A new method for the discovery of the best threshold value for finding positive or negative association rules using Binary Particle Swarm Optimization

Abdoljabbar Asadi[1],Azad Shojaei [2] ,Salar Saeidi[3] , Salah Karimi [4] and Ebad Karimi [5]

[1] University of Applied Science and Technology,Culture and Art Kurdistan Branch

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

[3] Department of Computer, Shareray Branch, Payam Noor University

[4] Department of Computer, Sannadaj Branch, Olom Tahghyghat University

[5] University of Applied Science and Technology, Jahad Daneshgahi Kurdistan Branch.

Abstract

In association rule mining most of former researches have worked on analytic optimizing method , but finding and specifying the advocate initiation limit influences on association rule mining's quality , which still is important hence this research wants to present a new algorithm for optimizing the analytic efficiency improvement including automatic analyze proper amount for initiation. Through former method this task had been performing based on positive rules but regarding that finding the negative ones were though for administrator, this research's privilege is that the initiation level automatically is analyzed for the first time; also it has high efficiency in large data base. Particle Swarm Optimization is observed for any particle's efficiency and as data turned in binary the advocate amount will be found. Results showed Particle Swarm Optimization could present better initiation level, and enhance the former algorithm's result a lot. Consequence will be comparing with Weka and Apriori..

Keywords: *data mining, association rule mining, minimal support, minimal confidence, particle swarm optimization*

1. Introduction

Regard to information technology development many databases have been developed to store the related data. Analyses of these databases to mine hidden rules have incremental importance [1]. One of the noticeable techniques help managers to make good decisions is data mining. This technique makes great tools available for users during current decay to extract meaningful information and useful patterns from databases [2], and this knowledge should be exact, readable and easy to understand [4]. But in spite of vast area of applications for data mining it still needs some manual operations, not automatic, to complete [4].

One of the important data mining techniques is association rule finding. It can extract hidden rules and dependent properties which have important role in decision making [1]. Apriorri Algorithm is the most famous one to extract association rule mining. But this algorithm has a major weakness which cannot calculate the minimal value of support and confidence and these parameters is estimating intuitively [4]. There are several algorithms to improve performance and accuracy of Apriorri. In traditional algorithms of association rule mining both of support and confidence parameters minimal value is chosen by the user try and error and this has an important effect on algorithm performance [1]. This approach can also produce many rules in a large database, millions, which probably many of them are not useful; it can be implied that it doesn't have enough efficiency. So we need a method to find best values of support and confidence parameters automatically specially in large databases. Most of surveys have pointed to positive association rules. While negative one is so prominent for data mining and also assessing them is quite tough task. It is possible that a negative rule seems as important as positive one. But recent algorithms are not capable to discover these rules [3].

Main goal of this paper is to presenting an optimal method to find suitable values of minimum threshold for support and confidence efficiently in positive or negative rules. This aim is achieved using particle swarm optimization (PSO) algorithm. PSO as an optimization method [5, 6] can be used for optimization of association rule mining [4]. A special type of PSO, named Binary PSO, is used for our work regard to its efficiency for local and large interval domains [7].

2. Literature Review

Our task is [5] complementary, in that method it has been used the PSO binary to threshold value through positive rules. Results of this article show that PSO binary bears high amount of stamina to discover associating rules. Weak point of 5 is that they have regarded the threshold value only for negative one. Data mining of both positive and negative in data base [8] has been stated. here some strategies of cutting and criteria about to evaluate the data bases have been revealed since taking decision about pragmatic cases like posing the product dissolving that and assessing and investing conations many factors necessity to minimize the detrimental rules us just for maximizing the rule. Hence associating data base of negative ones is matter of importance dramatically decision in [9] has been discussed .the author has used the hierarchical graph. This style effectively extracts the rules from data base.

Regarding to negative data mining importance it has been extracted for [10, 11, 3] only. Through [3] an overall method for casting all associating rules named NRGA has been delivered. this makes all clandestine rules relying on algorithm APRIORI .here to show the negative rules names such as CNR,ANR,ACNR.

Because of long runtime of Apriori algorithm to find association rules, its operational efficiency has a considerable importance. Several papers have presented different association rule mining algorithms to improve Apriori algorithm. Savasere et al. [12] developed Partition Algorithm for association rule discovery which is basically different form classic algorithm. This algorithm first scans the database to find strong item sets. Then support value is calculated for all item sets. Validity hint of Partition algorithm is that any strong item set appear in a section at least one time. Park et al. [13] introduced DHP at 1995. DHP is a derivative of Apriori plus some extra controls. It uses hash table to restrict candidates. DHP has two main properties: effective make of item sets and efficient reduction of database size by dropping adverse attributes.

Toivonen et al. [14] presented sampling algorithm at 1996. This algorithm is about finding association rules according to reduce database operations. DIC algorithm by Brin et al. at 1997 [14] splits database into some parts called start point. It determines support value for item sets belong to each start point and so extracts the patterns and rules. Bender search algorithm [16] by Lin et al. developed at 1998 can discover rules from most frequent itemsets. Yang et al. offered an efficient hash based method named HMFS which combines DHP and Bender search algorithms results in reduction of database scan and filtering repeated itemsets to find greatest repeated itemset [14]. This can shorten overall computation time for finding greatest repeated itemset.

Genetics algorithm has been applied to association rule mining during recent years. [15] Utilizes weighted items to distinguish unique itemsets. Value of different rules is determined using weighted items in fitness function. This algorithm can find suitable threshold value for association rule mining. Saggar [16] et al. presented a method for optimizing extracted rules, using genetics algorithm. The importance of the work is that it can predict rules with negative value.

Kuo et al. [1] in 2011 developed a PSO based method for automatic finding threshold value of minimal support. Their work shows that basic PSO can find values faster and better than genetics algorithm. Gupta [4] also offered a method at 2011 for automatic finding of threshold value using weighted PSO. His results show high efficiency of PSO for associative rule mining. This approach also can gain better values of threshold in comparison with previous ones.

3. Basics

3.1 Association rule mining

Agrawal et al. raised associative rule mining idea at 1993 [17]. A positive association rule presented as *if* $A \rightarrow B$ which A and B are subsets of *itemset(I)* and each itemset includes all of the items $\{i_1,i_2, \ldots,i_n\}$; It can be shown that in database $D= \{T1, T2, . . ., Tk\}$ a customer buys B product after buying A one if $A \cap B \neq \emptyset$. Association rule mining should be based on the following two parameters[17]:

1. Minimum support: finding item sets with the value above threshold
   $Support(A \rightarrow B)=P(A \cup B)= \frac{A \cup B}{D}$ (1)
2. Minimum Confidence: finding item sets with the value above threshold
   $Confidence(A \rightarrow B) =p(B|A)=\frac{A \cup B}{A}$(2)

Better rules have greater support and confidence value. Most famous algorithm for association rule mining is Apriori, offered by Agrawal et al. It repeatedly determines candidate itemsets using minimal support and confidence to filter itemsets for finding repeated ones with more frequency [1].

3.2 Particle Swarm Optimization Algorithm

PSO algorithm first developed at 1995 by James Kennedy, Russell C. Eberhart. It uses a simple mechanism inspiring from simultaneous motion of birds and fishes fly and their social life. This algorithm has successful applications recent years [6, 18]; mainly neural network weighting and control systems and everywhere that genetic algorithms can be use. PSO is not only a tool for optimization but also a tool for human social recognition representation. Some scientists believe that knowledge will optimize in effect of mutual social behaviors and thinking is not only a private action, indeed it is a social one. There are some entities in search space of the function which we are going to optimize it, namely particles [19]. PSO as an optimization algorithm provides a population based search which every particle change its position according to the time. Kendy in 1998 represented that each particle can be a possible answer that can move randomly in problem search space. Position change of each particle in search space is affected by experience and knowledge of itself and its neighbors [20].

Suppose we have a *d* dimension space and *i*'th particle from the swarm can be present with a velocity vector and position vector. Position change of each particle is possible by change in position structure and previous velocity. Position of each particle is x_i and it has information about best value which has reached yet, named *pbest*. This information is obtained from particles attempt to reach the best answer. Also any particle knows the best answer obtained for *pbest* from others in the swarm, named *gbest*. Each particle tries to change its position in order to reach the best solution using the following parameters:

x_i current situation, v_i the velocity, destination between the current position and *pbest*, destination between current position and *gbest*.

So the velocity of each particle changes as follows:

$$V_i^{k+1} = wv_i^k + c_1r_1 \,.\, (pbest_i - x_i^k) + c_2r_2 \,.\, (gbest - x_i^k) \quad (3)$$

Which V_i^k is the velocity of each particle in *k*'th repeat, *w* is the inertia weight, c1 and c2 are learning coefficients, r_1 and r_2 are random variables in the *[0,1)* interval with the unique distribution, x_i position of each particle *i* in the *k*'th repeat, $pbest_i$ which is *pbest* of *i*'th particle and *gbest which is gbest* of the group. Maximum of velocity (V_{max}) is to prevent velocity from increasing unlimitedly [21,22]. Position of each particle is determined as follows:

$$X_i^{k+1} = x_i^k + v_i^{k+1} \quad (4)$$

Equations 1 and 2 are form primitive version of PSO algorithm. PSO algorithm is so easy and has low computational, speed and memory load. It is using to solve continues problems while our work needs discrete version of the PSO.

The particles movement style towards the best one has been illustrated in Figure 1 [4]. One of the discrete versions is binary PSO which has developed by Kennedy and Eberhart at 1997 [6]. They did a small change on the algorithm to support discrete quantities also. Velocity is used as a probabilistic threshold value here and can be 0 or 1. $\boldsymbol{X_j^i}$, value of *j'th* bit from binary vector, shows the *i'th* particle position.

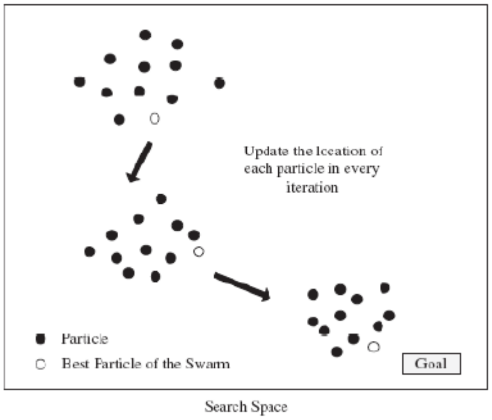

Update the lccation of each particle in every iteration
Particle
Best Particle of the Swarm
Goal
Search Space

Fig1.How the particles move towards the best particle

So the following describes Binary PSO function [7]:

$$X^i_j[t] = \begin{cases} 1 & ,\sigma < s(v[\mathrm{t}]) \\ 0 & ,otherwise \end{cases} \quad (5)$$

Which **σ** is a random number with the uniform distribution in [0,1] interval. **s(.)** is also the Sigmoid function described as follows:

$$S(z) = \frac{1}{1+\exp(-z)} \qquad (6)$$

Velocity change in Binary PSO is the same way as standard PSO.

4. Method

Process stages are according to Fig.1. Suggested algorithms include two sections: preprocess and assess. At first stages are collected and preprocess is assessed, then algorithm is used to discover association rules. Hardest section are collecting and preparing date.

Through all articles and researches which have been fulfilled in data mining, collecting data and pre assessing of that are of the most prominent parts, and also allocate the largest amount of time and expenditure. Within this article likewise these two stages have been revealed as the primary stage of algorithm. Throughout third stage we try to impose the proposed algorithm upon pre assessed data. In this stage, user first picks the assessment kind: positive or negative and then data will be read from the data base; flowingly the data binary equivalent will be revealed. In coming stage the binary PSO algorithm will be imposed in order to find positive or negative efficient association rule mining. At this level first particles primary population is constructed and any particles fitness is assessed then GBest will be chosen among the early population. Subsequently we provoke the articles in the space towards the most optimized particle .any particle stands for a rule.

At the end of algorithm the safety amount and support of t he best particle is revealed as the threshold and this could be found out by the other associating rules .here to find the threshold for positive rules it has been used the same method as in [5].

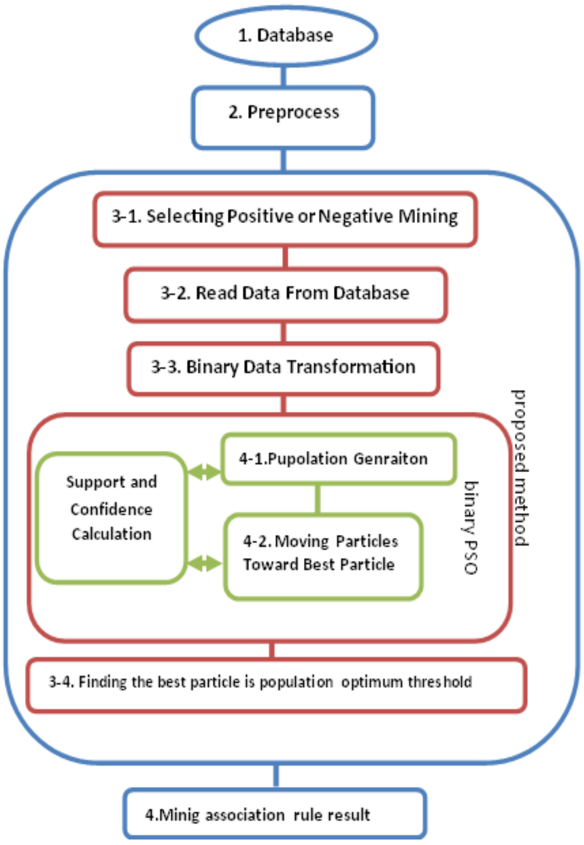

1. Database
2. Preprocess
3-1. Selecting Positive or Negative Mining
3-2. Read Data From Database
3-3. Binary Data Transformation
proposed method
4-1.Pupolation Genraiton
Support and Confidence Calculation
4-2. Moving Particles Toward Best Particle
binary PSO
3-4. Finding the best particle is population optimum threshold
4.Minig association rule result

Fig2. Steps of the proposed method

Primary substance used through data-mining is data. Hence a good data mining milestone are using and accessing primary data collecting and preparing is quite tough task [23]. In this research saved data in database Bank Marketing [24] have been used: this includes 452000 records of Bank Market study about costumers and each record bears ten features. We have used a 3500 record sample quite occasionally of this data base.

In suggested algorithm for concealed positive and negative rules it has been used binary PSO.

We have used optimal binary PSO to improve positive and negative rule production. Each particle represents a positive rule; consist of a predecessor and a successor. Figure 3 shows a particle; orange color is predecessor and blue one is successor. Every box represents a field from database. Containment of the boxes presents the value of a field in the database in the binary format.

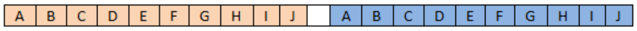


Fig3. Presentation of a particle

IF(AGE = Old AND Housing = yes) →(Marital = married, contact = telephone)

For example Fig.4 shows a rule with the following specifications:

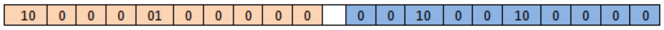


Fig4. Example of a particle in the database

Implementation of the proposed method has been done using *R2010b* version of MATLAB software. Movement representation of the particles toward the best goal is prepared form MATLAB also. Guiding a particle from the swarm population to an optimal answer is done by the fitness function. The particle with the greatest value of fitness usually supposed as the best particle [1] and [4]. In the proposed method *A* and *B* are collections of properties participating at predecessor and successor obtained from decoding respective particle according to what is explained. We calculate support and confidence values as follows:

In order to producing positive rule in the form of *if A* → *B*, two criteria form *cost(p)*function has been used to evaluate association rules quality.

$$Support = \frac{Supp(AUB)}{N} \qquad (7)$$

$$Confidence = \frac{Supp(AUB)}{Supp(A)} \qquad (8)$$

In which N is whole number of transactions and supp (A) is the number of Item A, B repetation through all of transactions. To create negative rules as if A → ~B two criteria have been used to ensure the quality of association rules mining..

$$Support = \frac{Supp(A) - Supp(A \cup B)}{N} \qquad (9)$$

$$Confidence = \frac{Supp(A) - Supp\ (AUB)}{Supp(A)} \qquad (10)$$

Likewise to create negative rules as if ~A→B two criteria at cost (P) to measure the association rules mining quality have been used.

$$Support = \frac{Supp(B) - Supp(AUB)}{N} \qquad (11)$$

$$\text{Confidence} = \frac{Supp(B) - Supp\ (AUB)}{N - Supp(A)} \qquad (12)$$

To create negative rule as if ~A → ~B two cost (P) to measure association rules mining quality are used.

$$\text{Support} = \frac{\text{N} - \text{Supp(A)} - \text{Supp(B)} + \text{Supp(AUB)}}{\text{N}} \qquad (13)$$

$$\text{Confidence} = \frac{\text{N}-\text{Supp(A)}-\text{Supp(B)}+\text{Supp(AUB)}}{\text{N}-\text{Supp(A)}} \quad (14)$$

After sending the particles to the fitness function, particle with the greatest fitness level will be used to move other particles toward the most optimal rule. Fitness function is defined as follows:

Fitness= α_1**Support +*α_2**Confidences –*α_3**NA*

Which *NA* is the number of properties used in the rule and coefficients, α_1, $\alpha 2$, $\alpha 3$, is used to parametric control of fitness function and customized by the user. First and second parts of this function is related to support and confidence values. It is essential to take into account both parts simultaneously. Because only one of support or confidence values cannot be a criteria for quality assessment of produced rules. It is evident that the more the value of both factors simultaneously the better the quality of the rule. We know that long rules will probably result to low quality productions also. So we try to produce relatively short, readable rules with more concept and quality which has special importance in data mining [4].

First *n* particles are creating quite randomly, each one representing a rule. Then fitness value of each one will be evaluated using the function noticed before. Binary PSO search algorithm will run until reaching the end condition; i.e. the best particle has founded and we can show the rules that support and confidence value of them are grater from minimal support and minimal confidence.

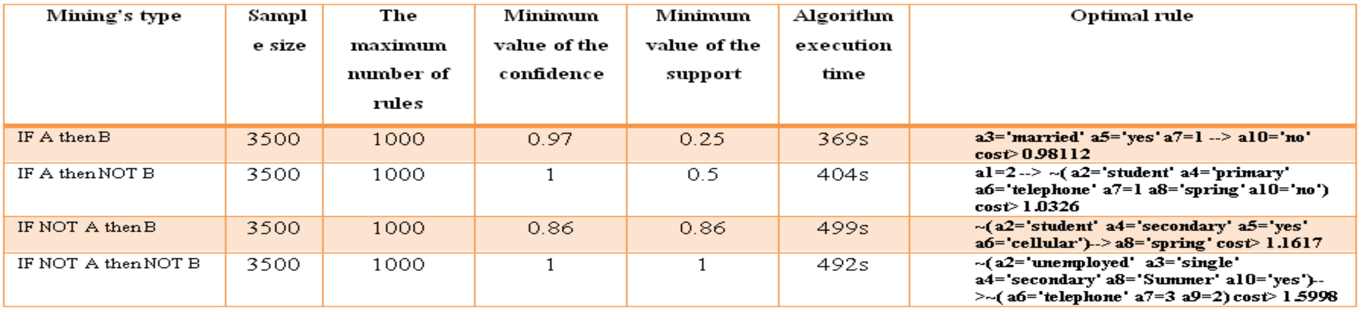
We used new method over many collections of data that the results were quite rewarding. Here for example result over a collection named Bank Marketing which bear 45200 records of transactions and each record includes 10 features. We have chosen a 3500 record sample quite occasionally. We set the new algorithm parameters as Table 1 [5]:

Table 1 –PSO Algorithm Parameters

Repeat Numbers	Learning rate of C_1,C_2	α_3	α_2	α_1
50	2	0.2	0.8	0.8

We evaluate all four association rules mining. The results are shown Table 2.

Table 2 - Results of the proposed algorithm



Mining's type	Sample size	The maximum number of rules	Minimum value of the confidence	Minimum value of the support	Algorithm execution time	Optimal rule
IF A then B	3500	1000	0.97	0.25	369s	a3='married' a5='yes' a7=1 --> a10='no' cost> 0.98112
IF A then NOT B	3500	1000	1	0.5	404s	a1=2 --> ~(a2='student' a4='primary' a6='telephone' a7=1 a8='spring' a10='no') cost> 1.0326
IF NOT A then B	3500	1000	0.86	0.86	499s	~(a2='student' a4='secondary' a5='yes' a6='cellular')--> a8='spring' cost> 1.1617
IF NOT A then NOT B	3500	1000	1	1	492s	~(a2='unemployed' a3='single' a4='secondary' a8='Summer' a10='yes')-->~(a6='telephone' a7=3 a9=2) cost> 1.5998

Regarding the results of this algorithm it is to say the negative data base bear safety and support better than positive ones. hence deans could access better rules and make more appropriate decisions .also this rules could be used in rule discovering assessments .particles movement towards the most optimized method has been shown.
Regarding to results it is possible relying on made algorithms of both positive and negative one be exposed with for much better consequences for example we use this in Weka which is of most important data mining software. positive data mining in that is according to Apriori. we regarded the early amount as zero for minimum amount of support and also we have utilized the algorithm APRIORI over the Apriori which due to memory limitation the algorithm could discover 40000 rules of those many are useless and picking the useful one is so tough act.

In continue we have utilized the positive moods to Apriori and Weka results of algorithm 2.consequently 164 useful rules with support amount of 25% and the security of 97have been created the results show that proposed algorithms enhance the associating rules efficiency.

5. Conclusion

In traditional algorithms of association rule mining both of support and confidence parameters minimal value is chosen by the user try and error and this has an important effect on algorithm performance. So we introduce a method to find best values of support and confidence parameters automatically specially in positive or negative association rule mining for the first time. Regarding the results of this algorithm it is to say the negative data base bear safety and support better than positive ones. We compared our results with Apriori in Weka. The results have shown that this new algorithm is efficiency. In the future works we can introduce a tow level algorithm for association rule mining.

References

[1] *R.J. Kuoa, C.M. Chaob and Y.T. Chiuc .Application of particle swarm optimization to association rule mining: Applied Soft Computing 11 (2011) pp:326–336*

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

[3] *Rupesh Dewang, Jitendra Agarwal. 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] *Manisha Gupta.Application of Weighted Particle Swarm Optimization in Association Rule Mining . International Journal of Computer Science and Informatics (2011) vol.1 pp:69-74.*

[5] Abdoljabbar Asadi, Mehdi Afzali ,Azad Shojaei, Sadegh Sulaimani. New Binary PSO based Method for finding best thresholds in association rule mining. Life Science Journal 2012;9(4). ISSN:1097-8135*.pp:260-264.*

[6] Kennedy, J., & Eberhart, R. C. A discrete binary version of the particle swarm algorithm. In *Proceedings of the conference on systems, man, and cybernetics*. (1997).pp: 4104–4109. Piscataway: IEEE.

[7] R. C. Eberhart and J. Kennedy. A new ptimizer using particle swarm theory. 6th Int. Symp. Micromachine Human Sci., Nagoya,Japan, 1995, pp. 39–43.

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

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

[10] W. Teng, M. Hsieh, and M. Chen.On the mining of substitution rules for statistically dependent items: ICDM(2002) pp:442-449.

[11] A. Savasere, E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of customer transactions:ICDE (1998) pp: 494-502.

[12] A. Savasere, E. Omiecinski, S. Navathe, An efficient algorithm for mining association rules in large database, in: Proceedings of the 21st VLDB onference,1995, pp. 432–444.

[13] Park, J. S., Chen, M., & Yu, P. An effective hash-based algorithm for mining association rules(1995). Pp: 175–186 .International Conference on Management of Data.

[14] H. Toivonen, Sampling large databases for association rules, in: Proceedings of the 22nd VLDB Conference, 1996, pp. 134–145.

[16] D.I. Lin, Z.M. Kedem, Pincer search: a new algorithm for discovering the maximum frequent set, in: Proceeding of the 6th International Conference on

[17] R. Agrawal, T. Imielin´ ski, A. Swami. Mining association rules between sets of items in large databases:ACM SIGMOD Record 22 (2) (1993) pp:207–216.

[18] j.kennedy and r.c.eberhart. particle swarm optimization : IEEE Int.Conf. Neural Netw. Perth, Australia(1995) vol. 4 pp: 1942-1948

[19] Riccardo poli , James Kennedy , Tim Blackwell .Praticle swarm optimization An overview :Springer Science.swarm intell(2007) pp:33-57

[20] Kennedy, J. The behavior of particles:porto,v.w,Saravanan,N.,Waagen.D.,andEiben,A.E(eds.) ,In:Evolutionary Programming VII,Springer (1998) pp:581-590

[21] Y. Shi , R. Eberhart. Parameter selection in particle swarm optimiza-tion: 7th Int. Conf. Evol. Program., NCS (1998) vol. 1447 pp: 591–600.

[22] R. Eberhart , Y. Shi.Comparing inertia weights and constrictionfactors in particle swarm optimization: IEEE Congr. Evol.Comput (2000) pp: 84–88.

[23] Philippe Lenca, Patrick Meyer, Bonoit vaillant, Stephae lallich. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid : European Journal of operation research (2008)184 610 – 626.

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

Abdoljabbar Asadi got his B.Sc. degree in Computer Engineering form Shahid Bahonar University in 2008, Recently he got his M.Sc degree of Software Engineering from the Islamic Azad University. He is lecturer and researcher at Information Technology and Computer Engineering Department of University of Applied Science and Technology,Culture and Art Kurdistan Branch.

Azad Shojaei is researcher in Department of Computer, Saghez Branch, Islamic Azad University.

Salar Saeidi is student in Department of Computer, Shareray Branch, Payam Noor**.**

Salah Karimi is student in Department of Computer, Sannadaj Branch, Olom Tahghyghat University.

Ebad Karimi got his B.Sc. degree in Computer Engineering form Shahid Bahonar University in 2008, He is lecturer and researcher at Information Technology and Computer Engineering Department of University of Applied Science and Technology,Jahad Daneshgahi Kurdistan Branch.