International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.2, 2015

RESEARCH ARTICLE

ISSN: 2321-7758

DISCOVERING RARE DATA SETS USING INFREQUENT WEIGHTED ITEMSET MINING

AYSHWARYA.R[1], DIVYA.M[2], DIVYA RAYAN.J[3], V.SATHYA PREIYA[4]

[1-3]Students, [4]Assistant Professor Department of Computer Science and Engineering
Panimalar Engineering College

Article Received: 13/03/2015 | Article Revised on; 19 /03/2015 | Article Accepted on:24/03/2015

ABSTRACT

To focus on the rare data sets mining, which has become an active research in data mining applications. When the need is to minimize cost and maximize profit, finding infrequent data correlations ,i.e., discovering itemsets whose frequency of occurrence in the analyzed data is less than or equal to a maximum threshold, is more effective than mining frequent data. Infrequent itemset discovery is applicable to data coming from different real-life applications. IWI miner algorithm is proposed which is designed specifically for finding rare itemsets Experimental results show efficiency and effectiveness of the proposed approach.

Index Terms: Clustering, classification, and association rules, data mining

I. INTRODUCTION

ITEMSET mining is an exploratory technique widely used for discovering valuable correlations among data due to its successful application in various data mining scenarios. The first attempt to perform itemset mining was focused on discovering frequent itemset, i.e.,p atterns whose frequency of occurrence in the analyzed data is above a given threshold. Frequent itemset find application in a number of real-life contexts like consumer market-basket analysis, inference of patterns from web page access logs and iceberg-cube computation. However, many traditional methods ignore the influence of each item or transaction within the analyzed data. To allow treating items differently based on their relevance, the idea of weighted itemset has been introduced. A weight is associated with each data item and characterizes its local significance within each transaction.

The significance of a weighted transaction is commonly evaluated in terms of the corresponding weights of the item. Furthermore, the main itemset quality measures have also been tailored to weighted data and used for driving the frequent weighted itemset mining process. In recent years, the attention of the research community has also been focused on the infrequent itemset mining problem, which is finding itemsets whose frequency of occurrence in the observed data is less than or equal to a maximum threshold. Infrequent itemset discovery is applicable to data coming from different real-life applications such as (i) mining of negative association rules from infrequent itemsets , (ii) statistical disclosure risk assessment where rare patterns in anonymous census data can lead to statistical disclosure , (iii) fraud detection where rare patterns in financial or tax data may suggest unusual activity associated with fraudulent behavior, and (iv)bioinformatics where rare patterns in microarray data may suggest genetic disorders. However,traditional infrequent itemset mining algorithms still suffer from their inability to take local item interestingness into account during the mining phase. Measures taken to mine frequent items are not applicable to find rare datasets and also it cannot handle weighted itemsets. This paper addresses the discovery of infrequent and weighted itemsets which has got wide range of applications in mining.

II. Related work

Frequent itemset mining is a widely used data mining technique that has been introduced. In the traditional itemset mining problem items belonging to transactional data are treated equally. To allow differentiating items based on their interest or intensity within each transaction, the authors focus on discovering more informative association rules, i.e., the weighted association rules (WAR),which include weights denoting item significance. However, weights are introduced only during the rule generation step after performing the traditional frequent itemset mining process. The first attempt to pushing item weights into the itemset mining process has been done. It proposes to exploit the anti-monotonicity of the proposed weighted support constraint to drive the Apriori-based itemset mining phase. However, weights have to be pre assigned, while, in many real-life cases, this might not be the case. To address this issue, the analyzed transactional data set is represented as a bipartite hub-authority graph and evaluated by means of a well known indexing strategy, i.e., HITS, in order to automate item weight assignment. Weighted item support and confidence quality indexes are defined accordingly and used for driving the itemset and rule mining phases. This paper differs from the above-mentioned approaches because it focuses on mining infrequent itemsets from weighted data instead of frequent ones. Hence, different pruning techniques are exploited. A related research issue is probabilistic frequent itemset mining. It entails mining frequent itemsets from uncertain data, in which item occurrences in each transaction are uncertain. To address this issue, probabilistic models have been constructed and integrated in Apriori-based or projection-based algorithms. Although probabilities of item occurrence may be remapped to weights, the semantics behind probabilistic and weighted itemset mining is radically different. In fact, the probability of occurrence of an item within a transaction may be totally uncorrelated with its relative importance. For instance, an item that is very likely to occur in a given transaction may be deemed the least relevant one by a domain expert. Furthermore, this paper differs from the above-mentioned approaches as it specifically addresses the infrequent itemset mining task.

III. Proposed work

In the proposed system, Infrequent Weighted Itemset Miner is used to discover rare weighted itemsets. To address this issue, the IWI-Support measure is defined as a weighted frequency of occurrence of an itemset in the observed data. Occurrence weights are derived from the weights associated with items in each transaction by applying a given cost function.

Data mining is the process of finding information where knowledge is gained by observing the data in very large repositories, which are analyzed from various views and the result is summarized into useful information. Due to the significance of extracting knowledge from the large data repositories, data mining has become a very crucial and guaranteed branch of engineering affecting human life in various spheres in direct and indirect means. Data mining is a technique that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. The first and simplest step in data mining is to describe the data, summarize its attributes, visually survey it using graphs and charts, and look for meaningful links among variables. Collecting, exploring and selecting the right data are critically important in the process of data mining.

The most widely used data mining techniques were mining frequent and probabilistic frequent itemset before effort has been made to discover rare correlations among data. Mining Frequent Itemsets over Uncertain Databases deals with two different definitions. The first definition, referred as the expected support-based frequent itemset, deals with the support of an itemset to measure whether the itemset is frequent. The second definition, referred as the probabilistic frequent itemset, uses the probability of the support of an itemset to measure its frequency. The definition of expected support-based frequent itemset uses the expectation to measure the uncertainty, which is an extension of the definition of the frequent itemset in deterministic data. The definition of probabilistic frequent itemset includes the complete probability distribution of the support of an itemset. An approach to exploit unweighted infrequent data in mining association rules has also been made in traditional model. But this paper is different from other approaches as it focuses only on mining weighted infrequent data sets over certain and uncertain databases.

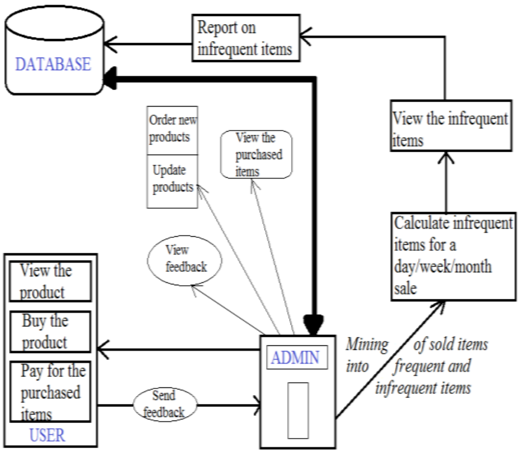
Weighted Association Rule Mining using Weighted Support and Significance Framework address the issues of discovering significant binary relationships in transaction datasets in a weighted setting. Traditional model of association rule mining is adapted to handle weighted association rule mining problems where each item is allowed to have a weight.

TABLE 1: Weighted transactional data set

Transactional dataset	Maximum IWI Support threshold		Observed data
date	customers	Product count	Item sets
-/-/-	A	25	j
-/-/-	B	63	l
-/-/-	C	55	b

In the proposed work, set of data is analyzed. Each itemset is associated with weight which differs based on each transaction. For instance, consider table 1 set of weighted itemsets on each transaction is observed. Transactional dataset (T), i.e., date is taken and count of customers and products is considered as maximum support threshold (ε). The conditions on maximum Infrequent Weighted itemset support threshold differ on each transactional dataset. Using IWI Miner algorithm, infrequent itemset (α) is mined from the overall sold items. IWI allows the expert to focus his attention on the underutilized or idle items and, thus, reduces the bias due to the possible inclusion of highly weighted items in the extracted patterns. Experiments, performed on both synthetic and real-life data sets, show efficiency and efficacy of the proposed approach. In particular, they show the characteristics and usefulness of the itemsets discovered from data coming from benchmarking and real-life systems, as well as the algorithm scalability

System Architecture


DATABASE
Report on infrequent items
View the infrequent items
Order new products
Update products
View the purchased items
Calculate infrequent items for a day/week/month sale
View feedback
View the product
Buy the product
Pay for the purchased items
USER
Send feedback
ADMIN
Mining into
of sold items frequent and infrequent items

Advantages

1. System resizing or resource sharing policy optimization.
2. Focus attention on the smallest sets that contain at least one underutilized/idle items.
3. Focus on cost minimization and profit maximization.
4. Infrequency leads to stock of expired items. Thus mining of infrequent items can minimize the expired stocks.
5. Has important usage in mining of negative association rules from infrequent itemsets.

Performance Analysis

We analyzed IWI Miner on standard synthetic and real data sets. In particular, we analyzed: (i) The impact of the equivalence procedure on the data set size (ii) the impact of the IWI-support thresholds on both the number of mined patterns and the algorithm execution time

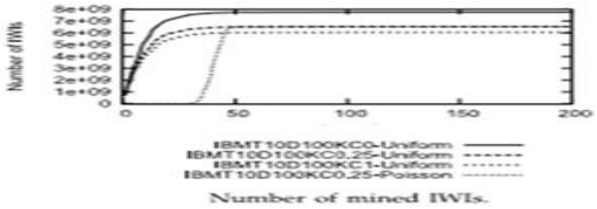


Fig.1.Impact of the maximum IWI-Support threshold on IWI Miner performance

Complexity analysis: The data set transformation procedure generates, for each transaction, a number of equivalent transactions at most equal to the original transaction length. A lower number of equivalent transactions are generated when two or more items have the same weight in the original transaction. The product of the original data set cardinality and its longest transaction length can be considered a preliminary upper bound estimate of the equivalent data set cardinality. However, in real data sets many transactions are usually shorter than the longest one and many items have equal weight in the same transaction. This reduces the number of generated equivalent transactions significantly. As confirmed by the experimental results achieved on real and synthetic data, the scaling factor becomes actually lower than the average transaction length, which could be considered a more realistic upper bound estimate. IWI Miner exploits the equivalence property, efficiently and effectively.

To reduce the complexity of the mining process, IWI Miner adopts an FP-tree node pruning strategy to early discard items (nodes) that could never belong to any itemset satisfying the IWI-support threshold. In particular, since the IWI-support value of an itemset is at least equal to the one associated with the leaf node of each of its covered paths, then the IWI-support value stored in each leaf node is a lower bound IWI-support estimate for all itemsets covering the same path.

Comparison with Traditional Non-Weighted Infrequent Itemset Mining

This paper is, to the best of our knowledge, the first attempt to perform infrequent itemset mining from weighted data. However, other algorithms are able to mine infrequent itemsets from unweighted data. Hence, to also analyze the efficiency of the proposed approach when tackling the infrequent itemset mining from unweighted data, we compared IWI Miner execution time with that of a benchmark algorithm, namely INIT. INIT is, to the best of our knowledge, the latest algorithm that performs unweighted infrequent itemset mining from unweighted data. For IWI Miner, we set all item weights to 1 in order to mine traditional (unweighted) infrequent itemsets. We compared IWI Miner and INIT performance, in terms of execution time, on synthetic and benchmark data sets with different characteristics.

In summary, when dealing with unweighted data (i) IWI Miner is shown to be orders of magnitude faster than state-of-the-art algorithms for all considered parameter settings and data sets, and (ii) IWI Miner is faster or competitive with state-of-the-art approaches, especially when setting lower maximum support thresholds or coping with denser data sets.

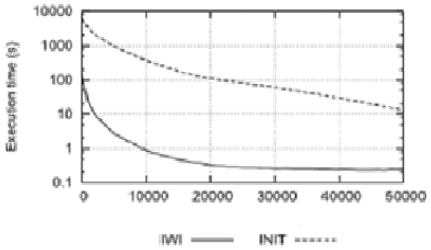
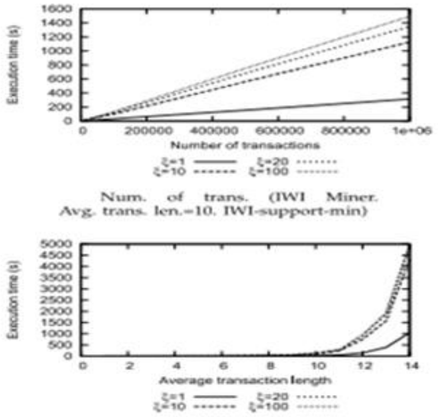


Fig.2.Comparison between IWI Miner and INIT in terms of execution time

Scalability Analysis

We also analyzed the algorithm scalability, in terms of execution time, on synthetic data sets. To test the algorithm scalability with the number of data set transactions, we generated data sets of size ranging from 0 to 1,000,000 transactions by following the procedure described. It reports IWI Miner execution times by varying the data set cardinality and by setting three representative IWI-support threshold values. The mining computational complexity appears to be strongly correlated with the cardinality of the extracted patterns. In general, IWI Miner execution time significantly increases for higher IWI-support threshold values due to the combinatorial growth of the number of extracted patterns.



Num. of trans. (IWI Miner. Avg. trans. len.=10. IWI-support-min)

Fig.3.IWI Miner scalability with different parameters. Correlation factor=0.25

IWI Miner execution time scale roughly linearly with the data set size for all the tested settings. We also analyzed the algorithm scalability, in terms of execution time, with the average transaction length. It reports the IWI Miner execution times by varying the transaction length and by setting three representative IWI-support threshold values. When increasing the average transaction length the algorithm execution time increases because of the non-linear increase of the number of possible item combinations.

IV. Methodology Used

Weighted Transaction Equivalence

The weighted transaction equivalence establishes an association between a weighted transaction data set T, composed of transactions with arbitrarily weighted items within each transaction, and an equivalent data set TE in which each transaction is exclusively composed of equally weighted items. To this aim, each weighted transaction $t_q \epsilon T$ corresponds to an equivalent weighted transaction set. Item weights in t_q are spread based on their relative significance, among their equivalent transactions in TE_q. The generated FP-table structure will be used to tackle the IWI mining problem effectively and efficiently.

Infrequent Weighted Itemset Miner Algorithm

Given a weighted transactional data set and a maximum IWI-support threshold, the Infrequent Weighted Itemset Miner algorithm extracts all IWIs whose IWI-support satisfies. IWI Miner is a FP-growth-like mining algorithm that performs projection-based itemset mining. Hence, it performs the main FP-growth mining steps: (a) Table creation and (b) recursive itemset mining from the database. Unlike FP-Growth, IWI Miner discovers infrequent weighted itemsets instead of frequent (unweighted) ones. To accomplish this task, the following main modifications with respect to FP-growth have been introduced: (i) A novel pruning strategy for pruning part of the search space early and (ii) a slightly modified structure, which allows storing the IWI-support value associated with each node. To cope with weighted data, an equivalent data set version is generated and used to populate the structure.

Algorithm Pseudo code

IWI-Miner (T, ε)

Input: T, a weighted transactional datasets

Input: ε, a maximum IWI support threshold

Output: α, the set of IWIs satisfying ε

```
α= null
/*Initialization */
countItemIWI-support(T)
Table <= a new empty table
for all weighted transaction $t_q$ in T do
    $TE_q$ <= equivalentTransactionSet($t_q$)
    for all transaction $te_j$ in $TE_q$ do
        insert $te_j$ in table
    end for
end for
α <= IWI Mining(Table, ε, null)
return α
```

Conclusion

This paper faces the issue of discovering infrequent item-sets by using weights for differentiating between relevant items and not within each transaction. FP-Growth-like algorithms that accomplish IWI mining efficiently is proposed. The usefulness of the discovered patterns has been validated on data coming from a real-life context with the help of a domain expert. As future work, plan has been made to integrate the proposed approach in an advanced decision-making system that supports domain expert's targeted actions based on the characteristics of the discovered IWIs. Furthermore, the application of different aggregation functions beside minimum and maximum will be studied.

REFERENCES

[1]. R.Agrawal, T.Imielinski, and Swami, "Mining Association Rules between Sets of Items in Large Databases,"Proc. ACM SIGMODInt'l Conf. Management of Data (SIGMOD '93) , pp. 207-216, 1993.

[2]. M.L. Antonie, O.R. Zaiane, and A. Coman, "Application of Data Mining Techniques for Medical Image Classification," Proc. Second Intl. Workshop Multimedia Data Mining in Conjunction with seventhACM SIGKDD (MDM/KDD '01), 2001.

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

[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.