rules, which has a very important weakness; these algorithms may produce millions of rules in a large database, most of which are

non-useful; therefore, it can be said that these algorithms are less efficient in large databases [6]. Thus, there is a need for a method which can discover efficient and optimal rules in large databases so that managers can make more effective decisions using these optimal rules. Genetic algorithm [3,7] is one of the optimization methods. Its working scope is very widespread and use of this method has been expanded for optimizing and solving problems with the progress of sciences and technology; this algorithm can be used for optimizing association rules [3].

In this paper, a method was presented for discovering efficient and optimal patterns and rules in databases using binary genetic optimization algorithm [8]. In this method, binary conversion was first done on the data and then the chromosomes were made considering binary data. Afterward, binary genetic algorithm was applied to discover optimal rules in the database. Strength of this algorithm was finding of efficient and optimal rules along with rare rules in the database. Rare rule is a rule which has the highest value considering cost function; this rule may not be discovered in traditional data mining algorithms although they can greatly help managers in decision making.

2. A review of the performed activities

Considering that Apriori algorithm takes long time finding association rules, its computational efficiency is an important issue. There have been many articles all about to substantiate Apriori's efficiency. Avasere et al. [9] in this way has presented an algorithm called partition one which basically different from traditional one. This algorithm scans the database once for finding strong item sets. Then, supporting value for all the item sets is calculated in the second turn of scanning. Accuracy point of Partition algorithm is that strong item sets appear at least in one section.

Park et al. [10] introduced DHP algorithm in 1995. DHP is a derivation of Apriori algorithm with a series of additional controls. On this basis, DHP uses Hash table, which helps candidates to be limited. DHP includes two important characteristics: efficient construction of strong item sets and effective reduction in database size by discarding its characteristics. Toivonen et al. [16] presented a sampling algorithm in 1996, which was related to finding association rules according to reduction in the database activity. DIC algorithm was introduced by Brin et al. [11] in 1997. DIC divides database to several sections called start points. In each start point, DIC algorithm specifies a support value for all the item sets and discovers patterns and rules.

Pincer search algorithm was introduced by Lin et al. (1998). This algorithm can efficiently discover item sets with the highest frequency [12]. In 2001, Yang et al. introduced an efficient method based on Hash which was called HMFS. HMFS combines two methods of DHP and Pincer search method. A combination of these two methods provides two important results: one is HMFS method which can reduce the number of database scan and another can filter iterative item sets for finding the largest iterative item set. These two algorithms can filter total time of computation for finding the largest iterative item set [13].

In recent years, genetic algorithm has been used for discovering association rules [15, 16]. In [14], weighted items were used for showing the importance of unique item sets. Using these weighted items in fitness function of genetic algorithm discovery, value of different rules is determined. This algorithm can find a suitable threshold limit for discovering association rules. In addition, Saggar et al. presented a method for optimizing discovered rules using genetic algorithm. The importance of their work was that it could predict the rules which had negative characteristics [17].

In [3], a new method was presented for making all positive and negative association rules called NRGA, which made all hidden rules using Apriori algorithm. In this paper, names such as CNR, ANR and ACNR were considered for representing negative rules. Also, correlation coefficient equation was modified; therefore, all of its obtained rules were promising. Weakness of this paper was long time and dependence on Apriori algorithm.

3. Genetic optimization algorithm

Genetic algorithm is inspired by genetics and Darwin's theory of evolution and is based on survival of the fittest or natural selection. A common application of genetic algorithm is its use as an optimizing function. Genetic algorithm is a useful tool for pattern recognition [3,18], feature selection, understanding image and machine learning. In genetic algorithms, genetic evolution of organisms is simulated.

Genetic algorithms can be regarded as a directed random optimization method which gradually moves toward the optimal point. Regarding features of genetic algorithm compared with other optimization methods, it can be said that it is the algorithm which can be applied to any problem and has proved efficiency for global optimum finding without any knowledge of the problem and any limitation on type of its variables. This method is able to solve complex optimization problems when classic methods are neither applicable not reliable for finding global optimum.

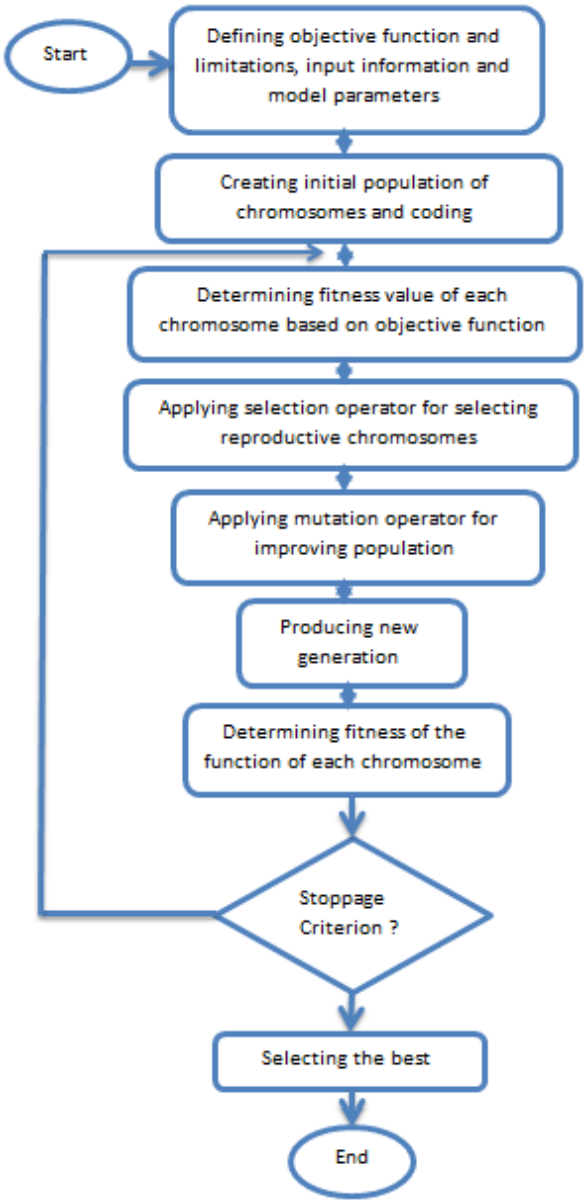

Fig.2 Flowchart of binary genetic algorithm [8]

In 70s, a scientist from University of Michigan, John Holland, introduced the idea of using genetic algorithm in engineering optimizations. The main idea of this algorithm is transfer of hereditary traits by genes. Genes are pieces of a chromosome which have the required information for a DNA molecule or a polypeptide. In addition to genes, there are different types of different regulatory sequences on chromosomes which participate in replication, transcription, etc. [19, 20].

Now, it can be mentioned that genetic algorithm is the tool, using which a machine can simulate natural selection mechanism. This action is performed by searching in problematic case for finding better and not necessarily optimal responses. Genetic algorithm can be called a general searching method which imitates natural biological evolution rules. In fact, genetic algorithms use Darwin's natural selection principles for finding an optimal formula for predicting or adapting patterns. Genetic algorithms are almost good options for regression-based prediction techniques [8].

Similarly, genetic algorithm method starts with specifying decision variables, objective function and limitations. In Figure 2, flowchart of binary genetic algorithm is presented. Objective function attributes a quantitative value to a set of special values for variables (chromosomes). Using objective function, optimization algorithm is directed toward improving values of the variables to reach optimal value of objective function. Genetic algorithm method starts with specifying a set of chromosomes; each of which is a chain of continuous genes and each gene indicates a decision making variable of the problem [8].

4. Discovering association rules

Agarwal et al. introduced issue of mining association rules in 1993 [21]. A positive association rule as if A→B where A and B indicates item set (I) and each item set includes all items of {i1, i2, …, in}. In database D = {T1, T2, . . ., Tk}), it can be shown that a customer purchases product B after A has purchased A provided that A∩B = Ø. Association rule mining should be based on two parameters of confidence and support. Each rule is useful when value of these two parameters is close to the threshold limit determined by the user. Definition of two parameters of confidence and support is as follows [3]:

1) Minimum support: Finding item sets, the support value of which is higher than the threshold limit.

$$Support\ (A \rightarrow B) = P\ (A \cup B) = \frac{A \cup B}{\Sigma} \qquad (1)$$

2) Minimum confidence: Finding the item sets, confidence of which is above the threshold limit:

$$Confidence\ (A \rightarrow B) = p\ (B|A) = \frac{A \cup B}{A} \qquad (2)$$

A better rule is the one which has high confidence and support values and the rules with high confidence and support values should be sought.

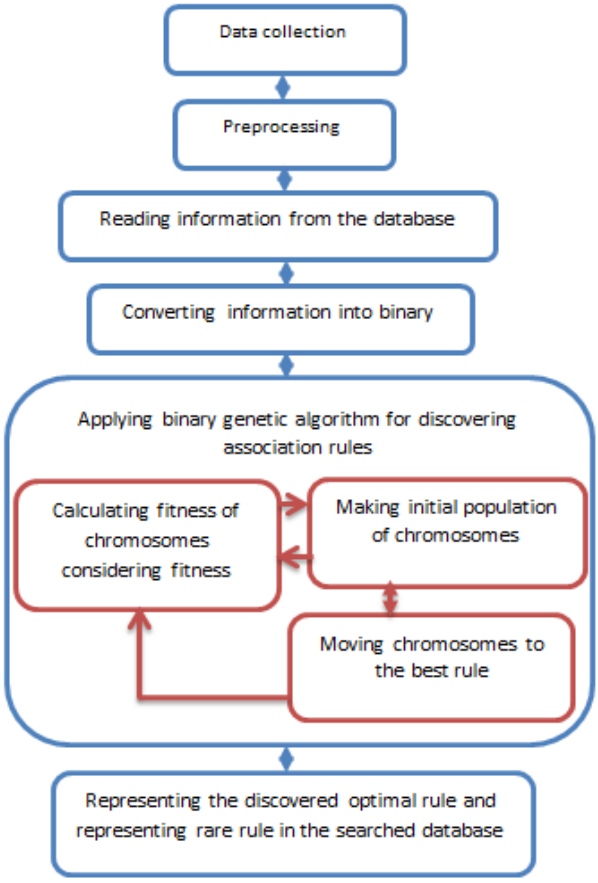

Fig.3 Flowchart of discovering association rules using binary genetic algorithm

5. The proposed algorithm

Performance stages of the present research are based on Figure 3. The proposed algorithm includes of three parts: 1) preprocessing; in all papers and researches which have been conducted on data mining, data collection and preprocessing are among the main stages and devote the longest time and highest cost to themselves [22], 2) binary conversion; in this stage, binary equivalent of the information is obtained and 3) application of binary genetic algorithm; in this stage, application of the proposed algorithm to the preprocessed data is discussed. In this stage, initial population of genes is made and fitness value of each gene is calculated. Finally, GBest is selected from among this population and is introduced as a rare rule. When applying the algorithm, each rule which is larger than the threshold limit is introduced as an optimal rule.

The initial material used in data mining is data. For this reason, the cornerstone of good data mining operations is application and access to good and suitable initial data. It is very difficult to collect and prepare data [22]. In the present research, data stored in Zoo database [3, 24] were used. This database was obtained from UCI site. In this site, different universities have placed databases for applying data mining algorithms. Zoo database had 101 records of information for different animals. Each record had 18 features. In this paper, six binary features were used.

As mentioned above, binary genetic optimization method was used for producing hidden rules in the proposed algorithm. Before a genetic algorithm can be executed, the related problem should be properly coded or represented. The most common method of representing chromosomes in genetic algorithm is binary strings. Each decision making variable becomes binary and then chromosome is created by juxtaposing these variables. Each chromosome in this proposed method indicated a rule and each rule was composed of two consequent and antecedent parts. Figure 4 shows a chromosome [22]; white section is antecedent and gray section is consequent. Each one of the cells in the antecedent or consequent indicates a field of database. Values of these cells are numbers on the basis of 2, which is a value for the desired field in database.

Field1	Field2	Field3	Field4	Field5
Field1	Field2	Field3	Field4	Field5

Fig.4 Representing a chromosome

For example, Figure 5 represents the following rule or condition:

IF (tail= 1 AND milk= 1) → (hair= 1, toothed= 1)

1	0	1	0	0
0	1	0	0	1

Fig. 5 An example of a chromosome in the desired database

As observed above, the information available in each field should be converted into its binary code. The following algorithm was used for converting information into binary:

For each column in Table of Data Base
count_field (i) = The number of DISTINCT value in each column
bitarray (i) = [log2(count_field(i))]+1

In this algorithm, all fields of the related table are met in a loop and the number of unique elements of i-th field is obtained and placed inside the variable count_field(i). Then, the number of necessary bits for i-th field is obtained considering count_feild(i). This value is added to 1 and placed in bitarray(i). The reason of summing is that the number of bits should be more to show the number of elements in i-th field to increase probability of lack of

involvement of a field in a rule. Now, the i-th field has bitarray(i) cell in each chromosome and each cell can be zero or one and i-th field of each chromosome can indicate zero number up to 2bitarray(i). Assume that numerical value corresponding to binary bits of i-th field is equal to M; then:

- If $0< M <=2^{bitarry(i)-1}$, then i-th field participates in the rule made by the chromosome.
- Otherwise, i-th field does not participate in the rule made by chromosome.

For example, in Bank Marketing database, after discretization, field Age has three values of 1 (15 to 34), 2 (35 to 50) and 3 (larger than and equal to 51). Now, 2 binary bits are considered for showing values of this field, in which numbers are like Table 1.

Table 1: An example of binary conversion

Binary value	Concept
00	Means no participation of field Age in the rule
01	Participation of field Age in the rule with value 1 (15 to 34)
10	Participation of field Age in the rule with value 2 (35 to 50)
11	Participation of field Age in the rule with value 3 (larger and equal to 51)

To implement the proposed algorithm, MATLAB software, version R2010b, was used. Considering the abilities, this software can implement the proposed algorithm. In the proposed algorithm, MATLAB diagrams were used for representing movement of chromosomes toward the best target.

5.1 Designing fitness function

Directing the chromosome inside the population toward an optimal response is done using fitness function. it is usually assumed that the chromosome with the largest fitness value is the best chromosome. In the proposed method, A and B were the characteristics participating in the antecedent and consequent parts of the rules, respectively, which were obtained using the corresponding chromosome decoding according to the previous section. In this paper, coefficients of confidence and support were obtained for the rules as follows:
To produce rules as *if* $A \rightarrow B$, the following two criteria were used in function *Cost (P)* to measure quality of association rules:

$$\text{Support} = \frac{Supp(A \cup B)}{N} \quad (3)$$

$$\text{Confidence} = \frac{Supp(A \cup B)}{Supp(A)} \quad (4)$$

Where N is total number of transactions and $SUPP(A \cup B)$ is frequency of items A and B in total transactions. Function *Cost (P)* obtains value of confidence and support for each chromosome. After sending particles to fitness function, the chromosome which has the highest fitness level is used for moving other chromosomes toward the most optimal rule. To measure value of each chromosome or rule, a criterion is required which specifies value of that chromosome or rule considering confidence and support values. Fitness function was defined in this research as follows [3, 22, 23]:

$$Fitness = \alpha_1 * Support + \alpha_2 * Confides - \alpha_3 * NA \quad (5)$$

where NA is the number of characteristics participating in the produced rule and coefficients α control effect of each parameter inside fitness function. If the user needs, they can be regulated as desired. As shown above, the first and second parts of this function are related to calculation of support and confidence values of the produced rule. It seems necessary to consider these two components together for calculating fitness of the produced function because confidence or support degrees alone cannot be a criterion for judging quality of the produced rule. It is evident that the rule has high quality when these two factors have high value. On the other hand, it is known that the probability of redundant characteristics which reduce quality of the produced response is high in rules with large length. As a result, the third section produces the rules with relatively short length and higher legibility, understandability and quality, which are of special importance in data mining.

6. Results of simulation and analysis of the proposed algorithm

Genetic algorithm applies survival of the fittest on a series of responses of the problem with the hope for obtaining better responses. In each generation, better approximates of the final response are obtained using the operators which have imitated natural genetics. This process causes new generations to be more compatible with conditions of the problem. Considering characteristics of the genetic algorithm, it can be said that the proposed method had high ability to discover association rules. Zoo database was used to test the proposed algorithm and the obtained results were compared with the algorithms [3]. The proposed algorithm

was called GBAR. First, parameters of the algorithm were regulated as follows:

Table 2. Parameters of GBAR algorithm for discovering association rules

Parameters	Population size	Number of frequencies	α_3	α_2	α_1	Threshold limit of support	Threshold limit of confidence
Values	70	100	0.2	0.8	0.8	0.2	0.5

The chart below shows, the proposed algorithm has been applied to Zoo database. In this section of the algorithm, iterative rules and rules with lower degrees of confidence and support than what was specified in Table 2 were eliminated from the produced rules and the remaining rules were introduced as final rules. Final optimal rule which was found by particles was introduced as rare rules. Then, results of the proposed algorithm were described. In Diagram 1, optimization manner of chromosomes toward the best state is shown.

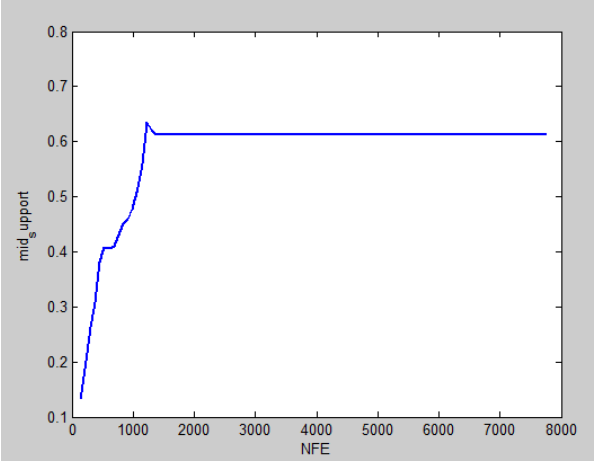

Diagram 1. Mean support of chromosomes

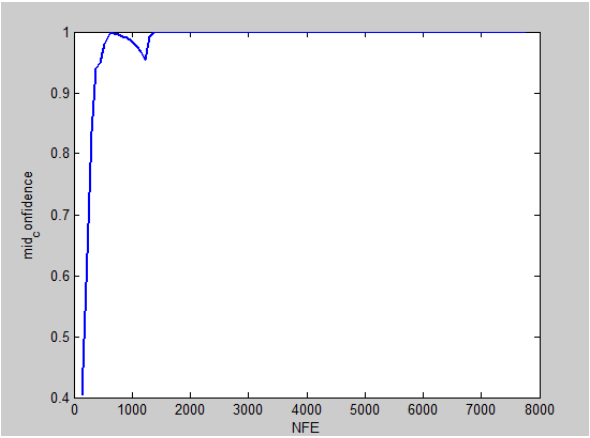

Diagram 2. Mean confidence of chromosomes

As mentioned above, Apriori algorithm may extract many rules from a database, most of which are not useful or may not be able to discover many rules that are important for managers. Apriori algorithm was applied to the desired database. Considering memory leakage, this algorithm extracted 40000 rules and, when the results were filtered based on the threshold limit of Table 2, this algorithm only could discover 16 useful rules. Mean confidence of the rules discovered by Apriori was equal to 0.7; but, mean confidence of the discovered rules was equal to 0.84, which indicated that the proposed algorithm was able to discover rules with better confidence coefficient than Apriori algorithm. In Diagram 3, the rules discovered by these two algorithms are shown.

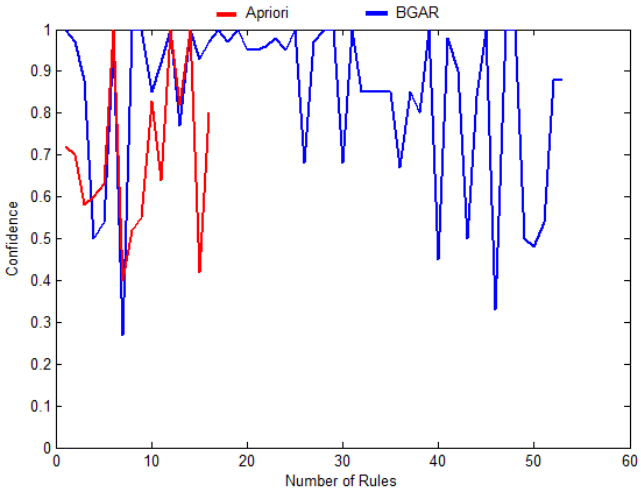

Diagram 3. Rules discovered by two Apriori and BGAR algorithms

7. Conclusion

Today, mining association rules is an important data mining algorithms. Traditional algorithms for discovering association rules like Apriori and FP-growth may extract millions of rules from database, most of which are not useful and cannot help managers to make decisions. One of the important challenges of discovering association rules is presentation of a method which can extract useful, optimal and efficient rules. In this paper, a new method was presented for mining useful and optimal association rules in database using binary genetic optimization algorithm. The obtained results showed that the presented algorithm had high ability to mine optimal and efficient association rules which could be very useful for managers in decision making. One of the strengths of this method compared with the previous methods was its ability to discover rare rules in large databases.

8. References

[*1*] Ren Jie Kuo, Chie Min Chao and Y.T. Chiuc. *"Application of particle swarm optimization to association rule mining*": Applied Soft Computing 11, (2011) pp: 326 –336.

[*2*] Sigurdur Olafsson, Xiaonan Li, and Shuning Wu *"Operations research and data mining*",in: European Journal of Operational Research 187, (2008), pp:1429–1448.

[*3*] Dewang Rupesh, Agarwal Jitendra. *"A New Method for Generating All Positive and Negative Association Rules":* International Journal on Computer Science and Engineering, *(2011)* ,Vol. 3, pp:*1649-1657.*

[*4*] Agrawal Rakesh, Tomasz Imielin´ ski, Arun Swami. *"Mining association rules between sets of items in large databases":*ACM SIGMOD Record 22 (2) *,(1993),* pp:*207 – 216.*

[*5*] Tan Pang-Ning, Steinbach Michael, Kumar Vipin. "*Introduction to Data Mining."* March 25, (2006).

[*6*] Veenu Mangat:"*Swarm Intelligence Based Technique for Rule Mining in the Medical Domain."* International Journal of Computer Applications, (2010), vol.4 pp: 19-24.

[*7*] David E Goldberg, Holland John H, *"Genetic Algorithms in Search, Optimization and Machine Learning*", Reading,MA: Addison-Wesley.

[*8*] Haupt, Randy L, Sue Ellen Haupt. *"Practical Genetic Algorithms",* Second Edition,ISBN *0-471-45565 -2.*

[*9*]Ashok Savasere, Edward Omiecinski, Shamkant Navathe. *"Mining for strong negative associations in a large database of customer transactions":* ICDE*, (1998),* pp: *494-502.*

[*10*] Jong Soo Park, Ming-Syan Chen, Philip S Yu. *"An effective hash-based algorithm for mining association rules."* International Conference on Management of Data*, (1995)* pp: *175 –186.*

[*11*] Toivonen Hannu. *"Sampling large databases for Association Rules" VLDB, (1996) pp: 1-12. Conference. India.*

[*12*] Dao-I Lin, Z.M. Kedem. *"Pincer search: a new algorithm for discovering the maximum frequent set",* in: Proceeding of the 6th International Conference on Extending Database Technology: Advances in Database Technology*,* (*1998*), pp.*105 –119.*

[*13*] D.L. Yang, C.T. Pan, Y.C. Chung, *"An efficient hash-based method for discovering the maximal frequent set, in":* Proceeding of the 25th Annual International Conference on Computer Software and Applications, (*2001*) pp*. 516 –551.*

[*14*] S.S. Gun, *Application of genetic algorithm and weighted itemset for association rule mining, Master Thesis, Department of Industrial Engineering and Management,* Yuan-Chi University, (*2002*).

[*5*] Ashraf El-sisi*. "Fast Cryptographic Privacy Preserving Assocaition rules mining on Distributed Homogenous database"*: The International Arab journal of information Technology, (2010), vol 7.

[*16*] Xiaohui Yuan, Buckles B. P, Zhaoshan Yuan, Jian Zhang.*"Mining Negative Association rules":* Proceedings of Computer and Communications, *(2002).*

[*17*] M. Saggar, A.K. Agrawal, A. Lad, *"Optimization of association rule mining using improved genetic algorithms",* in: Proceeding of the IEEE International Conference on Systems Man and Cybernetics, vol. *4*, *(2004),* pp. *3725 –3729.*

[*18*] Ya-ling Tang and Feng Qin: *"Research on Web Association Rule Mining structure with Genetic Algorithm":*Proceeding of the 8th world congress on Intelligent Control and Automation .China, *(2010).*

[*19*] J.H Holland, *"Adaptation in Nature and Articial Systems",* University of Michigan Press, *Arbor* Ann, (1975).

[*20*] Huang Wenqi. Jin Renchao. *Jin, the Quasiphysical Personiocation Algorithm for Solving SAT. Problem-Solar",* Science in China, Series E, no.2, 179-186 (in Chinese), *(1997).*

[*21*] Rakesh Agrawal ,Ramakrishnan Srikant. "*Fast Algorithms for Mining Association Rules*". Proceedings of the 20th VLDB Conference .Santiago, Chile*,* (*1994*).

[*22*] Abdoljabbar asadi, Mehdi Afzali. "*Providing a new method for detecting positive and negative optimal performance association rules in very large databases using Binary Particle Swarm Optimization*"*:* The sixth Iran Data Mining Conference / IDMC, Dec 01,01 / 2102, Tehran, Iran,(*2102*).

[*23*] Abdoljabbar Asadi, Mehdi Afzali, Azad Shojaei, Sadegh Sulaimani. "*New Binary PSO based Method for finding best thresholds in association rule mining".* Life Science Journal; 9(4).pp:*260 - 264,*(*2012*).

[*24*] http://archive.ics.uci.edu/ml/