Asian Journal of Information Technology 13 (10): 582-587, 2014
ISSN: 1682-3915
© Medwell Journals, 2014

Weighted Quantum Particle Swarm Optimization (WQPSO) and PSO Algorithm to Association Rule Mining and Clustering

A.D. Gokil and S. Rajalakshmi
Department of Computer Science and Engineering,
Jay Shriram Group of Institutions, Tirupur, Tamilnadu, India

Abstract: In the area of Association Rule Mining (ARM) the most major algorithms is Apriori algorithm. In the existing Apriori algorithm minimum support and confidence are determined subjectively or through trial and error method, so the algorithm lacks the objectiveness and efficiency. To improve the efficiency of association rules, Particle Swarm Optimization (PSO) algorithm is projected which gives feasible threshold values for minimum support and confidence. In the PSO algorithm, initially it looks for the optimum fitness value of each particle and then finds their corresponding support and confidence as minimum threshold values. The difficulty of PSO algorithm is that it guesses that the items have the same implication without taking into account of their weight/attributes within a transaction or within the whole item space. To overcome this drawback, this study proposes a Weighted Quantum Particle Swarm Optimization algorithm (WQPSO) with weighted mean best position according to fitness values of the particles. WQPSO algorithm provides faster local convergence, fallout in better balance between the global and local searching of the algorithm, so it generates good performance. The proposed WQPSO algorithm is experienced with several benchmark functions and compared with standard PSO. The experimental result shows the supremacy of WQPSO and it is verified by applying the FoodMart2000 database of Microsoft SQL Server 2000. Likewise, in clustering, there are many unsupervised clustering algorithms have been developed one such algorithm is K-Means which is simple and straightforward. The main drawback of the K-Means algorithm is that the result is sensitive to the selection of the initial cluster centroids and may converge to the local optima. This is solved by PSO as it performs globalized search and produces clusters with high intra class similarity.

Key words: Data mining, association rule mining, particle swarm optimization, K-Means, weighted quantum particle swarm optimization, clustering

INTRODUCTION

With the growth of information technology, there are various kinds of information databases such as scientific data, trading data, financial data and marketing contract data. To effectively analyze and apply these data and find the significant hidden information from these databases have become very important issues. Data mining technique (Han and Kamber, 2000) has been the most broadly discussed and frequently applied tool in current decades. It can be considered into a number of models, including association rules, clustering and classification. Among these representation, association rule mining is the most broadly applied method. The Apriori algorithm is one of the most representative algorithms. It consists of many modified algorithms that focus on improving its efficiency and accuracy. On the other hand, two parameters namely minimal support and confidence are always find out by the decision-maker him/herself or through Trial and Error Method, so, the algorithm lacks both objectiveness and efficiency. Therefore, the main reason of this study is to suggest a PSO (Song and Gu, 2004) algorithm which provides feasible threshold values for minimal support and confidence. But the PSO algorithm guesses that items have the same implication without taking into account of their weight/attributes within a transaction or within the whole item room. For that reason, this study suggest a WQPSO (Xi *et al*., 2008) algorithm with weighted mean best position according to fitness values of the particles. The proposed WQPSO algorithm is experienced with several benchmark functions and evaluated with standard PSO. For the point of assessment, this study first employs the embedded database of Microsoft SQL Server 2000 to assess the proposed algorithm. In another side, the K-Means algorithm is a well-known approach to

Corresponding Author: A.D. Gokil, Department of Computer Science and Engineering, Jay Shriram Group of Institutions, Tirupur, Tamilnadu, India

clustering. Its popularity depends on its simplicity and computational efficiency. However, that approach tends to fixate on local optima near the initial cluster centers which are assigned randomly. So, this study explores the applicability of PSO and its variants to cluster data vectors. In the progression of doing so, the aim of the study is:

- To propose WQPSO algorithm for association rule mining
- To compare the performance of PSO with WQPSO
- To show that the standard PSO algorithm can be used to cluster arbitrary data
- To compare the performance of PSO and its variants with standard K-Means algorithm

LITERATURE REVIEW

This study briefly illustrate about the active algorithm in association rule mining. Agrawal *et al*. (1993) and Han and Kamber (2000) introduced the Apriori algorithm, it is helpful to catch the frequent itemsets from a transaction dataset and obtain association rules. Once frequent itemsets are obtained then it is easy to generate association rules with confidence larger than or equal to a user specified minimum confidence. It says, if an itemset is not frequent, any of its superset is not at all frequent. Although, in some cases with a lot of frequent itemsets, huge itemsets, or very low minimum support, till now it suffers from the cost of generating a huge number of candidate sets. In view of the fact that the processing of the Apriori algorithm requires loads of time, its computational efficiency is a very important issue.

Savasere *et al*. (1995) proposed the Partition algorithm. The algorithm carried out in two phases. First phase, the Partition algorithm rationally divides the database into a number of non-overlapping partitions. The partitions are measured one at a time and all large itemsets for that partition are generated. At the end of phase I, these large itemsets are combined to generate a set of all potential large itemsets. Phase II, the actual support for these itemsets is shaped and the large itemsets are recognized. The partition sizes are selected such that each partition can be accommodated in the main memory so that the partitions are read only once in each phase. An important, role of the approach is that it drastically reduces the I/O overhead associated with previous algorithms. But the problem is, additional work is needed to accurately estimate the number of partitions.

Toivonen (1996) proposed the sampling algorithm. It relates the level-wise method to the sample, along with a lesser minimum support threshold, to quarry the superset of a large itemset. A quite clear way of reducing the database activity of knowledge discovery is to use only a random sample of the relation and to find approximate regularities. Samples are little enough to be hold totally in main memory can provide reasonably accurate results. This method creates exact association rules but in some cases it does not create all the association rules.

Genetic Algorithm (GA) (Pei *et al*., 1998) was developed by Holland. Genetic algorithm is stochastic search algorithm modeled on the process of natural selection which underlines biological progress. By using Genetic Algorithm (GAs) the system can predict the rules which contain negative attributes in the generated rules along with more than one attribute in consequent part. The major benefit of using GAs in the discovery of prediction rules is that they perform global search and its complexity is less compared to other algorithms. But, it have some drawbacks, they are GAs are very slow, in presence of noise, convergence is hard and the local optimization technique might be useless and models with many parameters are computationally expensive.

Ant Colony Optimization (ACO) was introduced by Patel *et al*. (2011) and Dorigo has evolved significantly. ACO algorithm is a Meta heuristic stimulated by the foraging behavior of ant colonies. ACO algorithm is useful for the specific problem of minimizing the number of association rules. Drawbacks of the ant colony optimization are theoretical analysis is difficult and Probability distribution of ACO changes by iteration.

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The process follows a straightforward and easy way to classify a given data set through a definite number of clusters (assume k clusters) fixed a priori. The main thought is to illustrate k centroids, one for every cluster. The centroids should be located in a cunning way because of different location causes diverse result. The main disadvantage of the K-Means algorithm is that the result is sensitive to the selection of the initial cluster centroids and may converge to the local optima.

ASSOCIATION RULE MINING

Particle Swarm Optimization (PSO): Kennedy and Eberhart (Kuo *et al*., 2011) projected the Particle Swarm Optimization (PSO) algorithm. The main idea of PSO is created from the study of fauna behavior. It imitates the behaviors of bird flocking. Consider the following: a group of birds are arbitrarily searching for food in a region. There is only one piece of food in the region being searched. No one knows where the food is. However, the birds do know how faraway the food is during all iterations. The most useful strategy is to chase the bird which is nearest to the food. PSO learned from such a situation and used to solve the optimization problems.

In PSO, each single solution is a "bird" or "particle" in the search region. All particles have fitness values which are assessed by the fitness function. Particles fly through the problem space by following the current best particles. PSO is initialized with a group of random particles (solutions) and then look for the optima by updating generations. During each iteration, each particle is updated by following the two "best" values. The first is the best solution (fitness) it has achieved so far. This value is called "pbest". The other "best" value is tracked by the particle swarm optimizer is the best value obtained so far by any particle in the populace. This best value is a global best and is called "gbest". After finding the two best values each particle updates its equivalent speed and location with Eq. 1 and 2 as:

$$v^{new}_{id} = v^{old}_{id}+c_1\ rand()\ (pbest-x_{id})+c_2\ rand()\ (gbest-x_{id}) \quad (1)$$

$$x^{new}_{id} = x^{old}_{id}+v^{new}_{id} \quad (2)$$

Where:

v_{id} = The particle velocity of the idth particle
x_{id} = The idth or current particle
i = The particle's number
d = The dimension of searching space
rand( ) = A random number in (0, 1)
c_1 = The individual factor
c_2 = The societal factor

Usually, c_1 and c_2 are set to be 2.

Particle Swarm Optimization (PSO) algorithm: The proposed algorithm comprises two parts, preprocessing and mining (Kuo *et al*., 2011). The first part provides procedures related to calculating the fitness values of the particle group.

The data are distorted and stored in a binary format. Then, the exploration range of the particle swarm is set using IR (itemset range) value. In the next part of the algorithm, the PSO algorithm is working to mine the association rules. In that, first starts with particle swarm encoding, it is similar to chromosome encoding of Genetic algorithms. Then, generate a population of particle swarms according to the calculated fitness value. The PSO searching procedure proceeds until the stopping condition is reached which implies the best particle is found. The support and confidence of the best particle can indicate the minimum support and minimum confidence. After that apply these threshold values for association rule generation. Figure 1 demonstrates the algorithm structure.

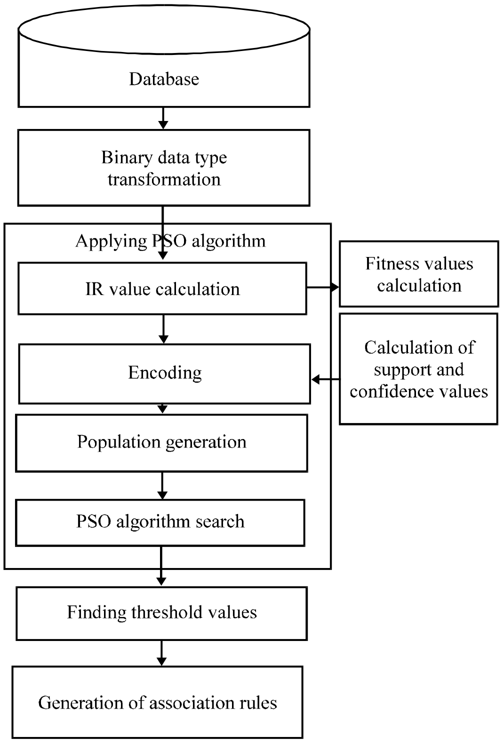
Database
Binary data type transformation
Applying PSO algorithm
IR value calculation
Encoding
Population generation
PSO algorithm search
Fitness values calculation
Calculation of support and confidence values
Finding threshold values
Generation of association rules

Fig. 1: PSO association rule mining algorithm

Methodology preprocessing: Preprocessing is a significant step for a successful data mining approach which spots the missing values for the given input. In some cases there may be a possibility of incomplete data, inconsistent data or noisy data in the data set. In this approach binary transformation technique is used as the preprocessing technique. This approach can speed up the database scanning operation and it is helpful to calculate support and confidence values more easily and quickly.

IR value calculation: Here, the IR value is calculated for the data received from the preprocessing technique. In order to increase the search efficiency IR analysis is used to choose the rule length generated by chromosomes in particle swarm progress. IR analysis avoids searching for large number of association rules which have no meaning in the process of particle swarm progression. This method deal with the front and back partition points of each and every chromosome and the range determined by these two points is called the IR which is exposed as follows:

$$IR = [\log (mTransNum(m))+ \log(nTransNum(n))]\left(\frac{Trans(m, n)}{TotalTrans}\right) \quad (3)$$

In the Eq. 3, the "m" and "n" value should posses the condition that m≠n and m<n. "m" denote the length of the itemset and TransNum(m) denotes the number of transaction records containing m products. "n" is the length of the itemset and TransNum (n) refers the number of transaction records having n products. Trans(m, n) refers to the number of transaction records obtaining m to n products. Total Trans represents the total number of transactions.

Applying PSO algorithm: The Particle Swarm Optimization algorithm has been proposed to calculate the threshold values. The algorithmic process is quite similar to that of genetic algorithms but the proposed procedures include only encoding, fitness value calculation, population generation, best particle search and termination condition. The process of generating the association rules using the PSO is done in the following ways. Initially, the encoding on the obtained itemsets has been performed. Then, the fitness values for the encoded itemsets are obtained. The equation for fitness value calculation is given below:

$$Fitness\ (k) = confidence\ (k) \times \log(support(k) \times length\ (k)+1) \quad (4)$$

The particle with the highest fitness value is taken and their support and confidence can represent the minimum support and minimum confidence and then use that threshold values for association rule mining.

Association rule mining: In this step, Association Rule Mining approach has been presented using the Apriori algorithm. The Apriori algorithm is an influential algorithm to generate the association rules. It is a two step process. First, frequent itemsets are generated by means of the join and prune step. Second, based on the frequent itemsets, the association rules are mined.

Weighted Quantum Particle Swarm Optimization (WQPSO): WQPSO works much like PSO but the difference is, it assigns some weightage to the particle which is having better fitness value. In PSO (Mishra and Omkar, 2011; Xi *et al*., 2008), each fitness value is given equal weightage. But, if researchers look at it from the social point of view, the elite members, i.e., the ones with higher fitness values are the major contributors to the development of the swarm's quality. Therefore, if researchers consider a case in which the better fitness values are given more weightage than the lower fitness values, it should lead to even more efficient convergence. So, the mean best position is replaced by a weighted mean best position keeping everything else the same. The first step in its implementation would be to determine whether a member is an elitist or not. It is carried out by evaluating the fitness's of the members. Hence, higher weightage shall be allotted to the member with a higher fitness value and lower to a lower fitness value. This shall take care of the convergence improvement. For that, first rank the members of the swarm in a descending order and then allocate the weightage also in decreasing order.

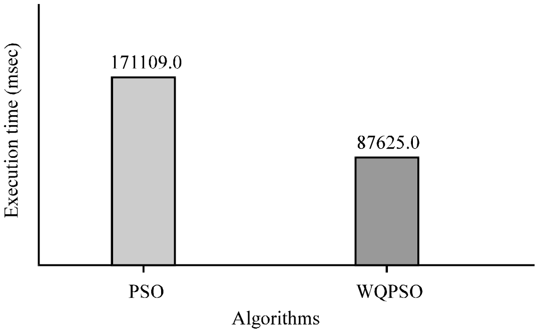

Fig. 2: Comparison of PSO and WQPSO

Results of ARM: PSO algorithm not only gives us a slight edge on the accuracy of the result it also reduces the computational cost. In the WQPSO procedure, the elite members are given more weightage compared to the ordinary members of the swarm. While, PSO gives us quicker and accurate results, WQPSO promises us even more accurate results but with a slightly higher computational cost.

In Fig. 2, the results obtained by WQPSO are compared with PSO, it shows WQPSO yields fairly quicker result with quicker convergence, i.e., WQPSO is successful in having better global search capability in comparison to PSO and hence a better optimal result is obtained using WQPSO. It is verified by applying the FoodMart2000 database of Microsoft SQL Server 2000.

CLUSTERING

Cluster analysis is a technology which can classify the similar sample points into the same group from a data set. The clustering aims at identifying and extracting significant groups in underlying data. In the field of clustering, K-Means algorithm is the most popularly used algorithm to find a partition that minimizes Mean Square Error (MSE) measure. Although, K-Means is an extensively useful clustering algorithm, it suffers from several drawbacks. The main drawback of the K-Means algorithm is that the result is sensitive to the selection of the initial cluster centroids and may converge to the local optima. This is solved by PSO (Satapathy *et al.*, 2009) as it performs globalized search.

PSO in clustering:

1. Initialize each particle to randomly selected cluster centroids
2. For each data vector Z_p
3. Calculate the Euclidean distance $d(Z_p,C_{i,j})$ to all cluster centroids C_{ij}:

$$d(Z_p, C_{ij}) = \sqrt{\sum (Z_p - C_{ij})^2} \quad (5)$$

4. Assign Z_p to cluster C_{ij} such that $d(Z_p,C_{i,j})$ should be minimumn
5. Evaluate the fitness function for each particle:

$$\text{Fitness calculation} = \frac{\sum_{j=0}^{N_c} \frac{d(Z_p, C_{ij})}{|C_{ij}|}}{N_c} \quad (6)$$

where, $|C_{ij}|$ is the number of data vectors belonging to cluster C_{ij} and N_c denotes the number of cluster centroids, i.e., the number of clusters to be formed

6. Compare every particle's fitness value with previous particle's best solution (pbest). If current solution is better than previous value (pbest) then update pbest with current solution
7. Compare fitness evaluation with the population's overall previous best. If current value is better than the gbest (the global version of the best value), then reset gbest to the current particle's value and position
8. The particle with gbest value is taken and their cluster is the best cluster with maximum intra cluster similarity
9. Repeat Step 2-6 until the predefined number of iterations is completed

Results of clustering: Manning *et al.* (2008) study shows that the aim of clustering is to attain high intra-cluster similarity and low inter-cluster similarity. To begin consider there are two clustering results:

C: The correct vector of bit masks = {c1, c2, c3, ..., cn}
K: The vector of bit mask results of some algorithm = {k1, k2, k3, ..., km}

Then, create a "matching matrix", M = $[a_{ij}]$. The matching matrix is just the number of cells that from result C are in cluster i and from result K are in cluster j. So, i goes from 1→n and j goes from 1→m. N is then the total number of cells. The F measure is then a measure of the algorithms precision and recall (Fig. 3-5).

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{Precision+recall}} \quad (7)$$

Where:

$$\text{Precision (P)} = \frac{\text{Cells correctly put into a cluster}}{\text{Total cells put into the cluster}} \quad (8)$$

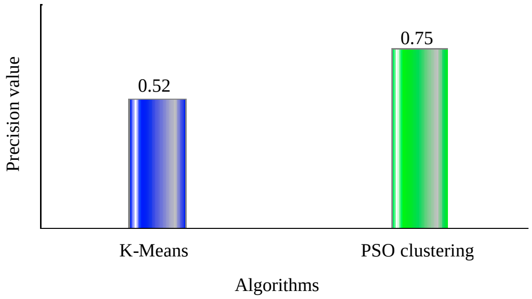

Fig. 3: Performance comparison based on precision value

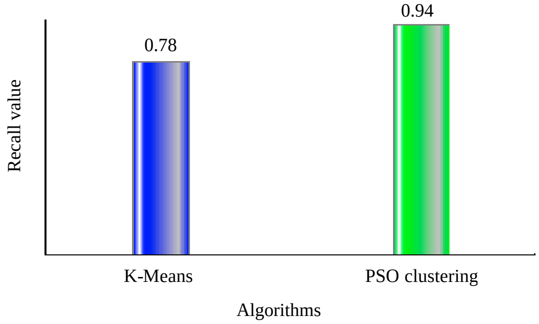

Fig. 4: Performance comparison based on recall value

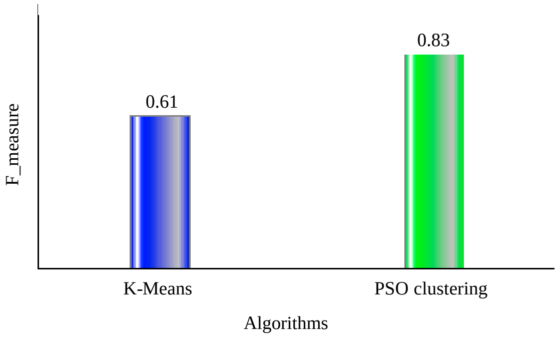

Fig. 5: Performance comparison based on F_measure

$$\text{Recall (R)} = \frac{\text{Cells correctly put into a cluster}}{\text{All the cells that should have been in the cluster}} \quad (9)$$

In form of equation:

$$P(c_i, k_j) = \frac{a_{ij}}{|k_j|}$$

which is the number of cells in that were in both cluster i and j (from correct answer C and clustering result K, respectively) divided by the number of cells that are in cluster j:

$$R(c_i, k_j) = \frac{a_{ij}}{|c_i|}$$

which is the number of cells in that were in both cluster i and j divided by the number of cells that are in cluster i (in this case the correct number of cells). So then:

$$F(c_i, k_j) = \frac{2\times R(c_i, k_j)\times P(c_i, k_j)}{R(c_i, k_j)+P(c_i, k_j)} \qquad (10)$$

This is the F score for the comparison of one cluster to another. The above experimental results show that PSO clustering performs better than the K-Means algorithm. Breast cancer dataset from UCI machine learning repository is used to calculate the precision, recall and F_Measure value.

CONCLUSION

An important research that takes place in the area of data mining is the process of extracting the required information based on the query. Thus, the effective information can be retrieved based on the efficient association rules. By focusing on the problem of the generating the association rules, in this study an effective approach of weighted quantum particle swarm optimization approach has been proposed. Here, the proposed approach has been evaluated with the existing concept of association rule mining. Experimental result shows that this approach provides an efficient association rule mining application for the information searching from the large databases. This approach can be further enhanced by applying the other association rule techniques that can outperform the proposed approach. Clustering is also a basis of many knowledge discovery tasks such as machine learning, statistics, data mining and pattern recognition. The well-known K-Means algorithm, suffers from several drawbacks due to its choice of initializations. In order to overcome K-Means shortcomings, PSO can be considered as a choice. PSO performs global search and K-Means is responsible for local search. The process of the proposed algorithm is such that the strength and ability of preventing from being trapped in local optimums is improved. Computational experiments show that the proposed algorithm of this study is effective, robust, easy to tune and tolerably efficient as compared with other approaches. To improve the obtained results of the proposed algorithm, it can increase local search ability around the best found position by the algorithm.

REFERENCES

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 25-28, 1993, Washington, DC., USA., pp: 207-216.

Han, J. and M. Kamber, 2000. Data Mining: Concepts and Techniques. Morgan-Kaufman Publishers, New York.

Kuo, R.J., C.M. Chao and Y.T. Chiu, 2011. Application of particle swarm optimization to association rule mining. Applied Soft Comput., 11: 326-336.

Manning, C.D., P. Raghavan and H. Schutze, 2008. An Introduction to Information Retrieval. Cambridge University Press, USA., ISBN-13: 9780521865715, Pages: 482.

Mishra, A. and S.N. Omkar, 2011. Singularity analysis and comparitive study of six degree of freedom stewart platform as a robotic arm by heuristic algorithms and simulated annealing. Int. J. Eng. Sci. Technol., 3: 644-659.

Patel, B., K.V. Chaudhari, R.K. Karan and Y.K. Rana, 2011. Optimization of association rule mining apriori algorithm using ACO. Int. J. Soft Comput. Eng., 1: 24-26.

Pei, M., E.D. Goodman and W.F. Punch, 1998. Feature extraction using genetic algorithms. Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning, Volume 98, December 16-19, 2007, Birmingham, UK, pp: 371-384.

Satapathy, S.C., G. Pradhan, S. Pattnaik, J.V.R. Prasad and P.V.G.D. Prasad Reddy, 2009. Performance comparisons of PSO based clustering. Int. Comput. Sci. Networking, 1: 18-23.

Savasere, A., E.R. Omiecinski and S.B. Navathe, 1995. An efficient algorithm for mining association rules in large database. Technical Report, https://smartech.gatech.edu/handle/1853/6678.

Song, M.P. and G.H., Gu, 2004. Research on particle swarm optimization: A review. Proceedings of 2004 International Conference on Machine Learning and Cybernetics, August 26-29, 2004, Shanghai, China, pp: 2236-2241.

Toivonen, H., 1996. Sampling large databases for association rules. Proceedings of the 22th International Conference on Very Large Databases, Septempber 3-6, 1996, Bombay, India, pp: 134-145.

Xi, M., J. Sun and W. Xu, 2008. An improved quantum-behaved particle swarm optimization algorithm with weighted mean best position. Applied Math. Comput., 205: 751-759.