ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

Mining Association Rules from Infrequent Itemsets: A Survey

Diti Gupta[1], Abhishek Singh Chauhan[2]

M.Tech Student, Department of CSE, NIIST, Bhopal (M.P.), India

Asst. Professor, Department of CSE, NIIST, Bhopal (M.P.), India

ABSTRACT: Association Rule Mining (AM) is one of the most popular data mining techniques. Association rule mining generates a large number of rules based on support and confidence. However, post analysis is required to obtain interesting rules as many of the generated rules are useless.However, the size of the database can be very large. It is very time consuming to find all the association rules from a large database, and users may be only interested in the associations among some items.So mining association rules in such a way that we maximize the occurrences of useful pattern. In this paper we study several aspects in this direction and analyze the previous research.So that we come with the advantages and disadvantages.

Keywords: ARM, Positive Association, Negative Association, support, confidence

1. INTRODUCTION

Mining association rules is an important task. Past transaction data can be analyzed to discover customer purchasing behaviors such that the quality of business decision can be improved. The association rules describe the associations among items in the large database of customer transactions. However, the size of the database can be very large.A large amount of research work hasbeen devoted to this area, and resulted in such techniques as k-anonymity [1], data perturbation [2], [3], [4], [5], and data mining based on [6], [7].

Association Rule Mining (ARM) is one of the most used research are in data mining. ARM can be used for discovering hidden relationship between items. By given a user-specified threshold, also known as minimum support, the mining of association rules can discover the complete set of frequent patterns. That is, once the minimum support is given, the complete set of frequent patterns is determined[8]. In order to retrieve more correlations among items, users may specify a relatively lower minimum support [8].ARM is also studied in terms of market basket analysis, which is the analysis of the itemset which can be analyzed after the customer purchasing in the mall [8]. It is just like the analysis of the customer of purchasing behavior. Association rules also used in various areas such as telecommunication networks, market, risk management and inventory control etc.[8][9].

In [10] author suggests that Data mining is used everywhere and large amounts of information are gathered: in business, to analyses client behavior or optimize production and sales [1].This signifies the research direction in several fields. We can use ARM and data mining application in health care, medical database, classification and combining these techniques with other approach extensively increases the potential behavior and applicability.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

In [11] author suggests that many of the researchers are generally focused on finding the positive rules only but they not find the negative association rules. But it is also important in analysis of intelligent data. It works in the opposite manner of positive rule finding. But problem with the negative association rule is it uses large space and can take more time to generate the rules as compare to the traditional mining association rule [12][13]. So better optimization technique can find a better solution in the above direction.

We provide here a brief survey on ARM. Other sections are arranged in the following manner: Section 2 introduces negative and positive association rules; Section 3 describes about Literature Review; Section 4 discusses about problem domain; section 5 shows the analysis; Section 6 describes Conclusion.

2.NEGATIVE AND POSITIVE ASSOCIATION RULE

Association rule mining can be represent in terms of A ⇒B [S, C] where A and B are sets of items; S is the support of the rules, defined as the rate of the transactions containing all items in A and all items in B i.e. Support (A ⇒B) = P (A ∪B) and C is the confidence of the rule, defined as the ratio of S with the rate of transactions containing A i.e. P (B / A). Support and confidence are measures of the interestingness of the rule. We calculate the support value for justifying the usefulness of the items present in the data set. A Higher support value indicates the effectiveness for the enterprise. The confidence signifies the decision theory, higher the confidence higher the decision accuracy.

Negative association rules of form A=>~B means supp(AU~B)≥ms.supp(AUB)= supp(A)-supp(AU~B). For most cases, the supp(A)<2*ms. Therefore supp(AUB)<ms, which means AUB is infrequent itemsets. So to find negative association rules, we need to find infrequent itemsets first.

Support shows the frequency of the patterns in the rule; it is the percentage of transactions that contain both A and B, i.e.

Support = Probability(A and B)
Support = (# of transactions involving A and B) / (total number of transactions).

Confidence is the strength of implication of a rule; it is the percentage of transactions that contain B if they contain A, ie.

Confidence = Probability (B if A) = P(B/A)
Confidence =
(# of transactions involving A and B) / (total number of transactions that have A).

3. LITERATURE REVIEW

In 2008, He Jiang et al. [14] suggest the weighted association rules (WARs) mining are made because importance of the items is different. Negative association rules (NARs) play important roles in decision-making. But according to the authors the misleading rules occur and some rules are uninteresting when discovering positive and negative weighted association rules (PNWARs) simultaneously. So another parameter is added to eliminate the uninteresting rules. They propose the support-confidence framework with a sliding interest measure which can avoid generating misleading rules. An interest measure was defined and added to the mining algorithm for association rules in the model. The interest measure was set according to the demand of users. The experiment demonstrates that the algorithm discovers interesting weighted negative association rules from large database and deletes the contrary rules.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

In 2009, Yuanyuan Zhao et al. [15] suggest that the Negative association rules become a focus in the field of data mining. Negative association rules are useful in market-basket analysis to identify products that conflict with each other or products that complement each other. The negative association rules often consist in the infrequent items. The experiment proves that the number of the negative association rules from the infrequent items is larger than those from the frequent.

In 2011, WeiminOuyang et al. [16] suggest three limitations of traditional algorithms for mining association rules. Firstly, it cannot concern quantitative attributes; secondly, it finds out frequent itemsets based on the single one user-specified minimum support threshold, which implicitly assumes that all items in the data have similar frequency; thirdly, only the direct association rules are discovered. They propose mining fuzzy association rules to address the first limitation. In this they put forward a discovery algorithm for mining both direct and indirect fuzzy association rules with multiple minimum supports to resolve these three limitations.

In 2012, YihuaZhong et al. [17] suggest that association rule is an important model in data mining. However, traditional association rules are mostly based on the support and confidence metrics, and most algorithms and researches assumed that each attribute in the database is equal. In fact, because the user preference to the item is different, the mining rules using the existing algorithms are not always appropriate to users. By introducing the concept of weighted dual confidence, a new algorithm which can mine effective weighted rules is proposed by the authors. The case studies show that the algorithm can reduce the large number of meaningless association rules and mine interesting negative association rules in real life.

In 2012, He Jiang et al. [18] support the technique that allows the users to specify multiple minimum supports to reflect the natures of the itemsets and their varied frequencies in the database. It is very effective for large databases to use algorithm of association rules based on multiple supports. The existing algorithms are mostly mining positive and negative association rules from frequent itemsets. But the negative association rules from infrequent itemsets are ignored. Furthermore, they set different weighted values for items according to the importance of each item. Based on the above three factors, an algorithm for mining weighted negative association rules from infrequentitemsets based on multiple supports(WNAIIMS) is proposed by the author.

In 2012, IdhebaMohamad Ali O. Swesi et al. [19] study is to develop a new model for mining
interesting negative and positive association rules out of a transactional data set. Their proposed model is integration between two algorithms, the Positive Negative Association Rule (PNAR) algorithm and the Interesting Multiple Level Minimum Supports (IMLMS) algorithm, to propose a new approach (PNAR_IMLMS) for mining both negative and
positive association rules from the interesting frequent and infrequent item sets mined by the IMLMS model. The experimental results show that the PNAR_IMLMS model provides significantly better results than the previous model.

In 2012, WeiminOuyang [20] suggest that traditional algorithms for mining association rules are built on the binary attributes databases, which has three limitations. Firstly, it cannot concern quantitative attributes; secondly, only the positive association rules are discovered; thirdly, it treat each item with the same frequency although different item may have different frequency. So he puts forward a discovery algorithm for mining positive and negative fuzzy association rules to resolve these three limitations.

In 2012, XiaofengZheng et al. [21] presented the theory, question and resolution of application of rough set in mining association rules. And it presented resolve the relation of support, confidence and the amount of rules by rough set analysis originally. According to the authors the entire conclusions were proved in data mining in provincial road transportation management information System.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

In 2013, AnjanaGosainet al. [22] suggest that the traditional algorithms for mining association rules are built on binary attributes databases, which has two limitations. Firstly, it cannot concernquantitative attributes; secondly, it treats each item with the samesignificance although different item may have differentsignificance[23]. Also binary association rules suffers from sharpboundary problems [24]. According to the authors many real world transactionsconsist of quantitative attributes. That is why several researchershave been working on generation of association rules for quantitativedata. Theypresent different algorithms given by variousresearches to generate association rules among quantitative data. They have done comparative analysis of different algorithms forassociation rules based on various parameters. They also suggest a future suggestion that there is the need of a framework of association rules for data warehouse that overcome the problems observed by various authors.

In 2013, Luca Cagliero et al. [25] tackle the issue of discovering rare and weighted itemsets, i.e., the Infrequent Weighted Itemset (IWI) mining problem. They proposed two novel quality measures to drive the IWI mining process. Two algorithms that perform IWI and Minimal IWI mining efficiently, driven by the proposed measures, are presented. Experimental results show efficiency and effectiveness of the proposed approach.

In 2013, Johannes K. Chiang et al. [26] aims at providing a novel data schema and an algorithm to solvethe some drawbacks in conventional mining techniques. Since most of them perform the plain mining based on predefined schemata through the data warehouse as a whole, a re-scan must be done whenever new attributes are added. Secondly, an association rule may be true on a certain granularity but fail on a smaller one and vise verse, they are usually designed specifically to find either frequent or infrequent rules. A forest of concept taxonomies isused as the data structure for representing healthcareassociations patterns that consist of concepts picked up fromvarious taxonomies. Then, the mining process is formulated as acombination of finding the large item sets, generating, updatingand output the association patterns. Their research presents experimentalresults regarding efficiency, scalability, information loss, etc. ofthe proposed approach to prove the advents of the approach.

4. PROBLEM DOMAIN

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. In the real time environment we can subdivide the problem in two parts.First is to find the set exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second stage is the occurrence generation from association rules. So if we apply dynamic minimum support then level wise decomposition is easy. If we think of the most appropriate and efficient data mining algorithm then we always think about Apriori algorithm. However there are two bottlenecks of the Apriori algorithm. First is the process of candidate generation which can increase the time as well as the space. So the second thing is generated from the first that it needed multiple scan when it in the iteration process. Based on Apriori algorithm, many new algorithms were designed with some modifications or improvements.

The computational cost of association rules mining can be reduced in the following ways:
- It can be reduced by reducing the passes.
- Need sampling
- Add extra constrains according to your requirements

In the traditional approach of discovering rules they need lot of search space, redundant dat and dangling tuples. The missing item set is called dangling tuples. To reduce the search space, and to improve the quality of the mined rules, it is fruitful to introduce additional measures apart from support and confidence which is achieved by negative associations. The

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

traditional Apriori-based implementations are efficient but cannot generate all valid positive and negative ARs. So we try to solve that problem without paying too high a price in terms of computational costs and reducing space with less time.

The main challenges:

- How to effectively search for interesting item sets
- How to effectively identify negative association rules of interest.
- Identification of negative association rules exist in the non-frequentitem set.
- In the traditional algorithms, the process of data mining for association rules generally split in two parts: first, mining for frequent item sets; and second, generating strong association rules from the discovered frequent item sets. But to club both of the sequences and generate the classified rule on-the-fly while analyzing the correlations within each candidate or non-candidate item set is missing which also avoids evaluating item combinations redundantly.
- Criteria of the discovered rules for the user requirements may not be the same. Many uninteresting association rules for the user requirements can be generated when traditional mining methods are applied.

5. ANALYSIS

After studying several research papers we come with the following analysis:

In [15] they conducted their experiments on a man-made dataset to study the behavior of the algorithms compared. The dataset had 200 transactions, when only the six largest categories were kept. Compared an algorithm WNRIF with traditional Algorithm PNWAR, the number of negative rules mined from the two algorithm is different.

Experiment one
The weighted minimum support (*wms*) is equal, but the weighted minimum confidence is set different values. Wms=3%

TABLE 1: RESULT 1[15]

wmc	50%	60%	70%	80%
PNWAR	214	199	194	175
WPNIF	354	335	330	306

The weighted minimum confidence (wmc) is equal, but the weighted minimum support is set different values. wmc=40%

TABLE 2: RESULT 2[15]

wms	3%	5%	7%
PNWAR	214	76	36
WPNIF	354	104	90

In [18] they assume that wmc is equal to 50%,mininterest is equal to 10, the values of multiple minimumsupport are as follows in Table 1 and the result of weightedfrequent and infrequent itemsets as follows:

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

TABLE 3: THE NUMBERS OF WEIGHTED FREQUENT AND INFREQUENT ITEMSETS [18]

		Number of Weighted Frequent and Infrequent Itemset					
Wms	WF /WIF	1set	2set	3set	4set	5set	6set
0.05	WF	8	14	25	14		
	WIF	0	0	0	26	51	
1set 0.05 2 set 0.05 3 set 0.04 4 set 0.03 5-set 0.03	WF	8	14	25	30	1	
	WIF	0	0	0	4	53	
1set 0.05 2 set 0.05 3 set 0.04 4 set 0.03 5-set 0.03		8	14	25	30	15	
		0	0	0	4	25	26

In [27] proposed algorithm tested with the T40I10D100Kand heart disease prediction. The proposed PVARM arecompared with traditional apriori, most interesting rule miningalgorithm and non-redundant algorithm. Their comparative study of no. of association rule mined fromT40I10D100K and heart disease prediction dataset. Theminimum support set as 50% and the minimum confidence setas 50%. The results are summarized in the Table below. It reducedthe 50% rules compared with non-redundant algorithm, 42%rules are reduced from most interesting rule mining algorithmand 80% of rules reduced to compare with traditional aprori.

TABLE 4: RESULTS [9]

Datasets	**Apriori**	**MIR**	**NRRM**	**PVARM**
T40I10D100K	40321091	678542	22700	11209
Heart disease prediction	3210	2470	1056	710

6. Conclusion and Future Suggestions

In this paper we survey about the finding of negative and positive association rules form the infrequent itemset. We come with our study with several advantages and the problem formulation which can be implemented in future. Based on our study we suggest that to extend the support-confidence framework in a dynamic fashion. In addition to finding confident positive rules that have a strong correlation, the algorithm discovers negative association rules with strong negative correlation between the antecedents and consequents. So we can devise an efficient which generates both positive and negative association rules. We also generates all types of confined rules, thus allowing to be used in different applications

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

where all these types of rules could be needed or just a subset of them. So we achieve better frequency result set for all the item set in both positive and negative associations.

REFERENCES

[1] Sweeney, L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557-570, 2002.

[2] Agrawal, R. and Srikand R. Privacy preserving data mining. In Proc. Of ACM SIGMOD Conference, pp. 439-450, 2000.

[3] Chen, K. and Liu. L. A random rotation perturbation approach to privacy data classification. In Proc of IEEE Intl. Conf. on Data Mining (ICDM), pp. 589-592, 2005.

[4] Xu, S., Zhang, J., Han, D. and Wang J. Singular value decomposition based data distortion strategy for privacy distortion. Knowledge and Information System, 10(3):383-397, 2006.

[5] Mukherjeee, S., Chen, Z. and Gangopadhyay, A. A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier related transforms. Journal of VLDB, 15(4):293-315, 2006.

[6] Vaidya, J. and Clifton, C. Privacy preserving k-means clustering over vertically partitioned data. In Prof. of ACM SIGKDD Conference, pp.206-215, 2003.

[7] Vaidya, J., Yu, H. and Jiang, X. Privacy preserving SVM classification. Knowledge and Information Systems, 14:161-178, 2007.

[8] Ms. KumudbalaSaxena, Dr. C.S. Satsangi, "A Non Candidate Subset-Superset Dynamic Minimum Support Approach for sequential pattern Mining", International Journal of Advanced Computer Research (IJACR),Volume-2 Number-4 Issue-6 December-2012.

[9] Dr. Manish Shrivastava, Mr. Kapil Sharma, MR. Angad Singh, "Web Log Mining using Improved Version of Proposed Algorithm", International Journal of Advanced Computer Research (IJACR),Volume 1 Number 2 December 2011.

[10] PragatiShrivastava, Hitesh Gupta," A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (IJACR), Volume-2 Number-3 Issue-5 September-2012.

[11] Nikhil Jain,VishalSharma,MaheshMalviya, "Reduction of Negative and Positive Association Rule Mining and Maintain Superiority of Rule Using Modified Genetic Algorithm", International Journal of Advanced Computer Research (IJACR) ,Volume-2 Number-4 Issue-6 December-2012.

[12] PreetiKhare, Hitesh Gupta, "Finding Frequent Pattern with Transaction and Occurrences based on Density Minimum Support Distribution", International Journal of Advanced Computer Research (IJACR),Volume-2 Number-3 Issue-5 September-2012.

[13] Leena A Deshpande, R.S. Prasad, "Efficient Frequent Pattern Mining Techniques of Semi Structured data: a Survey", International Journal of Advanced Computer Research (IJACR) Volume-3 Number-1 Issue-8 March-2013.

[14] He Jiang; Yuanyuan Zhao; Xiangjun Dong, "Mining Positive and Negative Weighted Association Rules from Frequent Itemsets Based on Interest," Computational Intelligence and Design,ISCID '08. International Symposium on , vol.2, no., pp.242,245, 17-18 Oct. 2008.

[15] YuanyuanZhao,He Jiang; RunianGeng; Xiangjun Dong, "Mining Weighted Negative Association Rules Based on Correlation from Infrequent Items," Advanced Computer Control,ICACC '09. International Conference on, vol., no., pp.270,273, 22-24 Jan. 2009.

[16] WeiminOuyang; Qinhua Huang, "Mining direct and indirect fuzzy association rules with multiple minimum supports in large transaction databases," Fuzzy Systems and Knowledge Discovery (FSKD), Eighth International Conference on , vol.2, no., pp.947,951, 26-28 July 2011.

[17] YihuaZhong; Yuxin Liao, "Research of Mining Effective and Weighted Association Rules Based on Dual Confidence," Computational and Information Sciences (ICCIS), Fourth International Conference on , vol., no., pp.1228,1231, 17-19 Aug. 2012.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

[18] He Jiang, Xiumei Luan and Xiangjun Dong," Mining Weighted Negative Association Rules from Infrequent Itemsets Based on Multiple Supports", International Conference on Industrial Control and Electronics Engineering, 2012.

[19] IdhebaMohamad Ali O. Swesi, Azuraliza Abu Bakar, AnisSuhailis Abdul Kadir," Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets", 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012.

[20] WeiminOuyang," Mining Positive and Negative Fuzzy Association Rules with Multiple Minimum Supports", International Conference on Systems and Informatics, 2012.

[21] XiaofengZheng and JianminXu," Studies on the Application of Rough set Analysis in Mining of Association Rules and the Realization in Provincial Road Transportation Management Information System", International Conference on Industrial Control and Electronics Engineering,2012.

[22] AnjanaGosain and ManeelaBhugra," A Comprehensive Survey of Association Rules On Quantitative Data In Data Mining", IEEE Conference on Information and Communication Technologies, 2013.

[23] WeiminOuyang and Qinhua Huang, Mining Direct and Indirect Weighted Fuzzy Association Rules in Large Transaction Databases, IEEE Eighth International Conference on Fuzzy Systems and Knowledge Discovery, 2011.

[24] Keon-MyungLee,Mining Generalized Fuzzy Quantitative Associaton Rules wih Fuzzy Generalization Hierarchies, IEEE 2011.

[25] Luca Cagliero and Paolo Garza," Infrequent Weighted Itemset Mining using Frequent Pattern Growth", IEEE Transactions on Knowledge and Data Engineering,2013.

[26] Johannes K. Chiang and Sheng-Yin Huang," Multidimensional Data Mining for Healthcare Service Portfolio Management", IEEE 2013.

[27] K.Rameshkumar, M.Sambath and S.Ravi," Relevant Association Rule Mining from Medical Dataset Using New Irrelevant Rule Elimination Technique",IEEE 2013.