Revealing a Novel Method for Detecting Positive and Negative Optimal Performance Association Rules in Very Large Databases Using BPSO

Salah Karimi Haji pamagh[1], Dr. Mehdi Afzali[2], 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

Association rules mining is one of the useful data mining algorithm in presenting meaningful information through database. One of the important challenges for association rules mining is that; it might be extracted millions of rules which mostly are idol, furthermore current methods only seek positive rules which finding out the negative ones is more prominent. In this method we mingle data mining and Evolutionary algorithms including association rules, Particle Swarm Optimization, whose goal is discovering the pattern and positive, and negative rules and also optimum one through large database, also this algorithm is capable to present scarce rules, which might be neglected by administrator. Consequences by recent algorithms could help administrator in making many resolutions.

Also this algorithms result has been compared with Apriori. The results indicate the algorithms efficiency. Collecting and preparing data in this survey has been performed in SQl server 2008 and algorithm performed in MATLAB software.

***Keywords*:** Data mining, Association rules mining, Evolutionary algorithms, large databases, Particle Swarm Optimization.

1. Introduction

Improving technology in information domain, several database for preserving data have emerged; hence analysis these databases seem important [1]. One of the techniques which help administrators is Data Mining. This has helped users to extract meaningful information and useful rules from databases [2]. Acquired knowledge is pretty significant and this should be precise, legible and comprehensible [3].

One of the important algorithms is Association rules mining, which could find dependent rules which are important in making resolutions [1]. Most of researches have implied the positive association rules mining but negative ones are important as well [3]. Observing negative rules might not be as important as positive ones, but current algorithms could not find these rules. Moreover traditional association rules mining algorithms like Apriori [1] could produce many rules, which mostly are useless and abeles these algorithms efficiency. Hence optimizing rules could be plenty significant [4,5]. Discovering an efficient system is quite prominent to administrators. In this research we reveal a positive and negative association rules mining. Binary Particles swarm optimization [6, 7, 8] algorithm of one of the optimizing methods which could be used for association rules mining.[4] In current finding association rules mining like Apriori only those which are if A→B would be found. Using minimum confidence and support, they filter Itemset, to find Itemset with high frequency. But in this article in addition to positive rules, negative ones are found as if ~A→~B ; if A→~B ; if ~A→B. This article also capable to find association rules mining in gigantic databases, whose outcomes could be so useful for administrators. A special type of PSO, named Binary PSO, is used for our work regard to its efficiency for local and large interval domains [9].

2. Literature Review

Rummaging the association rules mining both positive and negative has been expressed in [10]. Also some strategies for correcting and criteria to evaluate the database are stated. Since making resolution about pragmatic issues including posing product, decomposition, analyzing and investing bear many factors, it is necessary to abate the

detrimental rules the same time in order for increasing the benefit; hence rummaging association rules mining is quite crucial for making decisions.

Assessing negative rules has been discussed through [11]. Author has utilized hierarchical graph method. This approach effectively extracts rules from database. Regarding to remarkable importance of assessing negative rules, some researches [3, 12 and 13] have been analyzing these rules. Through [3] a new method named NRGA has been revealed to construct association rules. NRGA makes all hidden rules relying on algorithm Apriori. Through this article to show negative rules, names such as ACNR, ANR and CNR have been used. Also correlation equation has been reformed hence all resulted rules are hopeful.

Through [14] it has been used quantum swarm to association rules mining this algorithm extract the best rules but not always achieve the rules quite surely. This article's result shows that quantum swarm algorithm causes better results than genetic ones one. Through [15], swarm intelligence techniques have been used to rummage rules through pharmaceutical database. Swarm intelligence has been added to rules discovering, in a way that its mobility is capable through rules rummaging. A traditional method to assess rules creates many rules, which has made pharmaceutical un-useful. Using the reveal approach it is possible to find the optimum patterns.

2. Basics

2.1 Association rule mining

Agrawal et al. raised associative rule mining idea at 1993 [16]. 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 $D = \{T1, T2, \ldots, Tk\}$ 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:

1. Minimum support: finding item sets with the value above threshold

$$Support\ (A \rightarrow B) = P(A \cup B) = \frac{A \cup B}{D} \quad (1)$$

2. Minimum Confidence: finding item sets with the value above threshold

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

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].

2.2 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 [6,7]; 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 [17,18]. 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 [19,20].

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:

$$V_i^{k+1} = wv_i^k + c_1 r_1 . (pbest_i - x_i^k) + c_2 r_2 . (gbest - x_i^k) \quad (3)$$

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 [9,21-22]. 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. One of the discrete versions is binary PSO which has developed by Kennedy and Eberhart at 1997 [7]. 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. x_j^i, value of *j'th* bit from binary vector, shows the *i'th* particle position. So the following describes Binary PSO function [17]:

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

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

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

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

3. 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.

Fig1. Steps of the proposed method

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 .if particles secure and support scale is larger than threshold that particle will be shown as a rule .eventually the best particle s best algorithm will be introduced as scarce rule.

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.[18] 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. 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 2 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.

Fig2. Presentation of a particle

For example Fig. 3 shows a rule with the following specifications:

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

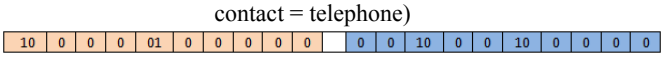

Fig3. Example of a particle in the database

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 [3]. 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* $A \rightarrow B$, two criteria form *cost(p)*function has been used to evaluate association rules quality.

$$Support = \frac{Supp(A) - Supp(A \cup B)}{N} \quad (7)$$

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

In which N is whole number of transactions and Supp(A) is the number of Item A, B repetition 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
.

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

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

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

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

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

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

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

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

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 [3].

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.

4. Results and Discussion

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 set the new algorithm parameters as follows:

Table 1 –PSO Algorithm Parameters

Repeat Numbers	Learning rate of C_1,C_2	α_3	α_2	α_1	Minimal support	Minimal confidence
7	2	0.2	0.8	0.8	0.04	0.4

Through association rules mining the repetitions and low quality rules are eliminated and the rest are known as ultimate ones, and the ultimate optimized were introduced as scarce ones. We evaluate all four association rules mining. The results are shown Table 2.

Regarding to results it is to say that the mood (if ~A→~B) reveals higher security and supporting rules. Then the mood (if ~A → B) reveals better rules the mood (A → ~B) likewise better rules than (if A→B). The important result is that through this database the amount of support and security of negative rules are quite better than positive ones, and this can help deans to make decisions. As the table above shows this algorithm wastes a long time to association rules mining.
As it was mentioned the algorithm Apriori might extracted many rules from data base which mostly are useless. For

Table 2 - Results of the proposed algorithm

Mining type	Sample size	number of rules	Number of rules generated	Average confidence level	Average amount of support	The algorithm execution time in seconds	Number of rules obtained by Apriori
IF A then B	45200	7000	58	0.43	0.098	11784	11799
IF A then NOT B	45200	7000	1200	0.94	0.15	12021	The ability to discover no negative rules
IF NOT A then B	45200	7000	683	0.58	0.57	16188	The ability to discover no negative rules
IF NOT A then NOT B	45200	7000	6913	0.89	0.87	19270	The ability to discover no negative rules

instance in diagram no 2 of algorithm Apriori, 11798 rules were extracted that makes the useful rules assess hard to administrators. Also one of the weak points of this is that it is not able to discover the negative rules whereas performing positive and negative associating rule mining using PSO binary, upon data base, the made negative rules are much better and simultaneously are more useful for administrators.

Through Fig.4-7 the approach of particle movement to best rules has been shown. Through these graphs, vertical axis stand for fitness function and the horizontal one stands for number of repeating the particle movements. Respect to amount of fitness function it is to say in this database, the mood (if ~A → ~B) creates the most optimize and also the most efficient rule, and the created positive rules have less efficiency.

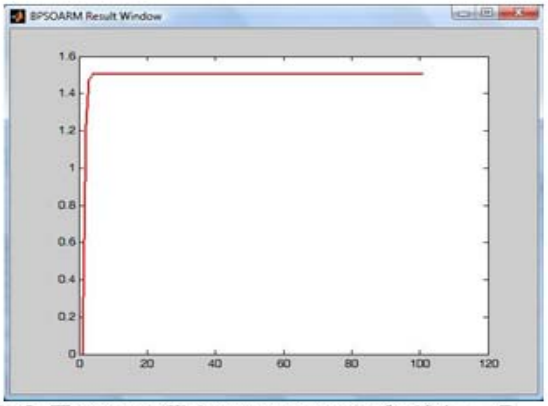

Fig 5: The most efficient way to move the if A → ~B

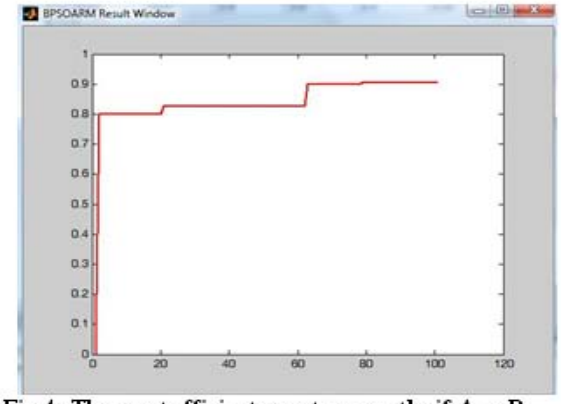

Fig 4: The most efficient way to move the if A → B

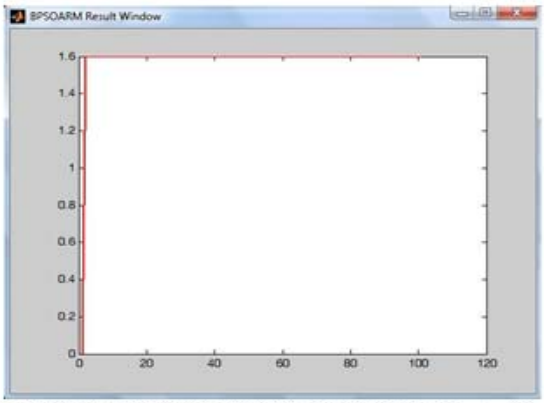

Fig 7: The most efficient way to move the if ~A → ~B

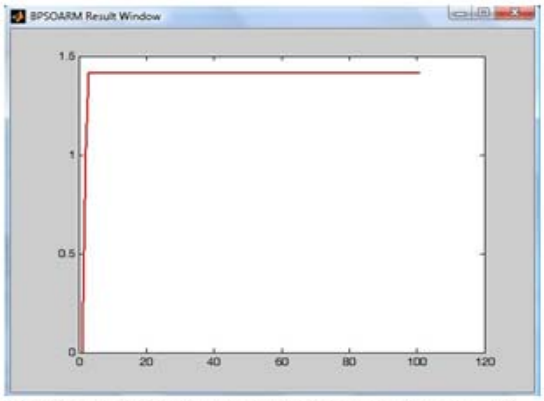

Fig 6: The most efficient way to move the if ~A → B

5. Conclusion

Here by combining Data mining and Evolutionary algorithms including PSO and association rules a new method is revealed. This method was used through a database to extract association rules mining, whose results show its high capability through discovering positive and negative association rules mining. The strong point of revealed method is to find scarce and high quality rules and also relying on Evolutionary algorithms to extract positive and negative rules. Results showed in some databases some negative rules are better and more efficient and make the administrators more successful. The scarce rules which are found by this algorithm could help deans to take the best decisions. Analysis shows that this algorithm is much better than the traditional associating data mining one. Weak point of the algorithm is the law speed of performing that we hope improve it and using faster method we become able to enhance this algorithm's efficiency. Essential problem through such survey is data and inaccessibility of data.

6. 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 Rule: International Journal on Computer Science and Engineering (2011) Vol. 3 pp: 1649-1657.

[4] Maragatham G, Lakshmi M. A RECENT REVIEW ON ASSOCIATION RULE MINING: Indian Journal of Computer Science and Engineering (2012) Vol. 2 pp:831-836.

[5] RUPALI HALDULAKAR, JITENDRA AGRAWAL. Optimization of Association Rule Mining through Genetic Algorithm: International Journal on Computer Science and Engineering Vol. 3 (2011) pp:1252-1259.

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

[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] Crina Grosan, Ajith Abraham, Monica Chis. *Swarm Intelligence in Data Mining:* SCI (2006)pp:1-20.

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

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

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

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

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

[14] Mourad Ykhlef.A Quantum Swarm Evolutionary Algorithm for mining association rules in large databases: Journal of King Saud University – Computer and Information Sciences (2011) pp: 1–6.

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

[16] 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.

[17] Riccardo poli , James Kennedy, Tim Blackwell .Praticle swarm optimization An overview :Springer Science.swarm intell(2007)1:33-57.

[18] 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.

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

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

[22] 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.

[23] X. Dong, S. Wang, H. Song, and Y. Lu.Study on Negative Association Rules: Transactions of Beijing Institute of Technology (2004) Vol.24 pp:978-981.

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