International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 16 (2017) pp. 5520-5524


# Efficient Frequent Pattern Mining using Particle Swarm Optimization

**Satish Muppidi [1], M. Ramakrishna Murty [2] and Suresh.Ch[3]**

*[1]Department of Information Technology, GMR Institute of Technology, Rajam-532127, Andhra Pradesh, India.*

*Orcid Id: 0000-0003-1714-1769*

*[2]Department of Information Technology, ANITS, Visakhapatnam, AP, India.*

*Orcid Id: 0000-0002-0124-9153*

*[3]Department of Information Technology, ANITS, Visakhapatnam, AP, India.*

*Orcid Id: 0000-0002-0861-8994*

**Abstract**

Association rule mining is one of the vital data mining tasks to extract knowledge from the data. In the process of association rule mining the foremost step is to find the frequent itemset. The frequent itemset is used to generate association rules. In general brute –force approach is expensive because there are exponentially many rules that can be generated from the data set. So that support count is the point to determining the frequency of occurrence for every candidate itemset that survives the candidate pruning of the rule generation. In general, support count chooses randomly by the user, and this random choosing of support count may not extract better association rules every time. In this work, applied particle swam optimization technique to choose appropriate support count, in order to extract efficient association rules.

**Keywords:** Frequent itemset, support count, PSO, association rule mining.

## INTRODUCTION

Data in business firms are increasing drastically due to the automation and digitalization in the recent past. Most of the organizations accumulate the large data and apply knowledge extraction techniques on them to get new business rules. In particular, to analyze retail customer's buying habits through the market basket analysis. Such valued statistics can be used to support diversity of business related applications such as customer relationship management, market promotions, and inventory management.

Association rule mining is one of the data mining techniques which is appropriate and suitable for the above said applications. Association rule mining is also called as frequent pattern mining was first proposed by Agrawal [1]. Frequent patterns are item sets, subsequences or substructures that appear in a data set with some frequency.

For example in the buying process of a customer's transaction database has a set of items such as pen, pencil and book appear frequently together in a transaction data set is called a frequent itemset [2][3]. In Frequent pattern mining analysis the customer buying habits by finding association between the different items those customers brought from the market place. For example as mentioned earlier if customers are buying pencil, how likely they are also buying notebook or may be any other item related to the pencil. This type of analysis can help retailers to increase the sales as well as production department.

The mathematical terms of representation of the frequent itemset is discussed below [4][5][6][21]. For example, let us take the data items in the super market data I={i1,i2,i3,….id} and performed transactions on the data as T={t1,t2,t3,……tN}. Each transaction ti is having a subset of items purchased from the items I. In the association analysis, a collection of 0 or more items is called as itemset. If the item set contains k-items, it is called as k-itemset. For example {pen, book, pencil} is an example of a 3-itemset [7][8]. The itemset does not contain any items hence it is called as null or empty itemset.

Transaction width is defined as the number of items presented in a transaction. In general, an itemset X is a subset of any of the corresponding transaction [9][10][11]. The important property of an itemset is a support count which refers to the number of transactions that contain a particular itemset.

In terms of mathematical representation of the support count

$$\sigma(X) = |\{ti\ /X \in ti, ti \in T\}| \quad (1)$$

Where the symbol |.| indicates the number of elements in a set.

The process of association rule mining decomposes into two major sub tasks.

- Frequent itemset generation
- Rule generation.

In the first subtask of frequent itemset generation the computational requirements are generally more expensive than the step of rule generation [12][13].

In general brute –force approach is expensive because there are exponentially many rules that can be generated from a data set. In explicit many rules may generate from the data set that contains d items based on the following equation

$$R = 3^d - 2^{d+1} + 1 \quad \text{-- (2)}$$

According to the above equation, in spite of small data set, it generates more number of rules which are not strong.

If data set has k-items, the data generates up to 2k-1 frequent itemsets, excluding null set [14]. In the real world data, the value of k is obviously very large and the search space of itemsets that need to be discovered is also exponentially large [15]. In the process of itemset generation, if itemset is frequent, all of its subsets must also be frequent. Therefore, to reduce inefficient frequent items with using support count [16][17]. The support count strategy trims the exponential search space based on support measure. The support of an itemset never exceeds the support for its subsets.

In the process of frequent itemset generation, it is generally used support count to reduce the search space too. The support count is being chosen by the user random value.

Support count is an important property for association rule mining. It refers to the number of transactions which contains a particular item set.

This paper is organized with four sections. In section 2 problem identification and basic idea of the problem is discussed. Explanation with the steps of the algorithm of PSO based in section 3 and dataset along with results in section 4.

## PROBLEM IDENTIFICATION

Generate association rules are considered as NP–hard problem as it needs to search a search space of 2n where n denotes number of items. The real world data has been increasing drastically, which has been drastically increasing search space. The exponential growth of search space to find the optimal solution is a challenging task. Evolutionary methods help to find the better optimal solutions even the search space increases drastically. Genetic algorithm (GA) can help to find the better solutions but computationally expensive according the studies. PSO is used in this work for choosing the optimal support count with the help of heuristic search techniques. The basic initiation towards improving the performance of association rule mining algorithms is to use the support and confidence requirements as object functions.

## PARTICLE SWARM OPTIMIZATION

Particle Swarm Optimization is a global optimization search algorithm presented by Kennedy and Eberhart. It is an evolutionary computation technique introduced through simulation of collective social behavior such as bird flocking and fish schooling. In PSO, particles represent candidate solutions in a solution space, and the optimal solution is found through moving the particles in the solution space. Each candidate solution is called particle and represents one individual of a population called swarm [10][11]. A particle changes their components and flies through the multi-dimensional search space. Particles compare themselves to their neighbors and intimate the best of those neighbors.

PSO is made ready with a population of random solutions (particles) using identical distribution. However, each particle in PSO traces a trajectory in an n dimensional search space, updating constantly a velocity vector based on best solutions found so far by that particle as well as others in the population (swarm).

$$V_i(t+1) = w\, v_i(t) + C_1\, \varphi_1\big(\text{pbest}_i\,(t) - x_i(t)\big) + C_2 \varphi_2 (\text{gbest}(t) - x_i(t)) \quad \text{-- (3)}$$

Here pbesti(t) is personally best experience the best value of the fitness function found by the i-th particle up to time t. gbest(t) is a global best experience, the best value out of pbesti(t) values of all particles in the swarm found swarm up to time t. w is the weighting factor, $\varphi_1, \varphi_2$ are uniform distributed random numbers that determines the influence of pbesti(t) and gbesti(t). C1 is particle's self –confidence, it controls the contribution towards the self-exploration. C2 is swarm confidence; control the contribution towards the global direction.

```
Input: Randomly initiated position and velocity of the particle: Xi(0) and Vi(0)
Begin:
    Reach up to termination condition (max number of iterations)
        Begin
            For i=1 to number of particles
                Calculate fitness f(xi)
                Update pbesti and gbesti
                Adapt velocity of the particle
                Update position of the particle
            Increment i value
```

**Figure 1:** Particle Swarm Optimization procedure

The position updates the ith particle based on the pbest and gbest values with the following equation.

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad \text{--(4)}$$

In the equation (3) the variable w is responsible for dynamically adjusting the velocity of the particles, and it is responsible for balancing between local and global search hence it requires fewer iterations for the algorithm to converge[16][17][18]. A low value of inertia weight implies a local search, while a high value leads to a global search.

The weighting factor w<1 only little momentum is preserved from the previous time-step [12]. The weighting factor w=0 the particle moves in each step totally ignores information about the past velocity. If w>1 particle can handle change their direction which implies a reluctance against convergence towards optimum. The w value is greater than 1 always used with Vmax to avoid swarm explosion. Vmax can be set to the full search range of the particle‘s position in order to allow global search.

Fitness function: In the PSO each particle is evaluated itself and is being move to the next position compared to its neighbors based on the fitness function. The fitness function is generated for the association rule mining based on the support count.

$$\text{fitness(i)} = \alpha 1 \left[\frac{\text{Sup(AUC)}}{\text{Sup(A)}}\right].\left[\frac{\text{Sup(AUC)}}{\text{Sup(C)}}\right].\left[1 - \frac{\text{Sup(AUC)}}{|D|}\right] + \alpha 2 \left[\frac{\text{Numberfields(i)}}{\text{MaxField}}\right] \quad \text{-(5)}$$

The goal of generating fitness function is to discover the association rules which are more interesting but those item occurrence is less comparatively other frequent itemset in the database. Therefore interestingness measure is used in the fitness function to find hidden information and it is used to generate better rules.

In the equation (5) α1and α2 are interestingness parameters may increase and decrease by the user as per parameters of the fitness function. |D| is the total number of records in the database.

In the equation (5) $\left[\frac{\text{Sup(AUC)}}{\text{Sup(A)}}\right]$ part indicates probability of creating the rule based on the antecedent and $\left[\frac{\text{Sup(AUC)}}{\text{Sup(C)}}\right]$ part indicates probability of creating the rule based on consequent.

Most of the interesting rules, the rate of acquired information is nearly same in both antecedent and consequent parts of the rule. $\left[1 - \frac{\text{Sup(AUC)}}{|D|}\right]$ , this part of the equation gives the probability of generating the rule depends upon the entire data set. Another parameter in the equation (5) is Numberfields(i) indicates number of attributes exists in the ith particle and maxfield indicates the maximum length of the rule.

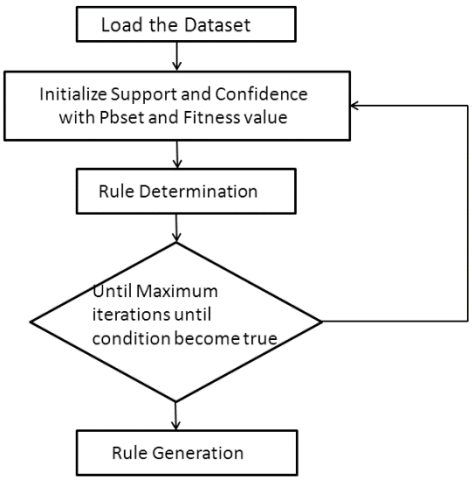


**Figure 2:** PSO based algorithm working process

## EXPERIMENTAL RESULTS

The experiments were simulated in Mat lab tool. The initial velocity was zero set to all the populations and learning factor values set to 2. Maximum number of iterations is 50. Experiments are conducted on the following bench mark data sets.

*Data set 1:*

BMSWebView1: This dataset contains 59,601 sequences of clickstream data from an e-commerce. It contains 497 distinct items. The average length of sequences is 2.42 items with a standard deviation of 3.22. In this dataset, there are some long sequences. For example, 318 sequences contain more than 20 items.

*Date set 2:*

FIFA: a dataset of 20,450 sequences of click stream data from the website of FIFA World Cup 98. It has 2,990 distinct items (webpages). The average sequence length is 34.74 items with a standard deviation of 24.08 items. This dataset was created by processing part of the web logs from the world cup offered

*Data set 3:*

Chess: This data set is prepared based on movements on the chess board. Source of this data set is from UCI machine repository.

The BMS web view1 database contains several months’ worth of clicking streams data from two e-commerce web sites. Each transaction in this data set consists of all the product detail pages seen in that session. That is, each product detail

view is an item. The goal of this datasets is to find associations between products viewed by visitors in a single visit to the web site. The FIFA dataset was created by processing part of the web logs from the world cup. The dataset details are also given below in the Table 1.

**Table 1:** Dataset Information

| Data set Name | Distinct Items | Avg. Length of sequence |
|---|---|---|
| BMSWebView1 | 497 | 2.42 |
| FIFA | 387 | 34.74 |
| Chess | 654 | 2.12 |

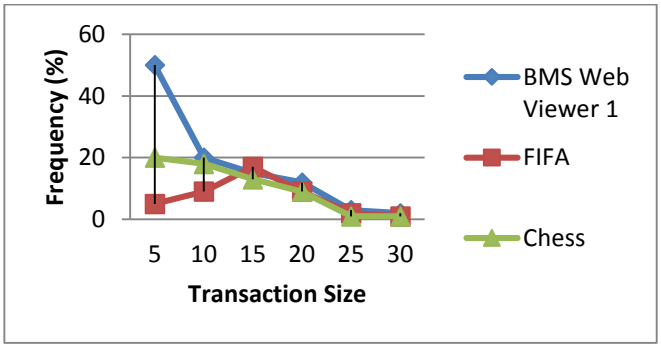


**Figure 3:** Data set size distribution

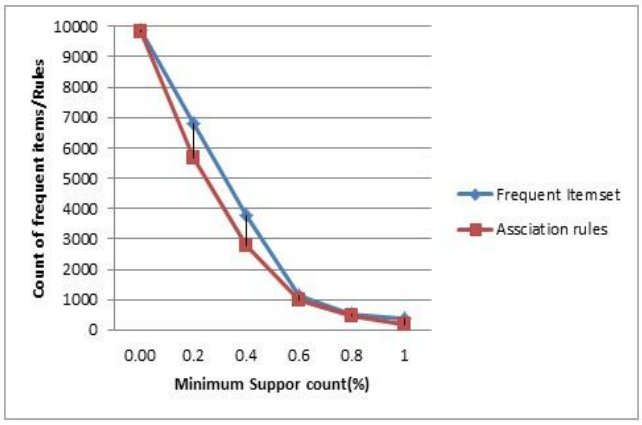


**Figure 4:** Frequent items/rules generation based on the support count threshold

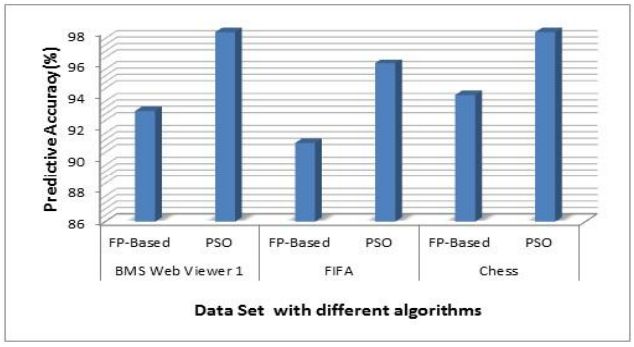


**Figure 5:** Predictive Accuracy of FP-Based and PSO algorithm

Firstly in the experimentation of PSO based association rule mining method specifies initial seeds here in this work initial seeds of dataset with different minimum support count values to be given and specify the number of iterations as input. Figure3 shows the data set distribution curve for different data set against to frequency. PSO based association algorithm works more effective here because there are different frequent items exits in the data set with different support count values. Experiments showed that the rules are generated based on the support count values choosing by the user while running the algorithm. According to results figure 4 should support counts increases the number of frequent items are reduces. Predictive accuracy is better for the PSO based association rule based algorithm against to other traditional association rule methods.

## CONCLUSION

Association rule mining is one of the important knowledge extraction techniques. The main objective of this paper is to generate efficient and effective association rules with the help of the evolutionary based technique called Particle Swarm Optimization. In the process of association rule generation minimum support count is one of the important parameters. Generally support count is choosing randomly, it may lead to escape some important frequent items from the data set. In this work applied particle swarm optimization to resolve the above mentioned issue.

## REFERENCES

[1] Jiawei Han, Hong Cheng, Dong Xin, Xifeng Yan "Frequent pattern mining: current status and future directions" Data mining and knowledge discovery, springer, vol 15, pp 55-86, 2007.

[2] Leandro dos Santos Coelho , Cezar Augusto Sierakowski "A software tool for teaching of particle swarm optimization fundamentals " advances in engineering software, Elsevier, pp 877-887, Vol 39, 2008.

[3] Chai.Chunlai. Li.Biwei. "A novel association rules method based on Genetic algorithm and Fuzzy Set strategy for web mining ". Journal of computers. Vol. 5, Issue 9, 2010.

[4] Eberhert.C.Russel. Shi.Yuhui. " Particle swarm optimization: developments, applications and resources" , IEEE, 0-7803-6657-3/01. 2001.

[5] R. Eberhart and Y. Shi, "Comparing Inertia Weights and Constriction Factors in Particle Swarm Optimization", Proc. of the Congress on Evolutionary Computation (CEC2000), pp.84-88, 2000.

[6] E. Ozcan and C. Mohan, "Analysis of a Simple

Particle Swarm Optimization System", Intelligent Engineering Systems through Artificial Neural Networks, Vol.8, pp. 253-258, 1998.

[7] M.Ramakrishna Murty, et al " Data Clustering using hybridization of Particle Swarm Optimization &Teaching Learning Based Optimization Techniques" International Journal of Applied Engineering Research, ISSN:0973-4562, Vol.10 ,No.81, pp 240-245, 2015.

[8] R. Agrawal and R. Srikant, 1995. Mining sequential patterns. Proceedings of the 11th International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan, pp: 3-14.

[9] R. Agrawal, T. Imielienski, and A. Swami. 1993. Mining Association Rules between Sets of Items in Large Databases. Proc. Conf. on Management of Data, 207–216. ACM Press, New York, NY, USA 1993.

[10] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000, pp. 1–12.

[11] Wang J, Han J BIDE: Efficient mining of frequent closed sequences. In: Proceeding of the 2004 international conference on data engineering (ICDE'04), Boston, MA, pp 79–90, 2004.

[12] Pei J, Liu J,Wang H,Wang K,Yu PS,Yang J Efficiently mining frequent closed partial orders.In: Proceeding of the 2005 international conference on data mining (ICDM'05), Houston, TX, pp 753–756, 2005.

[13] Balazs Racz,Ferenc Bodon,Lars Schmidt-Thieme., On Benchmarking Frequent Itemset Mining Algorithms, from Measurement to Analysis, 37-42, Chicago, Illinois, USA, august 2005.

[14] Webb, G.I. Efficient search for association rules. In Proceedings of the Sixth ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, NewYork, NY: ACM, 99-107.

[15] Zhan, Z.H., Zhang, J., Li, Y., and Chung. ―Adaptive particle swarm optimization. IEEE Transactions on Systems Man, and Cybernetics — Part B: Cybernetics‖, Vol.39 (6). pp. 1362-1381. ISSN 0018- 9472, 2009.

[16] Syed Khairuzzaman Tanbeer, Chowdhury Farhan Ahmed, Byeong-Soo Jeong , Young-Koo Lee "Efficient single-pass frequent pattern mining using a prefix-tree", Information Sciences, Elsevier 179 (2008) 559–583.

[17] K.Lahari, M. Ramakrishna Murty "Partition based clustering using Genetic Algorithms and Teaching Learning Based Optimization: Performance Analysis" International Conference and published the proceedings in AISC, Springer DOI: 10.1007/978-3-319-13731-5_22, Vol 2, pp 191-200.

[18] D.W. Cheung, S.D. Lee, B. Kao, A general incremental technique for maintaining discovered association rules, in: Proceedings of the Fifth International Conference on Database Systems for Advanced Applications, 1997, pp. 185–194.

[19] W. Cheung, O.R. Za, Incremental mining of frequent patterns without candidate generation or support constraint, in: Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS), 2003.

[20] E. Chen, H. Cao, Q. Li, T. Qian, Efficient strategies for tough aggregate constraint-based sequential pattern mining, Information Sciences 178 (2008) 1498–1518.

[21] 21.M Ramakrishna Murty, JVR Murthy, Prasada Reddy PVGD, Suresh Satapathy, "Statistical approach based keyword extraction aid dimensionality reduction", International conference Information Systems Design and Intelligent Application-2012, Springer-AISC (indexed by SCOPUS, ISI proceeding DBLP etc), ISBN 978-3-642-27443-5, Vol:132, 2012, PP: 445-454."