Comparative Analysis Of Non-Frequent Pattern Mining Approach

Karamjit Kaur, Rajeev Bedi, R.C.Gangwar

BCET, Gurdaspur, India
kammy_walia@yahoo.co.in, rajeevbedi12@gmail.com, rakeshgangwar@gmail.com

ABSTRACT: Data mining has many aspects like clustering, classification, anomaly detection, association rule mining etc. Among such data mining tools, association rule mining has gained a lot of interest among the researchers. Some applications of association mining include analysis of stock database, mining of the web data, diagnosis in medical domain and analysis of customer behaviour. In past, many algorithms were developed by researchers for mining frequent itemsets but the problem is that it generates candidate itemsets. So, to overcome it tree based approach for mining frequent patterns were developed that performs the mining operation by constructing tree with item on its node that eliminates the disadvantage of most of the algorithms. The paper tries to address the problem of finding frequent itemset by determining the infrequent itemsets in a transaction which would reduce the computation time. The proposed algorithm is compared with the existing weighted mining algorithm for performance evaluation.

Keywords: Association rule mining, Data Mining, Frequent pattern Mining, Infrequent weighted itemset, Weighted mining.

1 INTRODUCTION

Data mining has become an important as necessity of extracting the meaningful information from data has gained advantage for decision making and behavioral analysis [1]. It also focuses on analysing the relationship among the data and finds the hidden patterns in the data. The knowledge obtained with help of data mining tools can be utilized for solving complex problems such as detection of fraud identification to enhance customer buying behaviour. As most of the users are not professionally trained to analyse the patterns of the data, data mining tool in such cases resolve the problem to identify patterns for better decision making. The problem of rule extraction introduced in 1993 by Agrawal et.al [1] as stated below: Let *Item* be a set of items. A set $Y = \{item_1. \ . \ . \ item_k\} \subseteq Item$ is called an itemset, or an n-itemset if it contains n items. A transaction over *Item* is *T = (ID, Item)* where *ID* is the transaction identifier and *Item* is an itemset. A transaction *T = (ID, Item)* is said to support an item set $Y \subseteq Item$, if $Y \subseteq Item$. A transaction database *TD* is a set of transactions. If the support of an item *sup_item* is greater than the specified user-defined threshold value *thres,* the item is considered as frequent itemset in the data. The association rule mining helps in identification of the rule with the interestingness for decision making and market analysis. The need of rule mining becomes important in every sector. The availability and high dimensionality of data becomes a problem for finding the rules. Therefore, the large databases with various techniques for easy handling and so that extraction of frequent patterns can be done easily. Assigning maximum and minimum occurrence for the itemset [2] itself filters out the items with certain threshold value. Hence the frequent items set needs to be preserved for finding optimal items set.

2 LITERATURE REVIEW

In (2011), Weimin Ouyang et al. [3] illustrated three drawbacks existing in some of the conventional algorithms such as quantitative databases are not considered in some cases, the detection of the frequent itemsets is based on the minimum support that falsely accumulates the frequency of items to be same. Finally some algorithms performs mining where simple rule are obtained. To overcome these limitations, rule mining is done by assigning fuzzy value to the data items that have multiple support values. Ashish Gupta et al. (2011), [4] proposed a concept of tree-growth algorithm, that obtains minimum number of infrequent items. This approach is accomplished by the application of the residual trees that mines the database at each level with given threshold value. It makes sense to generate only minimal infrequent itemsets, i.e., those which are infrequent but whose all subsets are frequent. This approach extracts the minimum number of correlated infrequent itemsets for analysing large amount of data. The tree construction increases computational time of the mining process. In (2012), Yihua Zhong et al. [5] introduced an algorithm based on the concept of weighted dual confidence for extracting efficient weighted rules in the database because the traditional association rule approaches are based on the support and confidence metrics with attributes considering an equal weight, resulting in ineffective rules which are not suitable for taking decision making. This method extracts interesting negative association rules from the database through eliminating the meaningless association rules. In (2012), K.Suriya Prabha and R.Lawrance [9] recommended that the fuzzy set concepts integrate with the negative association rules for constructing a sub-tree which generates candidates from tree. This method takes less amount of time to compute, memory space and efficient search process for obtaining fuzzy frequent itemsets. [7] The fuzzification of itemsets of the data is distinguished into fuzzy regions improves the search process of data. The fuzzy FP tree construction does not require multiple scanning of the data resulting less amount of time computation. However, transformation of itemsets into the fuzzy regions results in more time complexity. In (2013), Johannes K. Chiang et al. [12] underlined some drawbacks in conventional mining techniques such as these can perform the mining process based on a predefined schemata, therefore scanning is required for addition of new attributes. Since the rule mining can be properly decided based on certain level, they are designed for extracting either frequent or infrequent rules. The author overcame these limitations by suggesting the concept which is used as a data structure for representing associations patterns of the data in the database. Therefore, the process is designed to integrate the itemsets for obtaining the large itemsets, updating and obtaining the resulting patterns. In (2014), Shipra Khare and Prof. Vivek Jain [12] proposed a mining of infrequent weighted itemset which provides less amount of computational time. This approach plays a

significant role in decision-making. To discard the infrequent rules, the support and confidence threshold values are necessary parameters that are required to the mine algorithm to avoid generating misleading association rules. Therefore, this algorithm is useful for discovering minimum number of nodes based on support and confidence values obtains interesting weighted negative association rules from the database. This approach extracts minimum number of frequent itemsets for analysing large amount of data. However the resultant negative association rules from the original transactional data cannot be recovered. In (2012), Christian Borgelt, et al. [13] highlighted that in traditional mining process, where a transaction is considered if the frequency of occurrence of item is more. The resulting item set is defined as boundary based decision item sets. But finding the significant groups of items in data rendering a challenging issue. To overcome these drawbacks, the paper described two approaches one is for obtaining item sets by extending the mining process over the similar attributes and measures the frequent distribution with the Eclat algorithm and the other approach is a clustering methodology based on the distance metric of the data. In (2013), Sujatha Dandu, et al., from the existing literature, it is observed that the existing Apriori approach for mining uncertain data requires huge computational time and cost [14]. The combined approach mainly focuses on mining of the patterns initially through the FP-growth approach followed by the Apriori algorithm. The mining of the frequent items is done following the Apriori property. The sub frequent items lead to the generation of the frequent items which are also frequent. This approach obtains the correlated item sets in less computational time. The multiple database scan is reduced in this approach. However, the approach follows top-to-down mining process, requires previous information for each node for retrieving the correlated item set. In (2013), Sowan et al. the paper describes the application of the fuzzy concept for enhancing the prediction and evaluate the performance [15]. Fuzzy C means approach is considered for formation of fuzzy sets and the identification of the rules are done with the Apriori Algorithm. The basic ideology of the minimum support count is replaced with multiple support values. The paper experimentally shows efficiency of proposed approach for prediction of future values than the existing approaches. In (2012), Zahra Farzanyar and Mohammadreza Kangavari, deal with algorithm for handling the large datasets [16]. The fuzzy rules are generated in the process. The mining process generates a huge number of candidate sets and handling such large number of candidate sets becomes difficult, results in ineffective for analysis the data. In the proposed methodology, these large numbers of items are pruned on the basis of behavioral attributes and ontology. Similar behavioral characteristics analyze the huge amount of data in a less computational time. As the correlated attributes are recognized, therefore the datasets get reduced which are applied for the mining process results in less amount of time complexity. As the membership of the methodology is dependent on the support value defined the user for pruning the dataset, hence the tuning of parameter should be done for obtaining set of association rules. In (2013), R. Prabamanieswari [17], the paper partitions the numerical dataset into fuzzy regions and forms cluster-based fuzzy set. Fuzzy frequent itemset is found from the fuzzy regions with application of the cluster-based approach. The paper also performs a comparison among the proposed and existing algorithms. The fuzzification of itemsets of the transactional database is differentiated into fuzzy regions resulting in database compression by providing a threshold value. The fuzzy FP tree construction does not require multiple scanning of the database which reduces the time complexity. The fuzzification of itemsets into the fuzzy regions is a time computing task.

Algorithm

The discovery of infrequent items from the set of items that are less than the specified threshold value have gained a lot of interest in current days. For obtaining the infrequent itemset from the weighted transactional dataset a weighted support measure has been considered for data pruning. The min-support threshold and max-support threshold are obtained from the weighted transactional database. The min-support threshold represents the minimum value of the item existing in the database. The max-support threshold represents the maximum value of occurrence of the item in the transaction. The minimum support value is defined as the observed frequency where each transaction item support isthe chosen weighted maximum or minimum function. An equivalent weighted transactional dataset is the union of all equivalent transactional set associated with each weighted transaction. The algorithm initiates with pruning of the items with the maximum threshold value. The mining procedure is similar to the FP-growth algorithm that performs projection on itemsets. The tree construction and recursive mining process is continued for the mining process. The equivalent dataset is generated for insertion to the FP-tree with the weighted assigned to each data item. However, to reduce the increasing time complexity the FP-tree pruning with weighted threshold are applied for pruning the data set that discards the items at beginning of the tree construction. As the tree-construction is done the miner is initiated for obtaining the itemsets and corresponding minimum patterns. In association rule mining, each itemset has certain occurrence frequency which could be termed as the weight of the itemset [19]. The weight could be positive, negative or null. Mining of such weighted transactional datasets for finding frequent patterns are called weighted itemsets. In state of art of the infrequent itemset mining algorithms, the ability of taking the small frequent itemset into consideration is negligible. The infrequent mining of the item sets is a FP-Growth algorithm finding the infrequent items from the given set of frequent items [7][9]. The algorithm performs the mining process in two steps: first by calculating the infrequent items by pruning the transactional database given the user-defined support values. The mining process is similar to the FP-Tree approach where the weight is associated with each item in the transactions. Apart from sorting of the items based on their support value, the algorithm sorts the items with their associated weight factor. An item in FP-Tree is pruned if it appears only in this tree paths from the root to a leaf node characterized by a weighted support value greater than a predetermined threshold value [8]. The pruning process continues until all the nodes are encountered, finally resulting in interested infrequent item sets.

3 PROPOSED WORK

3.1 Problem Definition

In association rule mining, each itemset has certain occurrence frequency which could be termed as the weight of

the itemset. The weight could be positive, negative or null. Mining of such weighted transactional datasets for finding frequent patterns are called weighted itemsets. In state of art of the infrequent itemset mining algorithms, the ability of taking the small frequent itemset into consideration is negligible. Finding the frequent itemset can be negated for the entire transactional dataset which would result in finding the infrequent patterns. The support measure in most of the algorithms treats the itemset equally even though they don't have the same relevance in the dataset.

3.2 Advantages of Proposed Methodology

In the existing approach the weights are assigned with local interestingness whereas in proposed approach the weights are assigned by a fuzzy membership value having a sigmoid distribution. In existing method, the support value is a user-defined value and in proposed method the support is calculated with max-min normalization, which makes the proposed methodology more independent of any parameter value. In existing approach the FP tree mining is done which would take large space and time complexity. The tree compression is done to reduce the space and time complexity by merging the similar path for a data item.

3.3 Existential Probability

The transactional data is characterized by the probability of existence of the item in the transactional database. Each transaction in the database contains some existential probabilities. The existential probability P(x, ti) of an item x in a transaction ti indicates the likelihood of x being present in transaction ti. There are two possible interpretation of an item x and a transaction ti: (i) the possible world W1 where x$\in$ ti and (ii) the possible world W2 where x'$\in$ ti. Although it is uncertain which of these two worlds be the true world, theprobability of W1 be the true world is P(x, ti) and that of W2 is 1 −P(x, ti).

3.4 Methodology

The proposed methodology encapsulates the probability of existence of a data item in a transaction. This encapsulation is achieved through the fuzzy conversion of the data items support count. As compared to the conventional method the existence of any data item can be either 0 or 1. A fuzzy decision value represents the degree of possibility of the existence of a data item which can be any value between the ranges of 0 to 1. The proposed methodology explores the set of infrequent items for finding the frequent itemset. The different steps of the proposed methodology are detailed below:

- The input data is scanned once for the support value to prune the data items that satisfy the minimum support value. In the transactional dataset, the maximum and minimum support value is computed for finding the support for the entire procedure.
- The transactional data is sorted according to the support count values that are pruned after the minimum support count obtained. The obtained sets of items are fuzzified with sigmoidal membership value since this membership is a smooth curve of value ranging from 0 to 1.
- The fuzzified database is now considered for the construction of the FP-tree. The tree construction is done from the database having the item and its corresponding weights. From the tree the similar paths are merged for obtaining the compact tree which reduces the computation time for mining the tree as well as the space complexity for storage of data.
- Finally the mining process is performed in the bottom-up manner which retrieves the paths for each item that satisfies the minimum support count for that item.
- Each conditional FP tree is generated satisfying the minimum support count. The conditional FP tree generation is repeated until all the combinations are attended.

3.5 Proposed Algorithm

Input: Transactional Database D
Output: Frequent itemset

Step1: Transform D into fuzzy value with mem_fn

Step2: Count the support of each item in for a given fuzzy value

Step3: Find the normalized supp_value of in for D

Step4: Compute the support of transactions supp_value and prune In>supp_value

Step5: Arrange the In as order of pruned in values

Step6: Construct the FP-tree with root and items In

Step7: For each level L Find similar transactions

Step8: Reduce the size of D by compressing obtained similar transactions

Step 9: Obtained the sub-conditional patterns of each item from the resultant Compact FP-Tree

Step10: Check for the Computed Support value of the item with the supp_value

Step11: Output the optimum Frequent Items

4 RESULTS AND ANALYSIS

All the experiments were performed on 2.65 GHz Intel core processor with 2 GB RAM running Windows 7. The algorithms were implemented in C programming. The Connect database having 67557 instances and 42 attributes in UCI Repository is considered. The infrequent itemsets drawn from the dataset is dependent of the support values. As the frequent itemsets negates, the infrequent items, resulting in increase of support value also increases the infrequent transaction count. Fig. 1. shows the graphical representational of the support values and the infrequent transaction count. In the plot for existing method the increase in support value results in a small range of variation among the transaction count whereas in the

proposed method with the increase in the minimum support value the number of transactions is large in such case. Fig. 2. plots the execution time of the algorithm with the variation of the support values. The execution time of the proposed method is reduced with introduction of the fuzzy parameter for support values as well as the assignment of the weight for each data item in the transaction is fuzzified as the existential probability. The execution time of the existing method is approx 45 sec with a support value of 50 whereas in the proposed method the execution time gets reduced to 20 sec. The execution time is reduced by half the existing method that results in outstanding performance of the proposed methodology.

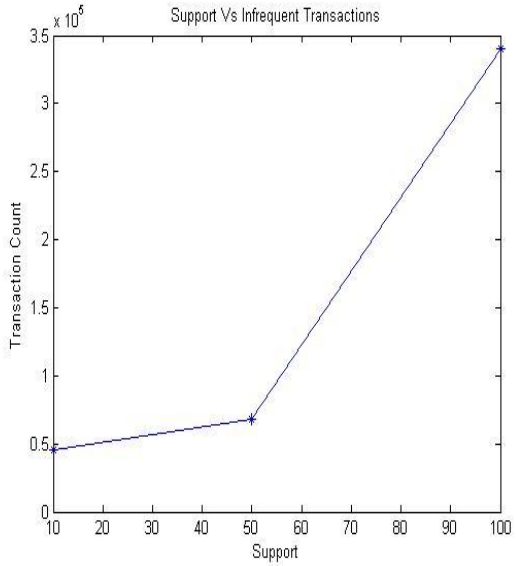

Fig.1 *Existing Method*

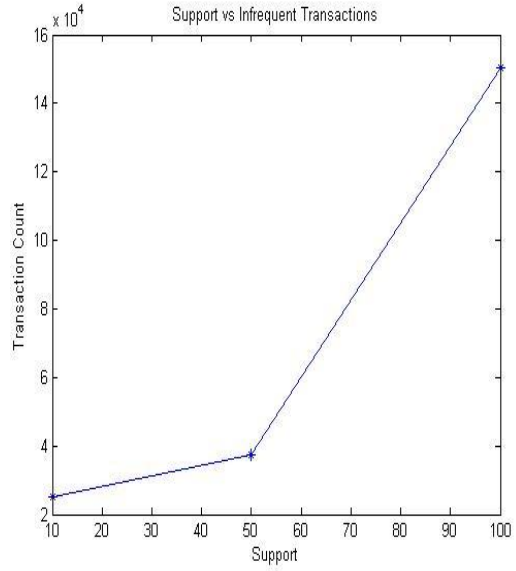

Fig.2. *Proposed Method*

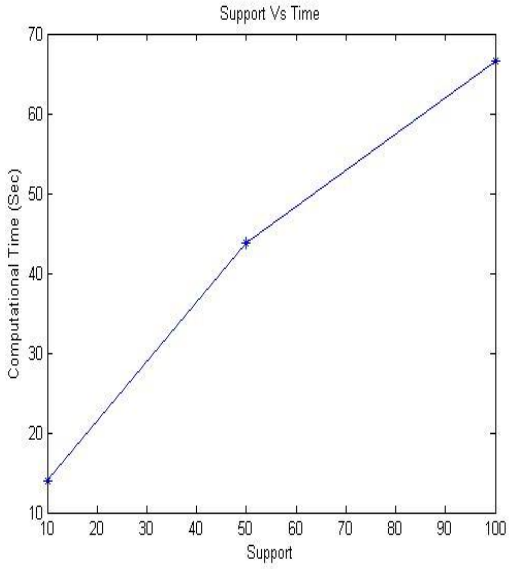

Fig.3. *Existing Method*

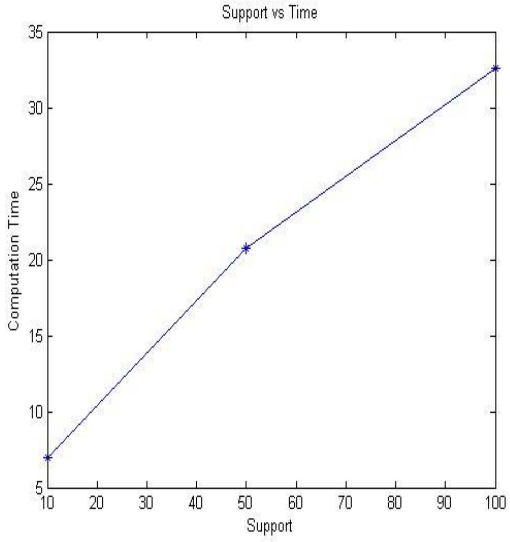

Fig.4. *Proposed Method*

5 CONCLSUION

In this paper, we have considered the important factors such as time and memory consumption for finding the infrequent itemsets. The literature survey reviews the performance of the various frequent mining algorithms on the basis of the approach made in each algorithm and the data set on which they are applied. The FP-tree mining algorithm and the Apriori algorithm are some of the benchmark algorithm in frequent pattern mining. We have analysed the drawbacks and tried to propose a new approach that overcomes the drawbacks of conventional frequent itemset mining process. The number of transaction count and execution time of the proposed method have been demonstrated as better than the existing method.

ACKNOWLEDGMENT

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

REFERENCE

[1] Ravichandran, I. 2003, Data mining and clustering techniques, Technical Report.

[2] J Han, M Kamber, “Data mining: Concepts and techniques 2nd edition”, Morgan Kaufman Publishes, 2006.

[3] 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.

[4] Ashish Gupta, Akshay Mittal and Arnab Bhattacharya, “Minimally Infrequent Itemset Mining using Pattern-Growth Paradigm and Residual Trees”, Proceedings of the 17th International Conference on Management of Data, pp.57-68,2011.

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

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

[7] 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.

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

[9] K.Suriya Prabha and R.Lawrance, “Mining Fuzzy Frequent itemset using Compact Frequent Pattern (CFP) tree Algorithm", International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.

[10] Xiao FengZheng 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.

[11] AnjanaGosain and ManeelaBhugra, “A Comprehensive Survey of Association Rules On Quantitative Data in Data Mining”, IEEE Conference on Information and Communication Technologies, 2013.

[12] Shipra Khare and Prof. Vivek Jain, “A Review on Infrequent Weighted Itemset Mining using Frequent Pattern Growth”, International Journal of Computer Science and Information Technologies, Vol. 5 (2) , pages1642-1647,2014.

[13] Christian Borgelt, Christian Braune, Tobias Kotter and Sonja Grun, “New Algorithms for Finding Approximate Frequent Item Sets”, Journal of Soft Computing, vol - 16, issue 5, pp. 903-917, Springer-Verlag, 2012.

[14] Sujatha Dandu, B.L. Deekshatulu & Priti Chandra, “Improved Algorithm for Frequent Item sets Mining Based on Apriori and FP-Tree”, Global Journal of Computer Science and Technology Software & Data Engineering, vol. 13 no. 2, 2013.

[15] Sowan, Bilal, Dahal, Keshav, Hossain, Alamgir, Zhang, Li and Spencer, Linda, “Fuzzy association rule mining approaches for enhancing prediction performance”, Expert Systems with Applications, vol. 40 no.17. pp. 6928-6937, 2013.

[16] Zahra Farzanyar and Mohammadreza Kangavari, “Efficient Mining of Fuzzy Association Rules from the Pre-processed Dataset”, Computing and Informatics, vol. 31, pp. 331–347, 2012.

[17] R. Prabamanieswari, “A Combined Approach for Mining Fuzzy Frequent Itemset”, International Journal of Computer Applications (0975 – 8887), 2013.

[18] Amir Ebrahimzadeh and Reza Sheibani, “Two Efficient Algorithms for Mining Fuzzy Association Rules”, International Journal of Machine Learning and Computing, vol. 1, no. 5, 2011.

[19] Luca Cagliero and Paolo Garza, “Infrequent Weighted Itemset Mining using Frequent Pattern Growth”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, no.1, pp.1, 2013.