International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

An Efficient Clustering Based High Utility Infrequent Weighted Item Set Mining Approach

Dr. N. Umadevi[1], A. Gokila Devi[2]

[1, 2]Research Supervisor, Department of Computer Science And Information Technology, Sri Jayendra Saraswathy Maha Vidyala College of Arts & Science, Coimbatore -5

Abstract: ***The conventional data mining techniques are mainly focused on discovering the correlation between items that are more frequent in the transaction databases. In recent years, the research area has focused on infrequent itemset mining whose frequency of occurrence is less than or equal to a maximum threshold. Most of the existing system introduced to mine infrequent item set. These methods do not consider the utility of item set. The main objective of this work is to find out the infrequent weighted item set from weighted transactional database and group them according to the utility value. An efficient high utility based clustering algorithm is used to group the high utility infrequent weighted item set by using k-mean clustering algorithm. The utility of items is determined by taking into account factors such as profit deal, temporal characteristics of items. The utility of weighted items is greater than the minimum utility threshold which is known as high Utility based Infrequent Weighted Item set. These item sets are clustered by using k-mean clustering algorithm. The proposed method achieves high performance in terms of scalability, accuracy and precision.***

Keyword: Association rule mining, support measure, utility function.

1. Introduction

Data mining is the process of finding out the data from various resources and summarizing it into precious information. Data mining tasks are to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers and clusters. Item set mining is one of the data mining methods which are used to find out the correlations among data. Discovering highly correlated item sets is known as frequent item set mining [1]. Items that are rarely found in the database are assumed to be uninteresting items and that are removed using the support measure. Such patterns are known as infrequent patterns [2].

In order to make use of weight in the mining process, several new concepts have been adapted. Support value is used in association rule mining [3], [4]. In weighted association rule mining (WARM), item sets are no longer simply counted as they appear in a transaction. This change of counting mechanism makes it necessary to adapt traditional support to weighted support. The goal of using weighted support is to make use of weight in the mining process and prioritize the selection of target item sets according to their significance in the dataset, rather than their frequency alone.

The trouble in mining infrequent patterns is:
(1) how to determine the interesting rare patterns, and
(2) how to professionally find out them in huge data sets.

The organization of this research work is given as follows: Section 1, detailed description about the introduction of the research work is given in this section. In section 2, various research works has been discussed in the detailed manner which was conducted in the previous work. In section 3, proposed work of this research methodology is discussed detailed. In section 4, performance evaluation tests were conducted with the consideration of the various performance measures. In section 5, overall research of this work is concluded.

2. Related works

Laszlo et.al [5] introduced a rare association rules to mine infrequent item sets. The proposed method is used to take out the rare association rules for mining frequent item set. The rare association rules are simply known as “mRI rules”. The support value of minimal rare item sets (mRIs) is calculated by Apriori algorithm. All restored infrequent items having two merits. First one is its highly informative and the second one is the amount of these rules is minimum.

Xindong Wu et.al introduced an efficient mining by utilizing both positive and negative association rules. The proposed system introduces a positive and negative rule for discovering frequent item set. They had also designed constraints for reducing the search space, and had used the increasing degree of the conditional probability relative to the prior probability for estimating the confidence of positive and negative association rules [6].

Abhang Swati Ashok et.al [7] proposed an Apriori algorithm which is used to Find Frequent Item Sets. The Apriori exploit a "bottom up" scheme, where frequent subsets are complete one item at a time, and collection of candidates is tested beside the data. The algorithm stops when no continuous winning extensions are found. The main concept of the Apriori Algorithm is to discover associations among various sets of data. It is occasionally referred to as "Market Basket Analysis". Each set of data set has a number of items which is known as transaction. The output of Apriori is sets of rules are referred how frequently items are enclosed in sets of data. However efficient item set mining is still investigated.

Luca Cagliero et.al [8] proposed a novel algorithm which is used to find out the Infrequent Weighted Item set from the transactional database. One of the famous data mining approaches is item set mining [9]. Infrequent Weighted Item set (IWI) and minimal Infrequent Weighted Item set (MIWI) mining algorithms are used to find out the rare item set in

the database. IWI-support-max measure and IWI-support-min measures are used to discovering an infrequent weighted item set. This system efficiently mine the rare item set from transaction database but it does not consider the utility of item set.

3. EHUC Algorithm based Infrequent Weighted Item Set Mining Approach

The proposed system consists of two phases. In the first phase, Infrequent Weighted Itemset Miner (IWI Miner) and Minimal Infrequent Weighted Itemset Miner (MIWI Miner) algorithms are used for mining Infrequent Weighted Itemset. In the second phase, efficient high utility based clustering (EHUC) algorithm is introduced to group the high utility infrequent weighted item sets. The utility of item is computed by using profit value of that item. The overall flow of the work is depicted in the following figure 1.

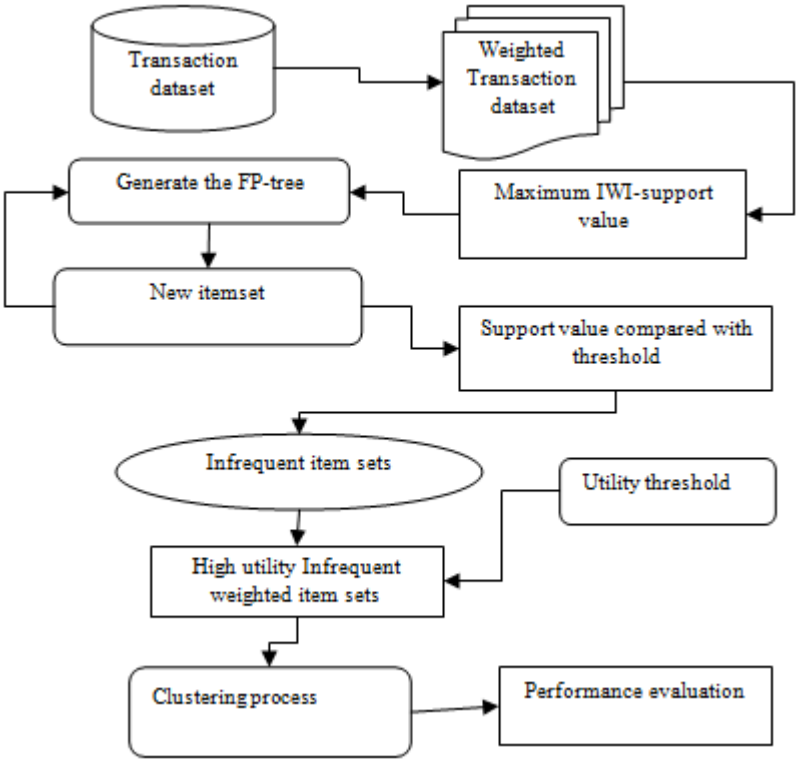

Figure 1: Overall Flow of the work

3.1 Weighted Transactional data base

The weighted transactional database (D) consists of number of transaction T. A number of item set is placed in every transaction and they are integrated with weight value.

$$T = \{t_1, t_2 \ldots t_n\} \quad (1)$$

$$I = \{i_1, i_2 \ldots i_n\} \quad (2)$$

$$W = \{w_1, w_2, \ldots . w_n\} \quad (3)$$

Where,
T - Set of transactions
I- Set of items
W- Set of weight value

The Item sets are mined from the weighted transactional data set is called weighted item set mining. Discovering item sets whose frequency of occurrence is less than or equal to a maximum threshold, which is known as infrequent item set mining. The trouble of mining item sets by considering weights associated with each item is called as the weighted item set mining problem. To overcome this mining problem an infrequent weighted item set mining and minimal infrequent weighted item set mining algorithms were introduced. It does not considered the utility of items and also not grouping the infrequent weighted item set which are having high utility.

3.2 Infrequent Weighted Items set Miner Algorithm

The IWI-support-min measure and IWI-support-max measure are used to find out the infrequent weighted item set. The IWI-support-min depends

on a minimum cost function, i.e., the occurrence of an item set in the transaction is weighted by the weight of its minimum interesting item. The IWI-support-max depends on a maximum cost function, i.e., the occurrence of an item set in the transaction is weighted by the weight of the huge interesting item. By using these IWI-support-min and IWI-support-max values, the Infrequent Weighted Item set Miner and Minimal Infrequent Weighted Item set Miner algorithms are discovering the rare item set on the transaction. These algorithms are FP-Growth-like mining algorithm. FP-tree creation and recursive item set mining are basic steps of FP-Growth-like mining algorithms. The Pruning process is

applied to remove frequent item set which does not satisfy the IWI-support threshold ξ.

Algorithm 1: IWI-Miner (T, ξ)
Input: T_w , ξ.
Output: Frequent item set
Step 1: Initialize D_T
Step 2: construction of FP tree
Step 3: Initialize T_w in FP tree
Step 4: Compute infrequent weighted item set mining
Step 5: Return IF
Step 6: End process
Infrequent weighted item set miner algorithm discovers the infrequent weighted item set by satisfy thresholdξ. To discover the minimal infrequent weighted item set, the same operation is performed by using MIWI miner algorithm.

Algorithm 2 IWI Mining (Tree, ξ)
Input: FP –tree, maximum IWI-support threshold
Output: F, the set of IWIs extending prefix
Step 1: Header table initialization
Step 2: for each item in HT
Step 3: compute new I with support value
Step 4 : items compared with threshold
Step 5: Select IF
// IF –I infrequent weighted item set
Step 6 : compute recursive mining

3.3 An efficient high utility based clustering (EHUC) algorithm

The utility of item is an important factor in real world situations. The utility of item sets is based upon user's perspective, such as cost, profit or revenue of significant importance. The proposed efficient high utility based clustering algorithm is used to find out the high utility Infrequent weighted item set by using minimum threshold values and user preferences. And grouping the high utility infrequent weighted item set by using k-mean clustering algorithm.

Algorithm 3:
Input: IWI
Output: group of high utility infrequent weighted item set
Step 1: IWI initialization
Step2: compute utility for each item set in IWI
$u(i_m, IWI) = l(i_m, IWI) \times p(i_m)$ (4)
// profit value of item i_m
Step 3: Calculate minimum utility
Minutil = $\delta \times \sum_{IWI} u(i_m, IWI)$ (5)
Step 5: if ($u(i_m) \geq$ minutil)
Step 6: Return high utility value
Step 7: Otherwise go to step 2
Step 8: Initialize μ_i
μ_i – Center of the cluster
Step 9: Compute mean of all High utility IWI belong to the cluster

$$\mu_i = \frac{1}{|c_i|} \sum_{j \in} c_i \; X_j \qquad (6)$$

X_j – Hgh utility items

Step 10: Group the high utility infrequent item set
If the utility of item is higher than the minimum utility ($u(i_m) \geq$ minutil) that can be considered as high utility weighted item set. The high utility value is computed from product of minimum utility threshold and utility of particular items in the tree. Then k-mean clustering algorithm is applied to cluster the high utility infrequent weighted item set.

4. Numerical Results

4.1 Data Set Description

The two data sets such as Real-life data sets and Synthetic data sets are taken for infrequent weighted item set mining. Real-life data set is used to validate the usefulness of the proposed algorithms technique. In Synthetic data sets also exploited a synthetic data set generator to estimate algorithm performance and scalability.

4.2 Result analysis

In the simulation, the existing IWI miner algorithms such as Infrequent Weighted Itemset Miner (IWI Miner) and Minimal Infrequent Weighted Itemset Miner (MIWI Miner) and the proposed EHUC algorithms are evaluated in terms of scalability, accuracy and Precision.

1. Scalability
The time taken for mining infrequent weighted items is known as Run time. The scalability is calculated for both existing approach and the proposed approach and compared in the following graph. And the actual run time that are obtained while mining infrequent weighted itemset are indicated in the table 1

Table 1: Scalability comparison

Number of Transaction	*Average Time Comparison*	
	IWI mining algorithms	*EHUC algorithm*
1000	10	5
2000	15	11
3000	35	26
4000	50	45
5000	59	55
6000	76	64

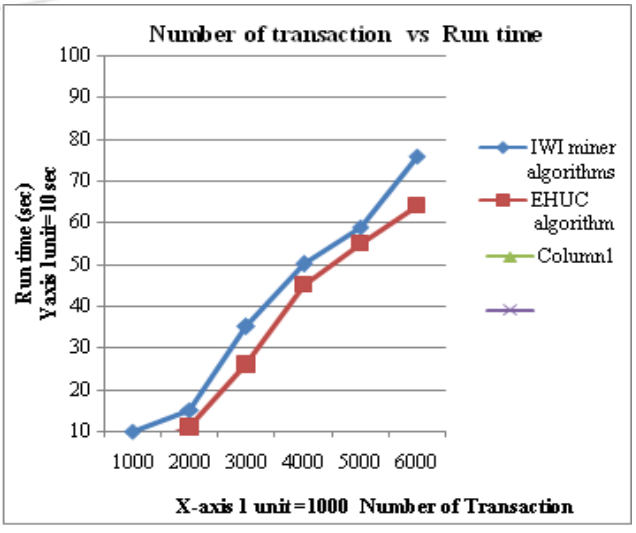

Figure 2: scalability comparisons

From the above graph it can be proved that the proposed methodology provides better results than the existing approach by discovering an infrequent item set. In this figure x axis plot the number of Transaction and y axis plot the Run time. The number of transactions is increasing the execution time of both algorithm are increased. As shown, the execution times of the proposed algorithm are effectively reduced.

2. Accuracy

Accuracy is defined as the degree of generating the experimental output that is matched with the expected output. The accuracy values that are obtained while mining infrequent weighted itemset are indicated in the table 2.

Table 2: Accuracy Comparison

Number of Transaction	*Accuracy Comparison*	
	IWI mining algorithms	*EHUC algorithm*
1000	10	15
2000	21	25
3000	33	38
4000	55	60
5000	74	79
6000	85	92

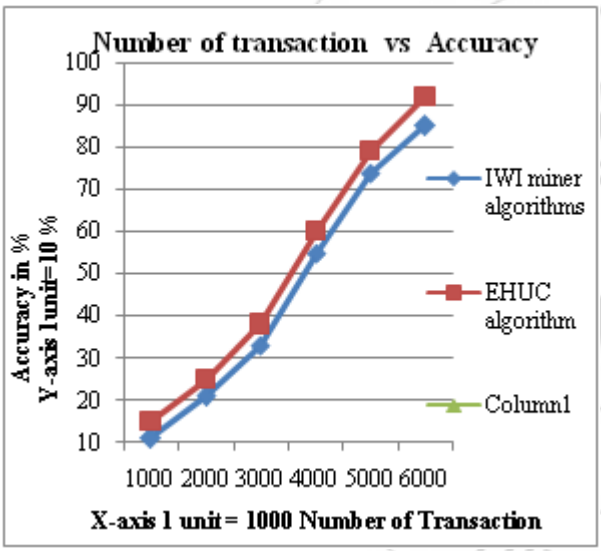

Figure 3: Accuracy comparisons

From the above graph it can be proved that the proposed methodology provides better results than the existing approach by discovering an infrequent item set. In this figure x axis plot the number of Transaction and y axis plot the Accuracy. The number of transactions is increasing the Accuracy of both algorithms are increased. From figure 2, the accuracy of the proposed algorithm is increased effectively.

3. Precision

Precision is defined as the Percentage of correct predicted results from the set of input transaction. The precision value should be more in the proposed methodology than the existing approach for the better system performance. The precision values that are obtained while mining infrequent weighted itemset are indicated in the table 3.

Table 3: Precision Comparison

Number of Transaction	*Precision Comparison*	
	IWI mining algorithms	*EHUC algorithm*
1000	0.2	0.3
2000	0.4	0.45
3000	0.5	0.56
4000	0.62	0.7
5000	0.7	0.8
6000	0.75	0.85

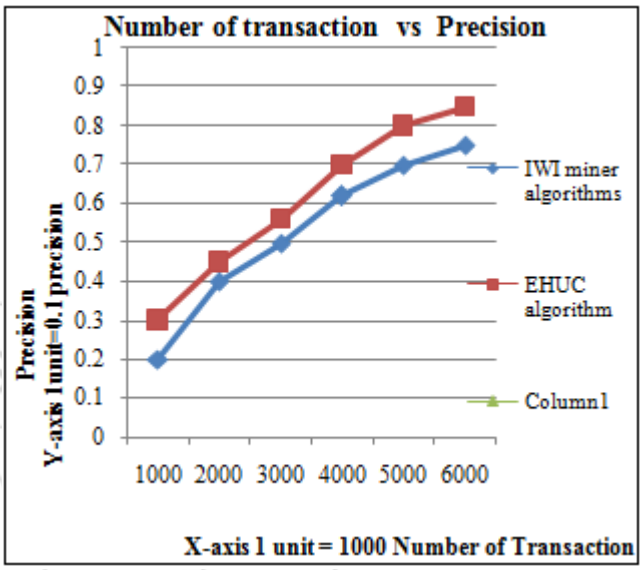

Figure 4: precision comparisons

In the above graph, the precision value is compared against the existing approach and the proposed approach. In the x axis numbers of transactions are taken. And in y axis, precision is taken. From this graph, it can be proved that the precision taken by the proposed methodology is high than the existing approach.

5. Conclusion

One of the most recent data mining research area is Utility Mining which prominence on all kinds of utility factors and incorporates the utility idea in data mining tasks. In this proposed system, we are evaluating the utility parameters of the rare item sets. The system considers a both the individual profit of each infrequent item and quantity of each one in the database simultaneously. The utility-based expressive data mining which aims at mining item sets whose utility value is high. It is termed as high utility weighted item set mining. Finally k mean clustering algorithm is used to group the infrequent weighted item set according to the utility of items. From the experimentation result, the proposed system can improve the performance of the system compared to the existing system.

In future, we can incorporate the IWI miner algorithm in an advanced decision making system that ropes domain expert's targeted actions based on the characteristics of the discovered IWIs. We can use the discovered infrequent patterns in the applications like error discovery and recovery, fraud recognition, finding outliers in the database.

References

[1] Sanjaydeep Singh Lodhi , Sandhya Rawat , Premnarayan Arya ," On demand efficient frequent itemset method in uncertain data, " Journal of Global Research in Computer Science research ,Volume 3, No. 7, July 2012 .
[2] Sakthi Nathiarasan, Kalaiyarasi, Manikandan, "Literature Review on Infrequent Itemset Mining Algorithms," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 8, August 2014.
[3] R. Agrawal, T. Imielinski, 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), pp. 207–216, May 1993,
[4] Suchismita Mishr, Pranati Mishr, "A Survey on Association Rule Mining Algorithms Performance Analysis," International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 8, 2014.
[5] Laszlo Szathmary, PetkoValtchev, Amedeo Napoli," Finding Minimal Rare Itemsets and Rare Association Rules," In Proceedings of the 4th International Conference on Knowledge Science, Engineering and Management (KSEM), 2010.
[6] Xindong Wu, C. Zhang, S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules," ACM Trans. Information Systems, vol. 22, no. 3, pp. 381-405, 2004.
[7] Abhang Swati Ashok, Jore Sandeep S., " The Apriori algorithm: Data Mining Approaches Is To Find Frequent Item Sets From A Transaction Dataset," International Journal of Innovative Research in Science, Volume 3, Special Issue 4, April 2014.
[8] Luca Cagliero, Paolo Garza," Infrequent Weighted Itemset Mining Using Frequent Pattern Growth, " IEEE transactions on knowledge and data engineering, vol. 26, no. 4, April 2014.
[9] Robin Singh Bhadoria , Rohit Bansal , " Analysis of Frequent Item set Mining on Variant Datasets", Vol 2 (5), pp.1328-1333, 2011.

Author Profile

N.Umadevi working as Head in the Department of Computer Science and Information Technology, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science. She has 3 years of industrial experience. Her area of interest are image processing and Data mining. Her publications include 8 international journals. She has presented papers in 6 international conferences and 4 national conferences.

A. Gokila Devi pursuing her M.Phil in Computer Science, Sri Jayendra Saraswathy Maha Vidhyalaya College of Arts and Science, Singanallur, Coimbatore. Under the guidance of Mrs. N. Umadevi working as Head of the Department, Department of Computer Science and Information Technology, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Singanallur, Coimbatore. Her area of interest is Data mining.