A Survey on Infrequent Weighted Itemset Mining Approaches

R. PRIYANKA, S. P. SIDDIQUE IBRAHIM

***Abstract*— Association Rule Mining (ARM) is one of the most popular data mining technique. All existing work is based on frequent itemset. Frequent itemset find application in number of real-life contexts e.g., market basket analysis, medical image processing, biological data analysis. In recent years, the attention of researchers has been focused on infrequent itemset mining. This paper tackles the issues of discovering the rare and weighted itemsets. The infrequent itemset mining problem is discovering itemsets whose frequency of the data is less than or equal to maximum threshold. This paper surveys various method of mining infrequent itemset. Finally, comparative way of each method is presented.**

***Index Terms*: Data Mining, Frequent itemset, infrequent itemset, FP- growth Algorithm.**

I. INTRODUCTION

Data is unprocessed or raw fact. It does not give any meaning. Information can be considered as an aggregation of data. It has usually some meaning and purpose. Information can be converted into knowledge about historical pattern and future trends.

Data Mining is defined as"Extraction interesting patterns or knowledge from huge amount of data". Data mining is the procedure for discovering data from different viewpoints and summarizing into useful information. Discovering of usual patterns hidden in a database plays an vital role in several data mining task. There are two predict kinds of models in data mining. One is predictive model which uses data which uses data with known result td develop a model that can use explicitly to predict values. Another is descriptive model, which describes the pattern in existing data.

Classification is a model or classifier is constructed to predict class label. It composed of two steps: supervised learning of a training set of data to create a model, and then classifying the data according to the model. It is based on predictive model. Regression analysis is a statistical methodology that is most often used for numerical prediction. It is used to predict missing or unavailable numerical data values rather than class labels. Many real world data mining applications can be seen as predicting future data states based on past and current data. Clustering, Summarization, Association rule and sequence discovery are descriptive in nature. Clustering involves identifying a finite set of categories to describe the data. Each member in a cluster should be very similar to other member in its cluster and dissimilar to other cluster. Sequential discovery is used to determine sequential patents in data. These patterns are based on a time sequence of actions. Association rule mining is popular and well researched data mining technique for finding interesting relationship between variables in a large database.

A. Applications of Data mining

Market basket analysis
Risk analysis
Fraud Detection
Biological data analysis...etc

B. Measures of Association Rule Mining:

Support*:* The rule X=>Y holds with support s if s% of transaction in D contains XUY. Rules that have a 's' greater than a user specified support is said to have minimum support. For example

TID	ITEMS	Support=Occurrence/Total Support
1	ABC	Total Support = 5 Support {AB} = 2/5 =40% Support {BC} = 3/% =60% Support {ABC}=1/5 =20%
2	ABD	
3	BC	
4	AC	

Manuscript received Feb, 2013.

Priyanka.R, Computer Science Engineering , Kumaraguru college of Technologye, Coimbatore, India.

S. P. Siddique Ibrahim, Computer Science Engineering , Kumaraguru college of Technologye, Coimbatore, India.

5	BCD	

Table 1 Example of Support Measure

Confidence: The rule X=>Y holds with support c if c% of transaction in D contain X also contain Y. Rules that have a 'c' greater than a user specified support is said to have minimum confidence. For example,

TID	ITEMS	Confidence=Occurrence{Y}/Occurrence{Y}
1	ABC	Total Support = 5 Support {AB} = 2/5 =40% Support {BC} = 3/% =60% Support {ABC}=1/5 =20%
2	ABD	
3	BC	
4	AC	
5	BCD	

Table 2 Example of Confidence Measure

C. *Itemset Mining*:

Itemset mining is an exploratory data mining technique widely used for discovering valuable correlation among data. Itemset mining of two types they are

- *Frequent itemset mining:*

Itemset mining focused on discovering frequent itemset. Itemset is frequent if its support satisfies given minimum support threshold. Frequent itemset find application in many real life contexts. For example buying a PC first, then a digital camera and then a memory card, if it occurs frequently in shopping history, then it is called frequent pattern. Market basket analysis is one of the frequent itemset mining applications.

- *Infrequent itemset mining:*

Patterns that are rarely found in database are considered to be unexciting and are eliminated using the support measure. Item set is infrequent if its support is less than or equal to predefined support threshold. This method has a great interest as they deal with rare but crucial cases. Applications in infrequent itemset mining include identifying rare diseases, predicting equipment failure, and finding association between infrequently purchased items.

II. LITERATURE SURVEY

Author in [1] has propose a new algorithm called MINIT (MINimal Infrequent Itemset), which is a used for mining minimal infrequent itemset (MIIs) [1]. This is the first algorithm for finding rare itemset. The itemset that satisfy a maximum support threshold and does not contain any infrequent subset, from transactional data set. It is based on SUDA2 algorithm. A Dataset property is to consider the matrix form is a main difference between MINIT and SUDA2. Matrix consists of binary entries for traditional itemset mining. But for SUDA2, the matrix entries can contain any integer. Minimal infrequent itemset problem is NP-complete.

Author in [2] has proposed a measure called w-support, which is used to find weight of itemset and weight of transaction does not require preassigned weight. These weights are completely based on internal structure of the database. HITS model and algorithm are used to derive the weights of transactions from a database with only binary attributes. A new measure w-support is defined to give significance of itemset and it differs from the traditional support in taking the quality of transactions into consideration based on these weights. An apriori-like algorithm is proposed to extract association rules whose w-support and w-confidence are above some give thresholds.

Probabilistic frequent itemset mining in uncertain transactional database. Based on world semantics Thomas Bernecker in [3] introduces new probabilistic formulations of frequent itemsets. In a probabilistic model, an itemset is said to be frequent if the probability that itemset happens in at least min support is higher than that of given threshold. In addition to probabilistic model, framework is presented which has a capacity to solve the Probabilistic Frequent Itemset Mining (PFIM) problem powerfully.

Based on interest/intensity of the item within the transaction Wei Wang in [4] has proposed by allowing weight to be associated with each item within the transaction. In turn, to associate a weight parameter with each item in a resulting association rules. Then it is called as weighted association rule (WAR). For example, bread[4,6] → jam[3,5] is a weighted association rule indicating that if a customer purchase a bread quantity between 4 and 6 pack, he is likely to purchase 3 and 5 pack of jam. This method produces a higher quality results than previous known method on quantitative association rules.

Author in [5] present SUDA2 a recursive algorithm for finding Minimal Sample Uniques (MSUs). It uses a new method for demonstrating the search space for MSUs and observe about the properties of MSUs to prune and traverse this space. It has a ability to identify the boundaries of the search space with an execution time which is numerous orders of magnitude faster than that of SUDA. SUDA2 is a good candidate for parallelism as a search can be divided up according to rank, it will produce efficient load balancing.

In weighted settings author in [6] deal with the problem of finding significant binary relationship in transactional dataset. In weighted association rule mining problem each item is allowed to have a weight. The main focus is to mining significant relationship relating items with significant weights rather than insignificant relationship. A new algorithm called WARM (Weighted Association Rule Mining) is developed. This algorithm proposed a “weighted downward closure property” as a substitution of original “downward closure property”. Weighted downward closure is a idea of replacing support with significance is proposed. WARM algorithm is both scalable and efficient in discovering significant relationship in weighted settings.

Ashish Gupta, Akshay Mittal, Arnab Bhattacharya [7] propose a new algorithm based on pattern-growth paradigm for finding minimally infrequent itemset. They introduce a new concept called residual tree. To mine a multiple level minimum support itemsets ,where different threshold are used for finding frequent itemsets for different lengtha of the itemset by using residual tree. For mining minimally infrequent itemsets (MIIs) author [7] introduce a new algorithm called IFP min. Here Apriori algorithm is proposed to find MIIs. Extension of the algorithm is designed for finding frequent itemset in the multi level minimum support (MLMS) model.

Luca Cagliero and Paolo Garza [8] deal with the issues of discovering rare and weighted itemsets, i.e., the infrequent itemset mining problem. Finding rare data correlations is more interesting than mining frequent ones. [8] is like FP-growth algorithm. The IWI-support measure is defined as a weighted frequency of occurrence of an itemset in the analyzed data. Occurrence weights are derived from the weights associated with items in each transaction by applying a given cost function. They mainly focuses on IWI- support-min measure and IWIsupport-max measure. Author proposes two IWI mining algorithm that carry out IWI and Minimal IWI mining efficiency. A result shows the efficiency and effectiveness of the proposed approach.

In [4] author focus on finding association among frequent itemsets. X. Wu, C.Zhang and S.zhang designed a new method for mining both positive and negative association rules efficiently. This approach is new and different from existing approach i.e., association analysis. This method reduces the search space and had used the increasing degree of the conditional probability to assess the confidence of positive and negative association rules. This result shows that the proposed approach is effective, efficient and promising [4].

Negative association rule (NAR) is used for discovering of interesting pattern during mining process. Apriori algorithm is used for mining negative association rule for frequent absence and presence (FAP) itemset. Association rule mining only explores positive relationship in the beginning. Positive relationship implies the purchase one item or itemset with the purchase of another item or itemset. Negative relationship implies the presence of items by the absence of other item in the same transaction. FAP itemsets effectively generate optimum number of rules including NAR compared to others. The search space can be significantly reduced in this approach.

III. COMPARATIVE TABLE

Sno	Author	Techniques	Merit	Demerits
1	Jiawei Han, Jian Pei, and Yiwen Yin	Frequent pattern tree (FP-tree) structure	Scalable and efficient result is achieved.	Computational overhead is increased
2	David j. Haglin and Anna M. Manning	Minimal infrequent itemset Mining	Search complexity is lower and better performance is obtained.	Running time and computational complexity is increased.
3	Wei Wang, Jiong Yang and Philip S. Yu	Weighted Association Rule (WAR)	Produce Higher quality result than previous known method on quantitative association rules.	Performance of the system is degraded in this system.
4	Feng Tao, Fionn Murtagh, Mohsen Farid	Weighted Association Rule Mining (WARM)	Scalable and efficient in discovering significant relationship in weighted settings.	Not reliable for high dimensional data.Thus the reliability of the system is reduced.
5	Ashish Gupta, kshay Mittal and Arnab Bhattacharya	Pattern-Growth and residual Tree	Performance is improved with less computation time.	For mining maximal frequent itemsets better scalability is not achieved.
6	A.M. Manning and D. J. Haglin	SUDA2	Capable of identifying the boundaries of the search space. SUDA2 is good candidate for parallelism.	Load balancing problem is occurred in this system.

IV. CONCLUSION

Frequent Itemset Mining has attracted plenty of attention but much less attention has been given to mining Infrequent Itemsets. This survey is focused on infrequent weighted itemset. Occurrence weights derived from the weights associated with each items in transaction and applying a given cost function. The related concepts of positive and negative correlated pattern and its association rules are mined. The major advantage for mining infrequent itemset was to advance the profit of rarely originated datasets in the transactions. Merits and demerits of each method are described in comparative table to efficiently differentiate the each methods functionality.

REFRENCES

[1] D. J. Haglin and A.M. Manning, “On Minimal Infrequent ItemsetMining,” Proc. Int’l Conf. Data Mining (DMIN ’07), pp. 141-147, 2007

[2] K. Sun and F. Bai, “Mining Weighted Association Rules Without Preassigned Weights,” IEEE Trans. Knowledge and Data Eng.,vol. 20, no. 4, pp. 489-495, Apr. 2008.

[3] C.-K. Chui, B. Kao, and E. Hung, “Mining Frequent Itemsets from Uncertain Data,” Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD ’07), pp. 47-58, 2007.

[4] W. Wang, J. Yang, and P.S. Yu, “Efficient Mining of Weighted Association Rules (WAR),” Proc. Sixth ACM SIGKDD Int’l Conf. Knowledge Discovery and data Mining (KDD ’00), pp. 270-274, 2000.

[5] A. Manning and D. Haglin, “A New Algorithm for Finding Minimal Sample Uniques for Use in Statistical Disclosure Assessment,” Proc. IEEE Fifth Int’l Conf. Data Mining (ICDM ’05), pp. 290-297, 2005.

[6] F. Tao, F. Murtagh, and M. Farid, “Weighted Association Rule Mining Using Weighted Support and Significance Framework,” Proc. nineth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’03), pp. 661-666, 2003.

[7] A. Gupta, A. Mittal, and A. Bhattacharya, “Minimally Infrequen Itemset Mining Using Pattern-Growth Paradigm

and Residual Trees," Proc. Int'l Conf. Management of Data (COMAD), pp. 57-68, 2011.

[8] Luca Cagliero and Paolo Garza," Infrequent Weighted Item set Mining Using Frequent Pattern Growth" IEEE Transactions On Knowledge And Data Engineering, Volume 26, No.4, April 2014

[9] X. Wu, C. Zhang, and S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules," ACM Trans. Information Systems, vol. 22, no. 3, pp. 381-405, 2004.

[10] Anis Suhailis Abdul Kadir, Azuraliza Abu Bakar, Abdul Razak Hamdan, "Frequent Absence and Presence Itemset for Negative Association Rule Mining," 11th International Conference On Intelligent System Design and Applications, 2011.

[11] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic Frequent Itemset Mining in Uncertain Databases,"Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '09), pp. 119-128, 2009.

[12] X. Dong, Z. Zheng, Z. Niu, and Q. Jia, "Mining Infrequent Itemsets Based on Multiple Level Minimum Supports,"Proc. Second Int'l Conf. Innovative Computing, Information and Control (ICICIC '07), pp. 528-531, 2007

[13] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 1-12, 2000.

[14] G. Cong, A.K.H. Tung, X. Xu, F. Pan, and J. Yang, "Farmer: Finding Interesting Rule Groups in Microarray Datasets," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '04), 2004.

[15] Mehdi Adda, Lei Wu, Sharon White(2012), Yi Feng " Pattern detection with rare item-set mining" International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.1, No.1, August 2012.

[16]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

[17] R. Agrawal, T. Imielinski, and A. Swami, ―Mining association rules between sets of items in large databases,‖ in Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93), May 1993, pp. 207–216.

R. PRIYANKA, ME (Computer Science & Engineering) Kumaraguru College Of Engineering and Technology, India .

Mr.. S. P. SIDDIQUE IBRAHIM, Assistant Proffessor (Department of Computer Science & Engineering) Kumaraguru College of Engineering and Technology, India