"OPTIMIZATION OF ASSOCIATION RULE BY MLMS-GA"

[1]SONALI P. MAHINDRE
Department of Computer Science & Engineering, PRMIT&R, Badnera, India
sona.mahindre@gmail.com

[2]DR. G. R. BAMNOTE
Department of Computer Science & Engineering, PRMIT&R, Badnera, India
grbamnote@rediffmail.com

ABSTRACT: *Association rule mining plays a very important role in various data mining process. The diversity of association rule mining spread in various field such as market bucket analysis, medical diagnose and share market prediction. Now a day's various authors and researcher focus on validation of association rule mining. For the validation of association rule mining used various optimization algorithm are used such as genetic algorithm, ACO and particle of swarm optimization also used. Some authors used trigonometric functions for validation of rule such functions are monotonic and non-monotonic. In the continuity of these used sine and cosine based constraints function. The sine and cosine functions work in certain range of interval. For the selection validation of these function used genetic algorithm. Genetic algorithm is very famous optimization technique in association rule mining. Multiple constraints and meta-heuristic function play major role in efficient association rule mining technique. The multiple constrains applied in form of inbound and outbound condition and valued the rule for real time database. Meta-heuristic function applied by many researchers in current research trend in data mining for pattern and rule extraction. These functions optimized rule and reduce the rules of redundancy for association rule mining.*

Keywords: Association rule mining, Genetic algorithm, ACO, meta-heuristic function.

1. INTRODUCTION

Association rule mining concept has been applied to domain and specific problem has been studied, the management of shopping malls and an architecture that makes it possible to construct agents capable of adapting the association rules is used. Data mining means extracting knowledge from large quantity of data. Interesting association could be identified among a large set of data items or set by association rule mining [1, 2]. The finding of interesting relationship between large amounts of business transaction records can help in many business decisions making process. Association rules mining is an important in data mining, and frequent item set mining is a key step of many algorithms for association rules mining. Lots of work done for mining of association rules [3]. When the dataset are huge, the rules generated may be very huge, but some of them are not interesting to the users, so, it is common to set some parameters to reduce the numbers of rules generated at the end, two common parameters support and confidence are used. An association rule R is of the form A — *B,* where A, B are disjoint subsets of the attribute set I. The support for the rule R is the number of database records which contain A U B (often expressed as a proportion of the total number of records).The confidence in the rule R is the ratio:

$$\frac{\text{Support for R}}{\text{Support for A}}$$

Which the support exceeds the required threshold. Such subsets arc referred to as "large", "frequent" or "interesting" sets. [7]

2. DATA MINING TECHNIQUES

There are many major data mining techniques that have been developed and used in data mining:

1) Association Rules
2) Clustering
3) Classification

1) ASSOCIATION RULE MINING

The general statement of association rule mining problem was firstly explained in [Agrawal et al. 1993] by Agrawal. Let's consider I=I1, I2, … , Im be a set of m different attributes, T be the transaction that contains a set of items such that T ⊆ I, D be a database with distinct transaction records Ts. An association rule is an implication in the form of X → Y, where X, Y⊆ I are sets of items called item sets, and X ∩ Y = ø. X is called antecedent and Y is called consequent, the rule means X implies Y[6,7]. There are two necessary basic measures for association rules, support(s) and confidence(c). As the database is huge and users concern about only those usually purchased items, usually threshold of support and confidence are already defined or predefined by users to drop those rules that are not so interesting or useful. Two thresholds are called minimal support and minimal confidence, additional constraints of interesting rules also can be specified by the users [8]. The two basic parameters of Association Rule Mining (ARM) are: support and confidence. Support of an association rule is defined as the percentage of records that contain X U Y to the total number of records in the database.

2) CLUSTERING

Clustering is a process of grouping to similar data. Clustering is the task of identifying a finite set of categories (or clusters) to explain the data [11, 12]. Thus, similar objects are assigned to the same category and dissimilar ones to different categories. Clustering is also called unsupervised learning because the data objects are depicted to a set of clusters which can be treated as classes as well. Clustering is the process of collecting the data records into meaningful subclasses (clusters) in a way which maximizes the similarity within clusters and minimizes the similarity between two separate clusters Other names for clustering are unsupervised learning(machine learning) and segmentation. Clustering is used to get an idea for a given data set Similarity between image regions or a pixel implies clustering in the feature space [17].

3) CLASSIFICATION

Classification is a data mining technique that involves three phases, a learning phase, a testing phase, an application phase. A learning model is to develop in learning phase. It may be in the form of classification rules, a decision tree, or a mathematical formula. As the class label of each training sample is provided, this approach is known as supervised learning. In unsupervised learning means (clustering), the class labels are unknown in advance. In the testing phase data are used to assess or check the accuracy of classifier. If the classifier passes the test phase, it is used for the classification of newly generated data tuple's. This is the application phase [13, 14]. The classifier predicts the class label for these new data samples. For classification algorithms, the two main problems on classifying a data stream are the infinite length and the concept drift. The second one makes the most static stream classification algorithms incapable of classifying a data stream with concept drifts for the changes occurred in the stream. For a time changing data stream, an incremental updating manner of the classifier is very important. A temporal model is used to capture the evolutions of the stream. Generally, the classification process is always accompanied by the course of model construction and test. [15] The classification model keeps changing with the progress of the data stream. If a static classifier is used to classify an evolving data stream, the accuracy of it will drop largely. For a sudden burst of concept drift in a time-changing stream, updated model always provides best accuracy. But for relatively stable time-changing streams, models built with long-term samples will be great. Both of the short-term behaviors and long-term behaviors of the stream are important for a classifier. It is decided by the stream itself and cannot be known a priori [16].

3. RELATED WORKS

[1] In this paper, Xiaobing Liu , Kun Zhai, Witold Pedrycz present a form of the directed item sets graph to store the information of frequent item of transaction databases, and give the trifurcate linked list storage structure of directed item sets graph. Furthermore, we develop the mining algorithm of maximal frequent item sets based on this structure. As a result, scanning a database only once, and improves storage efficiency of data structure and time efficiency of mining algorithm.

[2] In this paper, Ying-Ho Liu propose a new algorithm called U2P-Miner for mining frequent U2 patterns from univariate unfixed data, where each attribute in a transaction is associated with a quantitative interval and a probability density function. This algorithm is implemented in two phases. First, we construct a U2P-tree that compresses the information in the target database. Then author use the U2P-tree to discover frequent U2 patterns. Potential frequent U2 patterns are derived by joining base intervals and verified by traversing the U2P-tree. They also develop two techniques to boost speed of the data mining process. Since the proposed method is based on a tree-traversing strategy, it is both efficient and scalable.

[3] In this paper, Li Guang-yuan, Cao Dan yang, Guo Jian-wei present an algorithm for mining association rules with multiple constraints, the implemented algorithm simultaneously copes with two different kinds of constraints, it consists of three phases, first, the frequent 1 –item set arc generated, second, we exploit the properties of the given constraints to improve search space or save constraint checking in the conditional databases. Third, for each item set possible to satisfy the constraint, they generate its conditional database and perform the three phases in the conditional database recursively. Experimental results shows that the proposed method outperform the revised FP-growth algorithm. The problem of discovering all frequent item sets that satisfy constraints is a difficult.

[4] The aim of this study is to develop a new model for mining interesting negative and positive association rules out of a transactional data set. The proposed model is integration between two algorithms, the (PNAR) Positive Negative Association Rule algorithm and the (IMLMS) Interesting Multiple Level Minimum Supports algorithm, to propose a new approach (PNAR IMLMS) for mining both positive and negative association rules from the interesting frequent and infrequent item sets mined by the IMLMS model. The practical results show that the PNARIMLMS model provides significantly better results than the previous model. One demerit of this algorithm is negative rule extract from uninteresting pattern sets which is useless.

[5] In this paper author described The method is combined with the concept of hierarchical concept, data of the generalization sets processing, and uses neural network generalization into the database after the transaction, by way of introducing an internal threshold so there is no need to set the minimum support threshold.

By simulating the case shows that the method can not only efficient mining single and cross-layer association rules, but also the association rules is new ,easy to understand and meaningful.

[6] In this paper author described about the Association rule and clustering and the details are, A web based recommendation system include browsing history database attached with information related to the web pages that a user browsed. [5] This paper represents the Prediction of User navigation patterns of WUM using Association Rule and Clustering from web log data. In the first stage, separating the potential users is processed, and in the second stage clustering process is used to combine the users with similar interest, and in the last stage association and clustering is used to navigate the user future requests.

[7] In this paper Gavin Shaw, Yue Xu, Shlomo Geva propose two approaches which measure multi-level association rules to evaluate their interestingness. These measures of diversity and peculiarity can be used to help identify those rules from multi-level datasets that are useful. Association rule mining is technique that is widely used when querying databases, especially those that are transactional database, in order to obtain useful associations or correlations between sets of items. Much work has been done focusing on efficiency, effectiveness and redundancy. There has also been a focus on the quality of rules from single level datasets.

4. PROPOSED WORK

Validation of association rule mining is very important and critical phase of rule generation. The dependence of rule generation has two factors one is minimum support value and other is minimum confidence value. The process of validation satisfy the given condition and used for the process of rule generation. The generated rules come along with some negative rule and some positive rule in form of rule set. Now the process of strong rule generation used various constraints function such as monotonic and non-monotonic function for range validation of association rule mining. Some authors also used some multi-level approach for generation of association rule mining. Some also used some optimization technique for strong rule generation. In this dissertation proposed a sine cosine multiple constraints based association rule mining. The sine and cosine function work in

maximum and minimum range of value for the processing of given data.

Proposed algorithm for optimization of association rule mining, the proposed algorithm resolves the problem of negative rule generation and optimized the process of rule generation. Negative association rule mining is the great challenge for huge dataset. In the generation of valid rules association existing algorithm or method generate a series of negative rules, which generated rule affect a performance of association rule mining. In the process of rules generation various multi objective associations rule mining algorithm is used but all these are not solve. In this Paper we proposed MLMS-GA of association rule mining with min-max algorithm. In this algorithm we used MLMS used for multi-level minimum support for constraints validation. The scanning process of database divided into multiple levels as frequent level and infrequent level of data according to multi-level minimum support (MLMS). The frequent data logically assigned 1 and infrequent data logically assigned 0 in MLMS process. The divided process reduces the unnecessesory item in given database. The proposed algorithm is a combination of MLMS and min-max algorithm along this used level weight for the separation of frequent and infrequent item. The weight value act as Support length key is a vector value given by the transaction data set. The support value passed as a vector for finding a closest level between MLMS candidates key. After finding a MLMS candidate key the nearest level divided into two levels, one level take a higher order value and another level gain infrequent minimum support value for rule generation process. The process of selection of level also helps to reduce the passes of data set. After finding a level of lower and higher of given support value, compare the value of level weight vector. Here level length vector work as a fitness function for selection process of min-max algorithm.

Steps of algorithm (MLMS-GA)

1. Scanning of database used flowing steps

Some standard notation of pseudo code of algorithm such as D dataset, K level MLMS, Ls generation candidate

K = MLMS dataset (D)
n = Number of multiple level block
For i = 1 to n loop
Scan_k (Ki ∈k)
Li = gen__itemsets (ki)
For (i = 2; L^j i≠ φ, j = 1,2....,n; i++)
Ci ^{G} = ∪j = 1,2,...nL_i^j
End;
For i = 1 to n scan_kmap (ki∈K)
For all items C ∈CG generate block (C, ki)
End;
LG = {c ∈CG|}

2. **Generate multiple support vector value for selection process**

for all transaction LG do
generate count table TC
L_1 =(frequent 1-itemsets);
C_2 =L_1 ∞ L_1;
L_2 ={cEC_2 | sup(c)≥MinSupNum};
For(k=3;L_{k-1} ≠Ø ;k++)do begin
For (j=k; j≤m; j++) do
Generate $CIV_{ij}^{k-1;}$
C_k=candidate_gen (L_{k-1})
L_k = {cEC_k| sup(c) ⩾MinSupNum}; \
End

3. Set of rule is generated

Return L = U L_k;
Candidate_gen (frequent itemset L_{k-1})

a. for all(K-1)-itemsetlE L_{k-1} do
b. for all i_jE $L_{k-1\ do}$
c. //S is the result of the formula(2)
If for every r (1≤r≤k) such that S[r] ⩾k-1 then

L_1 = (frequent 1-itemsets);
C_2 =L_1 ∞ L_1;
L_2 = {cEC_2 | sup(c) ⩾MinSupNum};
For (k=3; Lk_{-1} ≠Ø; k++) do begin
For (j=k; j≤m; j++) do
Generate $CIV_{ij}^{k-1;}$
C_k=candidate_gen (L_{k-1})

4. Check MLMS value of table

5. If rule is not MLMS go to selection process

6. Else optimized rule is generated.

7. Exit

a) Data Encoding

The process of data in min-max algorithm needs some data encoding technique for representation of data. In this technique used binary encoding technique.

b) Fitness function

The population selection of Min-max Algorithm is a design of Fitness Function:

$$m(S) = \frac{Ai}{wi} + \frac{Bi}{L \times (1 - wi)}$$

Ai = {frequent item support}

Wi_1= {level of Wight value of MLMS}

Bi = {those value or Data infrequent}

The selection strategy based on the fitness and concentration, the probably of selection (pi) of individual whose fitness value is greater than one and m(s) is a those value whose fitness is less than one but near to the value of 1. The Min-max operators determine the search capability and convergence of the algorithm. Min-max operators hold the selection crossover and mutation on the population and generate the new population. In this algorithm it restore each chromosome in the population to the corresponding rule, and then finally calculate the selection probability (pi) for each rule based on above formula. In which single point are used. It divide multiple level domain of each attribute into a group and classifies the cut point of each continuous attributes into one group and the crossover carried out between the corresponding groups of two individuals by a certain or fixed rate rate. Any bit in the chromosomes is mutated by a certain or fixed rate, that is, changing "0"to"1","1"to"0". Now we explain complete process of algorithm shows block diagram of proposed algorithm using min-max algorithm.

5. CONCLUSION

In this Paper we proposed a novel method for optimization of association rule mining. Our propped algorithm is combination of min-max function and genetic algorithm. The min-max function and genetic algorithm work together and perform condition based rule generation process. The min-max condition based function operates in sine and cosine based trigonometric function for the processing of genetic fitness function.

We have observed that when we modify the condition new rules in large numbers are found. This implies that when min-max is solely determined through support and confidence, there is a high chance of eliminating interesting rules. With more rules emerging it implies there should be a mechanism for managing their large numbers. This largely generated rule is optimized with genetic algorithm.

We theoretically proved a relation between locally huge and globally large patterns that is used for local pruning at each site to reduce the searched candidates. We derive a locally large threshold using a globally set minimum recall threshold. Local pruning achieves a reduction in the number of searched candidates and this reduction has a proportional impact on the reduction of exchanged messages.

Our proposed algorithm is mixture or combination of MLMS and min-max algorithm. We have observed that when we modify the scan process of transaction generation of rule is fast. With more rules it shows that there should be a mechanism for managing their large numbers. The large generated rule is optimized with min-max algorithm.

Our proposed algorithm is performed better optimization in comparison of MLML-GA and monotonic condition based association rule mining.

6. REFERENCES

[1] Xiaobing Liu, Kun Zhai, Witold Pedrycz "An Improved Association Rules Mining Method" Expert Systems with Applications 2012, Pp 1362–1374.

[2] Ying-Ho Liu "Mining Frequent Patterns from Univariate Uncertain Data" Data & Knowledge Engineering, 2012. Pp 47-68.

[3] Li Guang-Yuan, Cao Dan Yang, Guo Jian-Wei "Association Rules Mining With Multiple Constraints" Elsevier Ltd. 2011, Pp 1678-1683.

[4] Idheb Mohamad Ali O. Swesi, Azuraliza Abu Bakar, Anis Suhailis Abdul Kadir "Mining Positive And Negative Association Rules From Interesting Frequent And Infrequent Item Sets" IEEE 9th International Conference On Fuzzy Systems And Knowledge Discovery, 2012. Pp 650-655.

[5] Huang Qinglan, Duan Longzhen "Multi-Level Association Rule Mining Based On Clustering Partition" IEEE Third International Conference on Intelligent System Design and Engineering Applications, 2013. Pp 982-986.

[6] Vidhu Singhal, Gopal Pandey "A Web Based Recommendation Using Association Rule and Clustering" International Journal of Computer & Communication Engineering Research, 2013. Pp 1-5.

[7] Gavin Shaw, Yue Xu, Shlomo Geva "Interestingness Measures For Multi-Level Association Rules' Proceedings of the 14th Australasian Document Computing Symposium, 2009. Pp 2-9.

[8] Aritra Roy, Rajdeep Chatterjee "Introducing New Hybrid Rough Fuzzy Association Rule Mining Algorithm" Proc. Of Int. Conf. On Recent Trends In Information, Telecommunication And Computing, ITC, 2014. Pp 167-175.

[9] Deepak A. Vidhate, Dr. Parag Kulkarni "Improvement In Association Rule Mining By Multilevel Relationship Algorithm" International Journal Of Research In Advent Technology, Vol-2, 2014. Pp 366-373.

[10] Noha Negm, Passent Elkafrawy, Mohamed Amin, Abdel Badeeh M. Salem "Investigate the Performance of Document Clustering Approach Based On Association Rules Mining" International Journal of Advanced Computer Science and Applications, Vol-4, 2013. Pp 142-151.

[11] Dong, X., Zheng, Z., Niu, Z., Jia, and Q.: Mining Infrequent Itemsets Based On Multiple Level Minimum Supports. Proceedings Of The Second International Conference On Innovative Computing, Information And Control (ICICIC 2007) (2007).

[12] Dong, X., Niu, Z., Zhu, D., Zheng, Z., Jia, and Q.: Mining Interesting Infrequent And Frequent Itemsets Based On MLMS Model. The Fourth International Conference On Advanced Data Mining And Applications, ADMA 5139.444-451(2008)

[13] Choo, Y.H., Bakar, A. A.., Hamdan, A. R.: Linguistic Association Rules Mining. Proceedings Of The International Conference On Electrical Engineering And Informatics Institute Technology Bandung, Indonesia.(2007)

[14] Cai Hongguo, Yuan Changan, Luo Jinguang, Huang Jinde. A Novel GEP - Based Multiple - Layers Association Rule Mining Algorithm (J). Computational Intelligence and Security. 2010, (13) : 68

[15] Wang Chunyu, Wang Xuehua, Zhang Xujuan, Wang Yanzhang. Criminal Cases Multiple Correlation Analysis Model Study (J). Intelligence Journal, 2011, 21 (1) : 45-50

[16] Wand Liangming. Web Log Mining Research of Fusion Multiple Fuzzy Matrix _SOFM (D). Nanchang. Nanchang University. 2010

[17] Yun-Huoy Choo, Bakar A.A., Hamdan A.R. 2008. The Fitness-Rough: A New Attribute Reduction Method Based On Statistical and Rough Set Theory, Intelligent Data Analysis. Vol 12(1). ISI Index. IOS Press. IMPACT FACTOR 0.929.