ISSN:2321-1156

International Journal of Innovative Research in Technology & Science(IJIRTS)

An Improved Association Rule Mining Approach Using Distance Weight and Ant Colony Algorithm

Girish Kumar Ameta,M.Tech Scholar,ACEIT,Jaipur; Dr. Vibhakar Pathak, Professor & Guide,ACEIT Jaipur;

Abstract

The growing rate of data is a challenging task for mined useful association rule in data mining field. The classical association rule mining generates rule with various problems such as pruning passes of transactional database, generation of negative rules and superiority of rule set. Time to time several researches modifies classical association rule mining with different approaches. But in recent scenario the association rule mining is suffering from superiority of rule generation. This problem of association is solved by multi-objective association rule mining, but still this process has been suffered by continuity of rule generation.

In this paper we are proposing a new algorithm distance weight optimization of association rule mining using Ant colony Algorithm. In this method, we find the near distance of rule set that uses equalize distance formula and generate two class higher class and lower class .The validation of class checked by distance weight vector. Basically distance weight vector maintain a threshold value of rule item sets. In whole process, we use the Ant Colony algorithm for optimization of rule set. Here we set population size is 1000 and selection process validate by distance weight vector.

I Introduction

This paper describes our new proposed algorithm *distance weight optimization of association rule mining*, implantation and working of algorithm. The proposed algorithm is implemented with the Ant colony and, compared with multi-objective association rule optimization using genetic algorithm and apriori algorithm. This paper also suggests that proposed algorithm is better rule set generator as compared to the MORA GA and aproiri method. *Section 2* of this paper describes about introduction of the association rule mining and challenges of finding the interested patterns among the item sets. *Section 3* describes about existing approaches and description of the related work for better association rule mining and methods to meet such challenges of the data mining. *Section 4* describes the proposed algorithm and new work done on the MORA with its usage in the association rule mining. *Section* 5 describes the implementation of new algorithm and result analysis.

II Association Rule Mining

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence[1]. Suppose one of the large item sets is Lk, Lk = {I1, I2… Ik}, association rules with this item sets are generated in the following way: the first rule is {I1, I2… Ik-1} ⇒ {Ik}, by checking the confidence this rule can be determined as interesting or not. Then other rules are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second sub problem is quite straight forward, most of the researches focus on the first sub problem. The first sub-problem can be further divided into two sub-problems: candidate large item sets generation process and frequent item sets generation process. We call those item sets whose support exceed the support threshold as large or frequent itemsets, those item sets that are expected or have the hope to be large or frequent are called candidate item sets. In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large[3]. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only "interesting" rules, generating only "no redundant" rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength. All methodology and process are not described here. But some related work in the field of association rule mining by the name of authors and their respective title.[4]

2.1“Improvement on the Constrained Association Rule Mining Algorithm of Separate” –

In this title authors describe the constrained technique for optimization of association rule mining as Separate is a desirable algorithm in terms of efficiency and candidate generation. However, Separate is not perfect due to deficiency of its joint function, especially when the length of item set or the number of candidate item sets is large. In this paper, three lemmas are proposed and proved mathematically; and based on these lemmas, a novel early stop function is designed elaborately. The early stop algorithm is capable of breaking the process of loop in the case of dissatisfying the join term, and by this means, performance is improved remarkably. Experiments have demonstrated that the proposed algorithm is more preferable compared with the currently-used join function. To improve the performance of Separate algorithm, a novel Early Stop algorithm is designed elaborately according to three lemmas. It has been validated experimental that Early Stop outperforms Join function in terms of execution time, although there is no any daunting programming effort involved. In the future work, the authors will consider the application of Early Stop in other Apriori-based algorithms.[5]

2.2 “An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules “-

In this title author describe a minimum support of multiple term for optimization of association rule mining. Rare association rules are the association rules containing rare items. Rare items are less frequent items. For extracting rare item sets, the single minimum support (minsup) based approaches like Apriori approach suffer from “rare item problem” dilemma. At high minsup value, rare itemsets are missed, and at low minsup value, the number of frequent item sets explodes. To extract rare item sets, an effort has been made in the literature in which minsup of each item is fixed equal to the percentage of its support. Even though this approach improves the performance over single minsup based approaches, it still suffers from “rare item problem” dilemma. If minsup for the item is fixed by setting the percentage value high, the rare item sets are missed as the minsup for the rare items becomes close to their support, and if minsup for the item is fixed by setting the percentage value low, the number of frequent item sets explodes. In this paper, we propose an improved approach in which minsup s fixed for each item based on the notion of “support difference”. The proposed approach assigns appropriate minsup values for frequent as well as rare items based on their item supports and reduces both “rule missing” and “rule explosion” problems. Experimental results on both synthetic and real world datasets show that the proposed approach improves performance over existing approaches by minimizing the explosion of number of frequent item sets involving frequent items and without missing the frequent item sets involving rare items. Most important, the proposed approach ensures that the difference between the support of an item and the corresponding minimum support remains constant for all items including rare items. As a result, it efficiently reduces the explosion of frequent item sets involving frequent items without affecting the extraction of frequent item sets involving rare items. We have evaluated the performance of the proposed approach by conducting experimental results on both synthetic and real world datasets. The results show that, as compared to existing approaches, the proposed approach prunes frequent item sets involving frequent items in a more efficient manner and without missing the frequent item sets involving rare items.

2.3 “Optimized Association Rule Mining with Genetic Algorithms“ -

The mechanism for unearthing hidden facts in large datasets and drawing inferences on how a subset of items influences the presence of another subset is known as Association Rule Mining (ARM). There is a wide variety of rule interestingness metrics that can be applied in ARM. Due to the wide range of rule quality metrics it is hard to determine which are the most ‘interesting’ or ‘optimal’ rules in the dataset. In this paper we propose a multi–objective approach to generating optimal association rules using two new rule quality metrics: syntactic superiority and transactional superiority. These two metrics ensure that dominated but interesting rules are returned to not eliminate from the resulting set of rules.[8] Experimental results show that when we modify the dominance relations new interesting rules emerge implying that when dominance is solely determined through the raw objective values there is a high chance of eliminating interesting rules. Keywords: optimal association rules, genetic algorithms, multi–objective interestingness metrics we have observed that when we modify the dominance relations new rules in large numbers are found. This implies that when dominance is solely determined through support and confidence, there is a high chance of eliminating interesting rules. With more rules emerging it implies there should be a mechanism for managing their large numbers and also to significantly improve the response time of the algorithm.[7]

III Description of Approaches

We proposed a novel algorithm for optimization of association rule mining, the proposed algorithm resolve the problem of negative rule generation and also optimized the process of superiority of rules. Superiority of association rule mining is a great challenge for large dataset. In the generation of superiority of rules association existing algorithm or method generate a series of negative rules, which generated rule affected a performance of association rule mining. In the process of rule generation various multi objective association rule mining algorithm are proposed but all these are not solve superiority problem of association rule mining.

In this paper we proposed distance weight optimizations of association rule mining with ant colony optimization. In this algorithm we used second odder quadratic equation and nearest neighbor classification technique for the selection of set of candidate of superiority of key for generation of rules. In the generation of rule selection of support value of transaction data set is play a important role , for this role we used heuristic search algorithm for better searching of support value for generation of optimized association rule.

In the process of novel algorithm for rule optimizations fist we discuss KNN and ANT algorithm and finally we proposed a hybrid method for optimization of association rule mining (DWORAM).

3.1 KNN

In the process of optimization of algorithm of association rule mining we used knn method for classification of superior support count and confidence value of itemset. Knn is a very famous algorithm for data classification. Here we describe process of knn methodology for classification of support and confidence.

Suppose each sample in our data set has n attributes which we combine to form an n-dimensional vector: x = (x1, x2. . . xn). These n attributes are considered to be the independent variables.

Each sample also has another attribute, denoted by y (the dependent variable), whose value depends on the other n attributes x[12]. We assume that y is a categorical variable, and there is a scalar function, f, which assigns a class, y = f(x) to every such vectors. We do not know anything about f (otherwise there is no need for data mining) except that we assume that it is smooth in some sense. We suppose that a set of T such vectors are given together with their corresponding classes: x(i), y(i) for i = 1, 2, . . . , T. This set is referred to as the training set. The problem we want to solve is the following. Supposed we are given a new sample where x = u. We want to find the class that this sample belongs. If we knew the function f , we would simply compute

v = f(u) to know how to classify this new sample, but of course we do not know anything about f except that it is sufficiently smooth. The idea in k-Nearest Neighbor methods is to identify k samples in the training set whose independent variables x are similar to u, and to use these k samples to classify this new sample into a class, v. If all we are prepared to assume is that f is a smooth function, a reasonable idea is to look for samples in our training data that are near it (in terms of the independent variables) and then to compute v from the values of y for these samples.

When we talk about neighbors we are implying that there is a distance or dissimilarity measure that we can compute between samples based on the independent variables. For the moment we will concern ourselves to the most popular measure of distance: Euclidean distance. The Euclidean distance between the points x and u is

$$d(\mathbf{x},\mathbf{u}) = \sqrt{\sum_{i=1}^{n}(x_i - u_i)^2}.$$

………..(4.1)

The simplest case is k = 1 where we find the sample in the training set that is closest (the nearest neighbor) to u and set v = y where y is the class of the nearest neighboring sample. It is a remarkable fact that this simple, intuitive idea of using a single nearest neighbor to classify samples can be very powerful when we have a large number of samples in our training set[12]. It is Possible to prove that if we have a large amount of data and used an arbitrarily sophisticated classification rule, we would be able to reduce the misclassification error at best to half that of the simple 1-NN rule. For k-NN we extend the idea of 1-NN as follows. Find the nearest k neighbors of u and then use a majority decision rule to classify the new sample. The advantage is that higher values of k provide smoothing that reduces the risk of over-fitting due to negative in the training data. In typical applications k is in units or tens rather than in hundreds or thousands. Notice that if k = n, the number of samples in the training data set, we are merely predicting the class that has the majority in the training data for all samples irrespective of u. This is clearly a case of over-smoothing unless there is no information at all in the independent variables about the dependent variable.[12]

3.2 Ant Colony Optimization

For the process of separation of class of candidate key for generation of association rule mining by KNN classification ,this classification whole class in two section ,in one section we classified only higher support vale and another section of class contain lower value of class. The process of searching of data according to given support of transaction table we used ant colony optimization for better searching of classified class and finally generated optimized rule. Here we discuss process of ant colony optimization. The ant colony optimization algorithm (ACO) is a heuristic algorithm for solving computational problems which can be reduced to finding good paths through graphs[18]. This algorithm is a member of ant colony algorithms family, in swarm intelligence methods. Initially proposed by Marco Dorigo in 1992 in his PhD thesis. It is the first algorithm was aiming to search for an optimal path in a graph; based on the behavior of ants seeking a path between their colony and a source of food. The original concept has since diversified to solve a wider class of numerical problems, and as a result, several different problems have emerged, bringing on several aspects of the behavior and impact of ants. Ant colony optimization algorithms are multi-agent systems, which consist of agents with the collective behavior of ants for finding shortest paths. Ant colony algorithms were inspired by the observation of real ant colonies[17]. Ants are social insects while insects that live in colonies and whose behavior is directed more to the survival of the colony as a whole than to that of a single individual component of the colony.

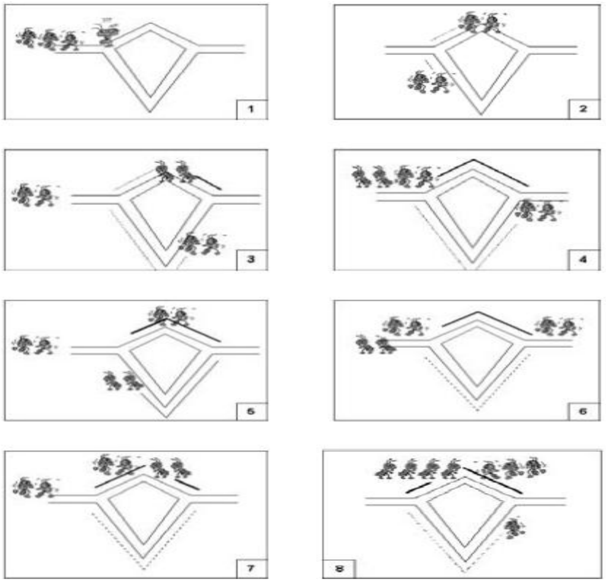

Figure 1 Simulation evolutions carried out by ants.

IV Proposed Method And Modification

We have proposed a novel algorithm for optimization of association rule mining, the proposed algorithm resolve the problem of negative rule generation and also optimized the process of superiority of rules. Superiority of association rule mining is a great challenge for large dataset. In the generation of superiority of rules association existing algorithm or method generate a series of negative rules, which generated rule affected a performance of association rule mining. In the process of rule generation various multi objective association rule mining algorithm are proposed but all these are not solve superiority problem of association rule mining.

In this paper we proposed distance weight optimizations of association rule mining with ant colony optimization. In this algorithm we used second odder quadratic equation and nearest neighbor classification technique for the selection of set of candidate of superiority of key for generation of rules. In the generation of rule selection of support value of transaction data set is play a important role , for this role we used heuristic search algorithm for better searching of support value for generation of optimized association rule.

We introduce a new feature sub set selection method for finding similarity matrix for rule mining without alteration of apriori rule mining. The proposed features sub set selection method based on ant colony optimization, ant colony optimization is very popular meta-heuristic function for searching for finding similarity of data. In this method we introduced continuity of ants for similar features and dissimilar features collect into next node. In that process ACO find optimal selection of features sub set. Imagine ants find features of similarity and equality in continuous root. Every ant of features compares their property value according to initial features set. When deciding data is negative and redundant should consider the two factors: importance degree and easiness degree of negative and redundant. While walking ants secrete phenomenon on the ground according to importance of the redundant and follow, in probability pheromone previously laid by other ants and the easiness degree of the negative.

Let D be data set and M be the number of ants ,importance degree a1,a2,……………………an is c1,c2,c3……………..cn, the appetency of solutions searched by two ants is defined as

App(I,j)=1 /C_i-C_j ………………………………………(1)

Where Ci and Cj is the importance of negative and redundant path. The concentration of the solution I defined as

Con(i+j)== $\frac{\delta i+\delta j}{m}$ ………………………………………(2)

Where $\delta i\ and\ \delta j$ is the number of ants whose appetency with other ants is bigger than α; α can be defined as m/10, then the incremented pheromone deposited by ants is

$_{\Delta}\tau i = Q.\beta i/Con(i+j)$…………………………(3)

Where Q is constantans .

Each level of pheromone constituted by means of a matrix $\tau\ where\ \tau ij(t)$ contains the level of pheromone deposited in the node I and j at time t ant k in node I will select the next node j to visit with probality

Pkij(t) $= \left\{\frac{\tau ij(tj)\alpha.(\mu ij)\beta}{\Sigma_{u\ \epsilon JK(i)}(\tau i)\alpha.\mu(i,j)\beta}\right\}\ if$

$j \in Jk(i)$…………………………………………..…(4)

otherwise 0;

Where $\mu ij\ reprents$ heuristic information about the problem which can defined as the easiness of the path.

at each iteration of the algorithm each ant ,using the pervious transition rule.

direct search in the best solution need global update rule applied as

$\tau ij(t+1) = (1-\rho).\tau ij(t) + \rho.\Delta\tau ij$............(5)

$\rho(0 < \rho \leq 1)$ represents a parameter which controls the pheromone evaporation .

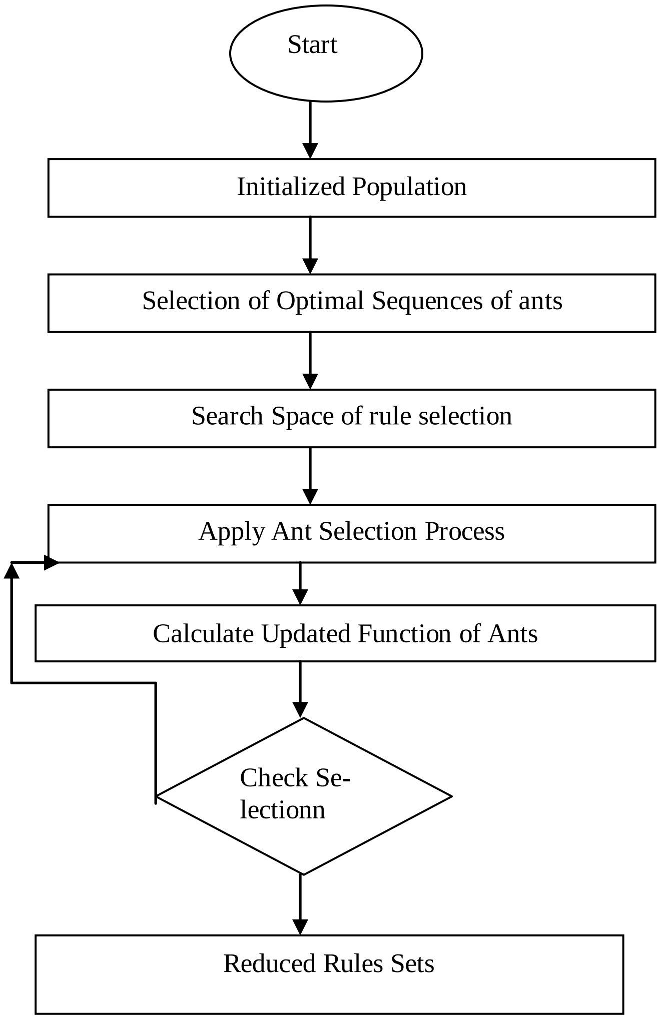
Start
Initialized Population
Selection of Optimal Sequences of ants
Search Space of rule selection
Apply Ant Selection Process
Calculate Updated Function of Ants
Check Selectionn
Reduced Rules Sets

Figure 2 :Flow Chart of Proposed DWOARM Using ANT Colony**:**

V Implementation and Results

To investigate the effectiveness of the proposed method implement in matlab 7.8.0 and testing of result we used wine data set, that data set provided by UCI machine laboratory. The experiment uses wine dataset obtained from UCI machine learning repository. The data set has 4177 samples [35]. It is composed of a discrete attribute and 13 continuous attribute. In this dissertation, we only mined such association rules X => Y that Y was contain of wine. The setting of parameters: The size of evolutionary population N=1000, crossover rate=0.006, mutation rate=0.001. The experiment was executed on Celeron(R) CPU 3.0GHz machine and software was MATLAB. MATLAB (matrix laboratory) is a numerical computing environment and fourth-generation programming language.

Attributes	A	B	C	D	E
Minimum Support	2	4	6	7	8
Minimum Confidence	.1	.2	.3	.4	.5
Execution Time	4.634764	4.683445	2.115712	2.132612	2.082319
No Of Rules	3262	3262	1932	1932	1932

Table 1:Comparisions of Result

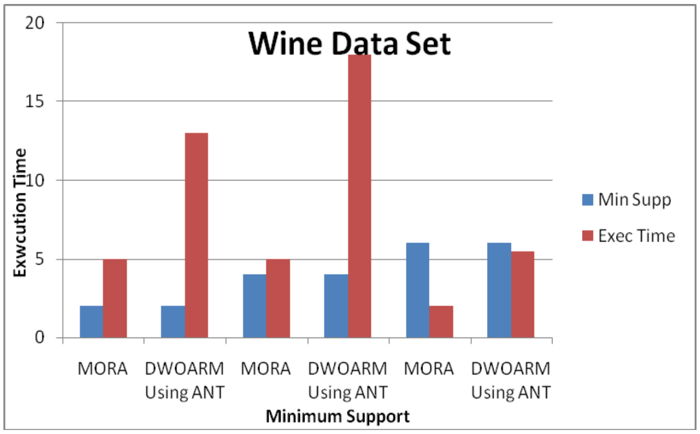

Figure 3 :Shows the comparative value of minimum support and execution time of MORA and DWOARM Using ANT algorithm for the processor extraction of rule.DWORAM using ANT takes more time in comparison of sample MORA algorithm.

V Conclusion

In this paper we have proposed a novel method for optimization of association rule mining. Our propped algorithm is combination of distance function and ant colony optimization.

We have observed that when we modify the distance weight new rules in large numbers are found. This implies that when weight is solely determined through support and confidence, there is a high chance of eliminating interesting rules. With more rules emerging it implies there should be a mechanism for managing their large numbers. The large generated rule is optimized with genetic algorithm.

We theoretically proofed a relation between locally large and globally large patterns that is used for local pruning at each site to reduce the searched candidates. We derived a locally large threshold using a globally set minimum recall threshold. Local pruning achieves a reduction in the number of searched candidates and this reduction has a proportional impact on the reduction of exchanged messages.

REFERENCES

[1] By Rakesh Agrawal Tomasz Imielinski Arun Swami Mining Association Rules between Sets of Items in Large Databases ACM SIGMOD Conference Washington DC, USA, May 1993.

[2] By Rakesh Agrawal Ramakrishnan Srikant_ Fast Algorithms for Mining Association RulesVLDB ConferenceSantiago, Chile, 1994.

[3] By Ramakrishnan Srikant* Rakesh Agrawal Mining Generalized Association Rules VLDB Conference Zurich, Swizerland, 1995.

[4] By Manjunath K.G, Kallinatha H.D. An Effective Indexing Method for High Dimensional Databases
(IJCSIT) Vol. 2 (5) , 2011, 2008-2018

[5] By Xiaofeng Yuan Hualong Xu , Shuhong Chen
Constrained association rule mining algorithm early stop algorithm large databases© IEEE

[6] By Q. C. Meng , T.J. Feng I , *2.* Chen I , C.J. Zhou , J.H. Bo2 Genetic Algorithms Encoding Study and A Sufficient Convergence Condition of GAS 0-7803-5731-0/9)sk$l0.00 0 IEEE 1999.

[7] By Pengfei Guo Xuezhi Wang Yingshi Han The Enhanced Genetic Algorithms
for the Optimization Design 978-1-4244-6498-2/10/$26.00 © IEEE 2010.

[8] By Masaya Yoshikawa and Hidekazu Terai A Hybrid Ant Colony Optimization Technique for Job-Shop Scheduling Problems Software Engineering Research, Management and Applications (SERA'06) 0-7695-2656-X/06 $20.00 © 2006.

[9] By Chi-Ren Shyu1,2, Matt Klaric1,2, Grant Scott1,2, and Wannapa Kay Mahamaneerat1 Knowledge Discovery by Mining Association Rules and Temporal-Spatial Information from Large-Scale Geospatial Image Databases 0-7803-9510-7/06/$20.00 © IEEE 2006.

[10] By LI Tong-yan, LI Xing-ming New Criterion for Mining Strong Association Rules in Unbalanced Events Intelligent Information Hiding and Multimedia Signal Processing 978-0-7695-3278-3/08 $25.00 © IEEE 2008.

[11] By Zhibo Chen, Carlos Ordonez, Kai Zhao Comparing Reliability of Association Rules and OLAP Statistical Tests Data Mining Workshops 978-0-7695-3503-6/08 $25.00 © IEEE 2008.

[12] By Lijuan Zhou Linshuang Wang Xuebin Ge Qian Shi A Clustering-Based KNN Improved Algorithm CLKNN for Text Classification Informatics in Control, Automation and Robotics 978-1-4244-5194-4/10/$26.00 ©IEEE 2010 .

[13] By XING Xue CHEN Yao WANG Yan-en Study on Mining Theories of Association Rules and Its Application Information Technology and Ocean Engineering 978-0-7695-3942-3/10 $26.00 © 2010 IEEE

[14] By Senduru Srinivasulu P.Sakthivel Extracting Spatial Semantics in Association Rules for Weather Forecasting Image 978-1-4244-9008-0/10/$26.00 © IEEE 2010.

[15] By TIAN He, XU Jing, LIAN Kunmei, ZHANG Ying Research on Strong-association Rule Based Web Application Vulnerability Detection 978-1-4244-4520-2/09/$25.00 © IEEE 2009.

[16] By Dieferson Luis Alves de Araujo' , Heitor S. Lopes', Alex A. Freitas2 A Parallel Genetic Algorithm for Rule Discovery in Large Databases 0-7803-5731-0/99$$10.00109 99 IEEE.

[17]Kuo R J and Shih C.W" Association rule mining through the ant colony system for national health insurance research database" in Taiwan. . In Journal of Computers and Mathematics with Applications, pp. 13 03-1318, 2007

[18]Atabaki G., Kangavari M. " Mining association rules in Distributed Environment through Ant Colony Optimization Algorithm, M.Sc thesis (in Persian), Iran University of Science and Technology, 2009.