The Significance of an Optimized Data Mining Technique – An Illustration

G. Maragatham*, Rajendran S
Department of Information Technology, SRM University, KTR, Chennai, India
***Corresponding author: E-Mail: maragatham.g@ktr.srmuniv.ac.in**

ABSTRACT

Association rule mining techniques have become most popular among the researchers in data mining community. One of the well-known technique is Apriori algorithm. Analyzing traditional frequent pattern mining algorithms reveals some drawbacks. That is, as the time expires certain rules mined from the database become obsolete. Next, while considering a retail dataset, the customers buying interest changes based on the purchase cost. These aspects must be considered as a data analyst for designing a better solution. To meet these two challenges in the mining process, the technique called UTARM. Utility based Temporal Rule Mining is used. To obtain an optimized results, a Weighted PSO is used. An example illustrating the combined approach is discussed to understand the importance of the approach in the data mining area. Hope, this article gives a better insight on using the technique with respect to the retail databases for better decision making.

KEY WORDS: Apriori, Association Rule Mining, UTARM - Utility based Temporal Association rule mining, Particle Swam Optimization.

1. INTRODUCTION

Data mining is basically defined as the process of finding out interesting, hidden patterns from the underlying databases. Lots of effort have been carried out on this aspect for finding out these hidden patterns. The mining of such hidden patterns has become most prevalent pattern discovery technique in KDD which was stated by Anil Rajput (2009), Han and Fu (1996), Han (2011). Out of several data mining tasks such as Association Rules, Sequential patterns, Classification etc., the data mining task- ARM gains its attention nowadays. The importance of knowledge discovery in databases is also addressed by researchers Fayyad UM (1996). For improving the mining process, the significance of utility miner was insisted by Jyothi Pillai (2011). Incorporating evolutionary techniques such particle swam optimization is attempted by Her-Shing Wang (2009) and Mourad Ykhlef (2011). Generally, Association rules are commonly produced in two steps. First, all the frequent itemsets in the database is found by adapting the FTU - minimum support, Frequent Temporal Utility (FTU). Secondly, from the frequent item sets generated, the rules are produced adapting to the FTU_confidence value. Simplicity and absence of model based assumptions are the main advantages of association rules. Generally, publication-like databases cannot be mined by the traditional approaches. First, as the time expires some of the discovered rules may become obsolete and secondly - in regard to the sales database, the customers buying interest may get changed on the purchase cost. With the aim of adapting these two challenges UTARM, Maragatham (2015) is proposed. For efficient decision making process, the rules generated must be reduced. Therefore, Optimization Technique called Weighted Particle swarm optimization. Maragatham (2011) is used. This method uses Fitness function which is based on the FTU_Support and FTU_confidence of UTARM technique. The integrated approach of UTARM with PSO is discussed in this article by considering a simple example to give a complete idea on the working mechanism. The Fitness function used by the authors effectively prunes out the unwanted rules and the final rule set is an Optimized rule set. This is done mainly to have an effective decision making process with respect to the application such as retail market business.

2. MATERIALS AND DISCUSSION

The approach uses two techniques (a) UTARM Method for Rule generation and (b) Weighted Particle Swarm Optimization Technique for optimizing the rules generated from the UTARM Technique. Initially, the database is partitioned based on the time granularity. The time granularity can be month, week or year. We have considered the time granularity as month. Based on the partitioned database the mining process. First the UTARM technique is used, the inputs required are the database, minimum support and the minimum confidence. For each partition separate utility table is supplied, which gives the weightage of each items. Then, the mining process starts by considering the consecutive partitions. For all the rules generated FTU support factor is computed (FTU-Frequent Temporal Utility). These rules are filtered based ion minimum support and then the FTU confidence is computed for the rules at the last step. Finally, the rules are pruned based on the minimum confidence value. These rules must be optimized. For optimization Weighed PSO technique is used.

UTARM Approach: UTARM is one among the key concept that we use in our concept in order to generate the association rules for the dataset. The concept mainly associated to temporal association rule mining, the temporal association rule mining is triggered with Utility mining. Utility mining is the problem of finding all item sets in a transactional database D with their utility values higher than the minimum utility threshold (τ). The input database D has a set of partitions, $D = \{P_1, P_2, P_3, \ldots, P_n\}$ where each partition has a set of transactions $T = \{T_1, T_2, \cdots, T_s\}$ and has a set of items; $X_D = \{x_1, x_2, \cdots, x_m\}$.Each item in the partition P_k contains different utility values. The transaction weighted utility value of item x_i TWU (x_i) in a partition P_k is calculated as, as the product of the maximum internal utility and the maximum external utility in the concerned partition. Initially, the process begins by supplying the input data to the first phase. We have a dataset D, which consists of a number of Item set.

$D = \{ I_1 , I_2, \ldots\ldots , I_n \}$

Each of the item sets consists of a number of items. The main function that are inevitable for our proposed approach are the support and the confidence of the item sets, which are used to find the rules.

The support function, which is defined, is derived from FTU support defined on the UTARM method. The support function defines how much the items in the item set give support to each other and this support function incorporates utility value with it. The confidence function is given below. Here l(x) gives the length of the item set in the partition and N gives the total number of transactions from partition *t* to *n*.

$$\mathrm{FTU}_{\mathrm{support(Xt,n)}} = \frac{\sum_{k=t}^{n} \mathrm{support}(X, P_k\ )}{l(X) * N^{t,n} * \sum_{k=t}^{n} \mathrm{TWU}(K)}$$

$$\mathrm{FTU\_}\mathit{c}\mathrm{onf}(\ X \Longrightarrow Y\ ) = \frac{\mathrm{FTU\ (support(\ } X \Longrightarrow Y\ ))}{\mathrm{FTU(support(}X))}$$

The Rules generated may contain redundancy. To remove this redundancy, weighted PSO technique is adapted. The rules generated are stored in the rule array in R.

$$R_D = \{ r_1, r_2, \ldots\ldots\ldots, r_n \}$$

Algorithm: UTARM (For generating Utility based Temporal Rule mining)

Input: Database D partitioned based on month granularity {P_1, P_2…..P_n}, minimum support and minimum confidence.

Output: utility based temporal association rules.

Steps:

- Generate all possible candidate 2-itemsets from Partition P1.
- Mining of FTU 2-itemsets from P1. (FTU-Frequent Temporal Utility)
- Generate candidate 2-itemsets form partition P2 and mining of FTU 2- item sets (P1+P2).
- Mining of FTU 2-itemsets (P1+P2+….+Pn)
- Generate all FTU 1-itemsets from FTU 2-itemsets.
- Mining of all FTU k-item sets (K>2)

Weighted PSO Approach: The main concept of PSO originates from the study of fauna behavior. The role of PSO algorithm is vibrant in the optimization of solutions which obtained through different methods. PSO is introduced by Kennedy, 1995. The PSO considers every solution in the search space as "Particles". The working of PSO algorithm begins with population of candidate solutions called particles, the population can be considered as swarm. In PSO the particles are moved around the search space according to some mathematical formulae. The movement of the particle is governed by the particles best known position and the swarm's best known position. A fitness function is defined over the problem to be optimized and the maximization of the fitness value determines the optimum solutions. In the following algorithm values x_i^n, x_i^o are new position and old position of the particles respectively, and in similar way v_i^n, v_i^o are the new velocity and old velocity of the particles. The process continues up to a termination criteria. Finally we get the value of g, the best known position of the swarm as a global minimum, from that value we optimize the different problem which are subjected for optimization.

Illustration of UTARM and PSO: The following Table 1 shows the list of sample transactions of sales data - month wise. The transactions are given in the form of pairs (item, qty purchased). (2, 3) means Item 2 is purchased and the quantity purchased is 5. Similary utilty (1, 4) means Item 1 is given weightage of 4.

The sample database Table.1 is given as an input to the UTARM technique and the results are shown in Table.2, Table.3, Table.4 and Table.5. The outcome of the UTARM is given to the PSO technique. The inputs and the fitness function computation results are shown in Table.6 and Table.7.

Table.1. Sample Database Transactions

S. No	Particle	Transactions	Utility
1	Set 1 (Jan-2016)	2,5 4,3 2,4 3,6 4,2 2,8 3,10 1,9 4,7	1 , 4 2 , 10 3 , 8 4 , 3
2	Set 2 (Feb-2016)	2,2 3,4 5,8 4,5 5,10 1,6 2,10 3,12 3,9 4,3 5,5	1, 2 2, 6 3, 5 4 , 1 5 ,8
3	Set 3 (March 2016)	2,3 3,1 5,4 6,12 2,8 6,5 1,12 4,6 2,2 3,4 6,3	1, 1 2 , 12 3 , 4 4 , 7 5 , 10 6, 10

Table.2. FTU 2-itemsets of partition P_1

C_2	Beginning	End	Frequency	Utility	FTU support
2,4	1	1	2	105	0.1544
2,3	1	1	2	248	0.3647
3,4	1	1	1	54	0.0397
1,4	1	1	1	57	0.0419

Table.3. FTU 2-itemsets of partition $P_1+P_2+P_3$

C_2	Beginning	End	Frequency	Utility	FTU support
2,3	1	3	6	480	0.0655780
3,5	2	3	3	213	0.0542
4,5	2	3	2	128	0.03636
2,5	3	3	1	76	0.0396
2,6	3	3	3	356	0.55625
3,6	3	3	2	170	0.177083
5,6	3	3	1	160	0.083333
1,4	3	3	1	54	0.028125

Table.4. Mined FTU itemsets

Candidate Items		Frequency	Utility	FTU support
C_1	$2^{3,3}$ $6^{3,3}$	3 3	156 200	0.78 1.00
C_2	$(2,6)^{3,3}$ $(3,6)^{3,3}$	3 2	356 170	0.89 0.88
C_3	$(2,3,6)^{3,3}$	2	230	0.417

From candidate 3 item set $(2, 3, 6)^{3,3}$, various possible rules are: $(2 \rightarrow 6)^{3,3}$, $(2\rightarrow 6,3)^{3,3}$, $(6 \rightarrow 2,3)^{3,3}$, $(6,3\rightarrow 2)^{3,3}$, $(3,6\rightarrow 2)^{3,3}$, etc are generated. For these rules, Confidence is calculated as per the formula 5**.** Confidence above 80% (minimum confidence) is chosen as finalized rules.

Accordingly the filtered rules are as follows: In the above tables C_2 represents candidate 2- itemsets.

Table.5. Final output of UTARM

SNo	Rules	FTU_confidence
1	$(2 \rightarrow 6)^{3,3}$	1.141025
2	$(6 \rightarrow 2)^{3,3}$	0.89
3	$(3,6\rightarrow 2)^{3,3}$	0.88

The output of UTARM is given as an input to the PSO Technique: These rules are considered as particles and for each of the rule initial velocity, position is assigned. These values ranges between [0,1] FTU support and FTU confidence are considered from the UTARM technique. r_{par}, r_{swm} are also random values between [0,1]. The following Table 6 Shows the initial condition of the PSO technique. xi is the initial position of each particle.

Table.6. Initial status of each particle

Rules	Velocity(ri)	Best_position(ri)	FTU_support (support)	FTU_Confidence (confidence)	rr_p	rr_s
1	0.658	0.85	0.89	1.141025	0.3	0.4
2	0.689	0.87	0.89	0.89	0.5	0.6
3	0.691	0.88	0.4177	0.88	0.6	0.5

After updating the initial velocity and position of each particle, the following Table7 shows the status of each particle.

Table.7. Updation of particle velocity, position and finding the fitness value of each particle based on Weighted PSO.

Rules	Velocity(ri)	Best_position(xi)	FTU_support (support)	FTU_Confidence confidence	Fitness value
1	0.67	1.52	0.89	1.141025	0.0867
2	0.699	1.569	0.89	0.89	0.0698
3	0.691	1.571	0.4177	0.88	0.0691

Fitness function Threshold is got by finding the average of the fitness values calculated: In the above table, the Fitness function threshold is 0.0752. Therefore, the particles whose fitness value above the threshold (0.0752) are selected. In the above example: the particle 1 (Rule 1) is selected. 0.0867 > 0.0752. The remaining rules such as rules 2 and 3 are rejected. Therefore, the optimized rule is rule1.

4. CONCLUSION

In this article, the authors have discussed an integrated approach (weighted UTARM with PSO) which combines association rules miner, utility and PSO. The integrated approach has brought a considerable improvements in generating optimized results. The method is well suited for the retailers to take better decision based on the current market scenario. On realizing the importance of the utility miner, further research could be extended with medical data. In the medical domain, the utility factor with respect to UTARM can be considered as the weight assigned to the severity of the disease. The weight factor could be suggested by the domain experts for accurate results prediction. Further, the authors conclude that the integrated approach could be applicable to any type of applications, accordingly the key factor-utility need to be defined.

REFERENCES

Anil Rajput, Ramesh Prasad Aharwal, Nidhi Chandel, Devendra Singh Solanki, Ritu Soni, Approaches of Classification to Policy of Analysis of Medical Data, International Journal of Computer Science and Network security, 9 (11), 2009.

Fayyad U.M, Smyth P and Uthurusamy R, Advances in knowledge discovery and data mining, Menlo Park, AAAI press, 21, 1996.

Han J and Fu Y, Exploration of the Power of Attribute-Oriented Induction in Data Mining, Simon Fraser University, 16, 1996.

Han J, Pei J and Kamber M, Data mining, concepts and techniques, Elsevier, 2011

Her-Shing Wang, Wei-Chang Yeh, Pei-Chiao Huang and Wei-Wen Chang, Using association rules and particle swarm optimization approach for part change, International journal of expert systems with applications, 36 (4), 2009, 8178-8184.

Jyothi Pillai, User centric approach to item set utility mining in Market Basket Analysis, International Journal on Computer Science and Engineering (IJCSE), 3 (1), 2011, 393-400.

Maragatham G and Lakshmi M, A Weighted Particle swarm optimization technique for optimizing Association rules, Proceedings of 4th International Conference on Recent trends in Computing, Communication and Information Technology, 2011, 9-11.

Maragatham G and Lakshmi M, UTARM, an efficient algorithm for mining of utility-oriented temporal association rules, International Journal of Knowledge Engineering and Data Mining, 3 (2), 2015, 208-223.

Mourad Ykhlef, A Quantum Swarm Evolutionary Algorithm for mining association rules in large databases, Journal of King Saud University Computer and Information Sciences, 23, 2011, 1–6.