Infrequent Weighted Item Set Mining in Complex Data Analysis

Sujatha Kamepalli
Research Scholar,
CSE Department,
Krishna University
Machilipatnam,
Andhra Pradesh, India.

Raja Sekhara Rao Kurra
Director,
Sri Prakash College of
Eng&Tech,
Thuni ,
Andhra Pradesh, India

Sundara Krishna.Y.K.
Professor, CSE Department,
Krishna University
Machilipatnam,
Andhra Pradesh,
India

ABSTRACT

Infrequent Weighted Association Mining (IWAM) is one of the main areas in data mining for extracting the rare items in high dimensional datasets. Traditional Association rule mining algorithms produce large number of candidate sets along with the database scans. Due to large number of transactions and database size, traditional methods consume more time to find the relevant association rules with the specified threshold. Prior and post database scans are required an additional effort to validate the association rules. Most of the existing weighted models are implemented for mining frequent itemsets, but finding infrequent itemset mining are useful in many recent fields like web,medical,cloud,complex databases,protein sequence etc. In weighted infrequent association rule mining, each item in the transaction is assigned a weight in order to mine high utility infrequent itemsets. In this proposed work, weighted association rule mining algorithm is proposed to find infrequent itemsets using weighted threshold measures. Proposed approach gives better results on real-time datasets compare to existing weighted models.

Keywords

Weighted association rules, Positive rule, Measures, Infrequent itemsets.

1. INTRODUCTION

Mining infrequent association rules is one of the vital issues in the field of data mining due to its wide range applications. Traditional association rules are derived from frequent item sets, which consider occurrence of items but don't reflect other factors, an example would be profit or price. Weighted Association rule mining has recently been proposed, by which transactions are attached with weighted values according to some measure. However, the exact significance of an item set couldn't be easily recognized by static measures. The problem of weighted association rule mining will be to extract the complete variety of association rules which satisfies a support constraint as well as a weight constraint within the dataset. After the computation of the weighted support of one rule, both the support and confidence are consider to discover the weights factors. Weighted Association rule mining has been proposed to find relevant rules by considering the weights of patterns. The idea of Weighted Association rule mining is appealing in which important patterns are discovered. We are able to make use of the term, weighted item set to represent specific weighted items. A simple strategy to find a weighted item set is to calculate the average value of the weights of the items in the item set. Most algorithms make use of a support and confidence constraints to prune the items space. This strategy provides basic pruning however the resulting patterns have weak affinity after mining datasets to take out frequent patterns. Even though minimum support can be increased, it isn't effective in generating patterns with increased weight and/or Support value. In Weighted Interesting Pattern mining user defined rules are identified using weighted frequent patterns. In weighted association rule mining each rule is verified against a new measure, called weight confidence, to consider weight affinity and stop the generation of patterns with substantially different weight stages. If the threshold measure is too high, then less number of item sets will be generated leading to loss of valuable association rules. Nevertheless, whenever the threshold is too low, than large number of frequent itemsets is generated, thereby making it difficult to select the important ones.

Positive frequent items are usually generated in two way process. Firstly, large candidate sets are generated and secondly positive frequent item sets are generated using these large candidate item sets. For example, a market analyst may purchase 10 pen drives and 5 DVDs and another may purchase 5 pen drives and 3 DVDs at a time. The traditional association mining approach treats these two transactions in the same way, which could lead to the loss of some important information. The item sets whose support is greater than the minimum support are referred as positive item sets. The item sets that are expected to be frequent or large are mentioned as candidate item sets. The disadvantage in validating the larger number of association rules that are generated is time consuming process. There's vast literature survey done to minimize the time, candidate sets and total number of association rules. Many of these algorithms stated that the association rules can easily be generated without duplicates or only interesting rules or based on measures.

2. RELATED WORK

Feng Tao et., al [7] addressed problem of discovering relevant binary decision rules in transaction datasets using weighted threshold measures. Traditional model of association rule mining is suffered to handle weighted association rule mining problems in which every item remains to possess a weight. The problem is iterative technique generates large item sets while pruning the item sets. The problem of invalidation “downward closure property” within the weighted measure is solved by introducing a new model of weighted function and exploiting a downward closure property.

In data mining, infrequent and frequent association rules play a key role and have been absolutely applicable in several areas. There are two problems in mining association rules. First one is finding relevant frequent item sets. Another is using those item sets to generate the decision rules. Later on frequent item sets have already been recognized, the corresponding rules could be derived easily. Information discovered from transactional databases in business applications like e-commerce really needs to be maintained, and an incremental updating technique should be developed for maintaining the discovered association rules by reviewing those databases. D.W Cheung[2-3] introduced a new algorithm popularly known as Fast Update algorithm, for efficient association rules when new transactions are incorporated. This system handles the incremental database in order to update the discovered rules. Incremental database should be scanned to achieve the item set filtering. While scanning the incremental database, important candidate item sets are extracted in association with their support that is caused by the incremental database.

Classified association algorithm that selects and analyses the correlation between high confidence rules, instead of relying on just one rule, has been implemented in [4-6]. This algorithm can benefit from a set of related rules to generate predictive rules by evaluating the correlation among them as shown in Fig 1.

However the negative association rules from infrequent item sets are ignored. Furthermore, they set different weighted values for items as stated by the significance of every single item.

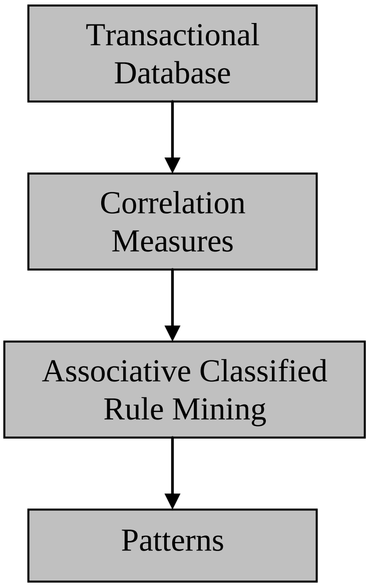

Fig 1. Basic Traditional Weighted Classification Based Association Rules

Traditional weighted frequent pattern mining algorithms are proposed in [4, 5, 6] are based on the Fptree and Apriori algorithm which uses train and tests mechanism. Patterns produced by WFIM have weak support and/or weight affinity patterns. Weighted frequent item set mining algorithm (WFIM) is used to get the most relevant rules out of large patterns. WFIM might use a weight range to regulate the total number of patterns. However, WFIM fails to provide methods to remove patterns that provide items with different support and/or weight levels. WFIM focused on the downward closure property while maintaining algorithm efficiency. It is certainly better if the weak affinity patterns could well be pruned first, causing fewer patterns after mining.

In Jiang et al. [7] support the technique that permits the users to specify multilevel minimum supports to reflect the interestingness of the item sets as well as their changed frequencies in the database. It is extremely effective for large databases to extract association rules in accordance to multiple supports. Existing models are mostly mining negative and positive association rules from frequent item sets.

3. PROPOSED APPROACH

The proposed architecture follows two phrases as stated in [8]: In the first phase equivalence property of the item sets are introduced. Evaluation of the equivalence property is applied on the FPTree of the weighted transactional datasets. In the second phase, each transaction rules in the FPTree is pruned using improved pruning method.

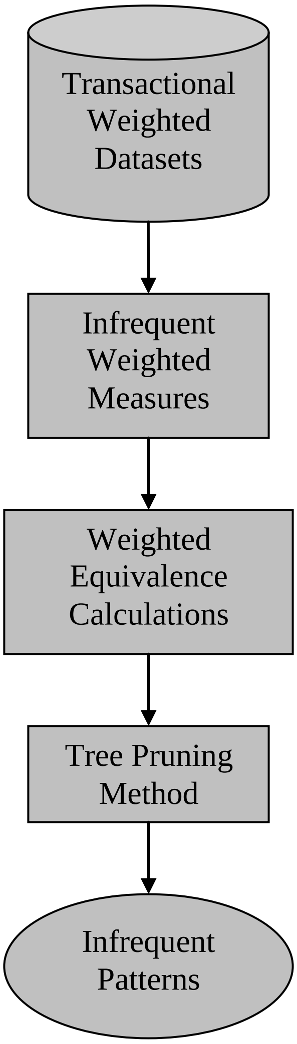
Transactional Weighted Datasets
Infrequent Weighted Measures
Weighted Equivalence Calculations
Tree Pruning Method
Infrequent Patterns

Fig 2. Proposed Flow Chart

Weighted Transaction Dataset: Each item in the transaction dataset is associated with weight. In this approach, a weighted transactional dataset is taken with limited number of transactions and limited items size.

Infrequent Weighted Measures: Let I=$\{i_1, i_2, i_3, ...i_n\}$ be the items, $W = \{w_{i1}, w_{i2}, w_{i3}....w_{in}\} / \mathrm{i} = 1, 2...\mathrm{m}$, be the weights associated with each item in the ith transaction and T=$\{\mathrm{t}_1, \mathrm{t}_2, \mathrm{t}_3 ......\mathrm{t}_n\}$ be the transactions in the dataset D. Infrequent Max and Min support to each item in the Transaction T can be defined as:

Infrequent Min Support= $\sum_{t' \in IS} Min(W_{t'} \text{ IS})$

Infrequent Max Support= $\sum_{t' \in IS} Max(W_{t'} \text{ IS})$

Where IS denote item set and t' is the items in the IS.

Weighted Equivalence Function: Let Transaction t1= {(a, 0), (b, 23), (c, 14), (d, 68)};

Minimum Equivalence Weight Function:

t1.a= {(a, 0), (b, 0), (c, 0), (d, 0)}
t2.b= {(b, 14), (c, 14), (d, 14)}
t3.c= {(b, 9), (d, 9)}
t4.d= {(d, 45)}

Similarly Maximum Equivalent Weight function is the reverse process of minimum equivalent weight function.

Tree Pruning Method:

In this procedure tree pruning is done in two steps as:
(a) FPtree construction and
(b) Mining infrequent item sets recursively from the Modified frequent pattern tree.

Proposed Miner finds infrequent item sets instead of frequent item sets in Fpgrowth. To process this requirement, the following vital changes W.R.T FPgrowth algorithm have been proposed (i) Robust tree pruning method to save the tree search space of items or item set.(ii) Change in tree construction using weighted item sets along with Infrequent support weights and index as shown in Fig 2.1 and 2.2.

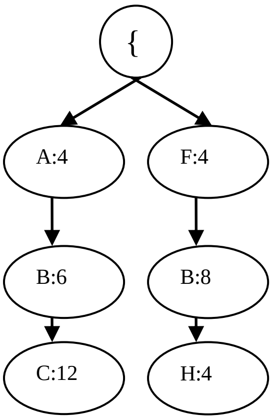

Fig 2.1 Before pruning

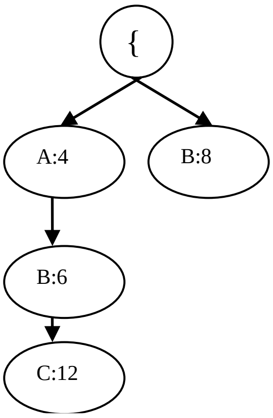

Fig 2.2 After pruning

4. EXPERIMENTAL RESULTS

All experiments are performed using eclipse and Netbeans IDE tool on Intel(R) Core(TM)2 CPU 2.13GHz, 4 GB RAM and the data mining framework. This framework requires third party libraries junit,jama.

```
Support (default 20%)    = 20.0
Confidence (default 80%) = 80.0
Number of records = 768
Number of columns = 38
```

Item sets with Minimal IWI:

```
A B C D E F G H I Minimal Item sets with IWI A  9
A B C D E F G H I Minimal Item sets with IWI A  10
A B C D F G H Minimal Item sets with IWI A  11
A B C D E F G H I Minimal Item sets with IWI A 12
A B C D E F G H I Minimal Item sets with IWI A 13
A B C D F H Minimal Item sets with IWI A 14
A B C D E F G H I Minimal Item sets with IWI A 15
A B C D E F G H I Minimal Item sets with IWI A 16
A B C D E F G H Minimal Item sets with IWI A 17
A B C D E F G H I Minimal Item sets with IWI A 18
A B C D E F Minimal Item sets with IWI A 19
A B C D E F G H I Minimal Item sets with IWI A 20
```

```
Generation time = 0.39 seconds (0.01 mins)
T-tree Storage          = 4832 (Bytes)
Number of frequent sets = 343
```

FPTree index with Item set:

```
[1] {19} = 742
[2] {9} = 700
[2.1] {9 19} = 674
[3] {23} = 634
[3.1] {23 19} = 633
[3.2] {23 9} = 610
[3.2.1] {23 9 19} = 609
[4] {27} = 548
[4.1] {27 19} = 548
[4.2] {27 9} = 542
[4.2.1] {27 9 19} = 542
[4.3] {27 23} = 522
[4.3.1] {27 23 19} = 522
[4.3.2] {27 23 9} = 521
[4.3.2.1] {27 23 9 19} = 521
[5] {1} = 473
[5.1] {1 19} = 472
[5.2] {1 9} = 471
[5.2.1] {1 9 19} = 470
[5.3] {1 23} = 467
[5.3.1] {1 23 19} = 466
[5.3.2] {1 23 9} = 466
[5.3.2.1] {1 23 9 19} = 465
[5.4] {1 27} = 427
[5.4.1] {1 27 19} = 427
[5.4.2] {1 27 9} = 425
[5.4.2.1] {1 27 9 19} = 425
[5.4.3] {1 27 23} = 425
[5.4.3.1] {1 27 23 19} = 425
[5.4.3.2] {1 27 23 9} = 424
[5.4.3.2.1] {1 27 23 9 19} = 424
```

Infrequent Association Rules

```
(1)  [I ] ->  [B ] 100.0%
(2)   [I , E ] ->  [B ] 100.0%
(3)   [D ] ->  [C ] 100.0%
(4)    [G , I ] ->  [G ] 100.0%
(5)   [G , B ] ->  [E ] 100.0%
(6)    [G , C ] ->  [I ] 100.0%
(7)    [H , D ] ->  [H ] 100.0%
(8)     [F , G , H , C ] ->  [F ] 100.0%
(9)    [E ] ->  [H ] 100.0%
(10)   [G , I ] ->  [C ] 100.0%
(11)   [H , C ] ->  [F ] 100.0%
(12)    [D , E ] ->  [E ] 100.0%
(13)   [F ] ->  [H ] 100.0%
(14)    [G , B ] ->  [C ] 100.0%
(15)   [D , C ] ->  [E ] 100.0%
(16)    [G , H , I ] ->  [I ] 100.0%
(17)    [G , D , E ] ->  [E ] 100.0%
(18)   [H , B ] ->   [I ] 100.0%
(19)    [I , E , B ] ->  [G ] 100.0%
(20)     [F , D , E , C ] ->  [B ] 100.0%
(21)     [F , G , H , D ] ->  [F ] 100.0%
(22)    [H , D , E ] ->   [D , I ] 100.0%
(23)   [D , C ] ->  [G ] 100.0%
(24)    [I , C , B ] ->  [F ] 100.0%
(25)   [G , E ] ->  [B ] 100.0%
```

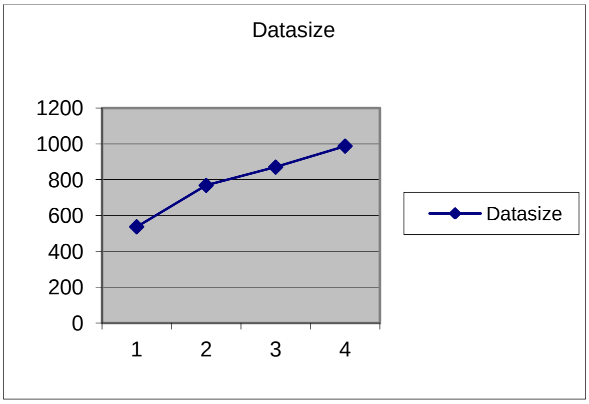

Fig 3. Different Experiments With variable sizes

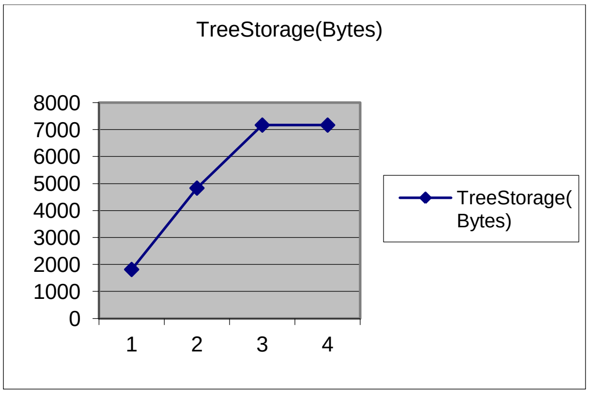

Fig 4. Tree Storage capacity with data sizes.

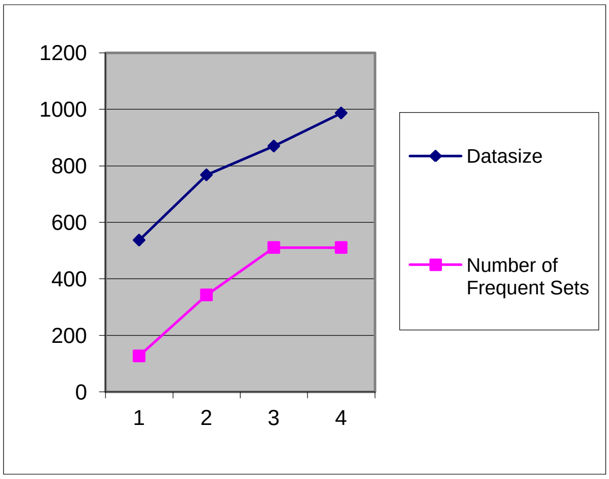

Fig 5. Frequent items vs datasizes

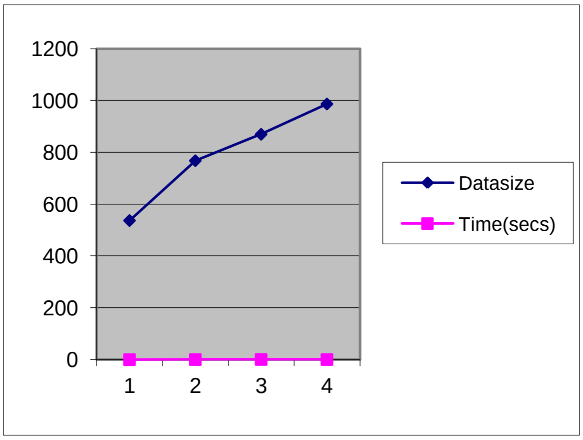

Fig 6. Time vs Datasizes

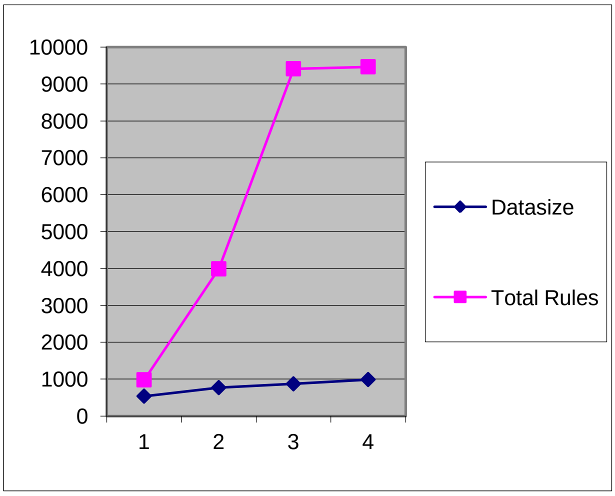

Fig 7. Total infrequent rules vs Data sizes

5. CONCLUSION

In this study, infrequent items from the weighted transactions database are identified with duplicate items. As the size of the database is complex then it is difficult to prune items within the FPTree. Max or Min, Equivalent weighted function produces best results on small datasets with limited items. Equivalent weighted function grows as the size of the item set increases. In future, a new weighted function on complex datasets will be used to eliminate duplicates and filter infrequent item sets.

Limitation in the Proposed Approach:

- If the data size increases, FPTree size also increases, as a result duplication of rules increases.
- Tree pruning mechanism proposed in this paper, completely eliminates the items which are highly relevant for decision making.
- Slight difference in the Max or Min, Equivalent weighted function values which produces large variant frequent item sets.
- Equivalent Weighted transaction is applied to real datasets with shorter item sets and shorter transactions. It takes more time for item sets with longer transactions.

6. REFERENCES

[1] Feng Tao, Fionn Murtagh, Mohsen Farid, “Weighted Association Rule Mining using Weighted Support and Significance Framework”, SIGKDD 2003.

[2] R. Agrawal, T. Imielinski and A. Swami, “Mining Association Rules between Sets of Items in arge Datasets”, International Conference on Management of Data, pp. 207-216.

[3] D. W. Cheung, J. Han, V. Neg and Y. Wong, “Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique”, , The International Conference on Data Engineering, (1996), pp. 106-114.

[4] Alaa Al Deen, Mustafa Nofal and Sulieman Bani-Ahmad, “Classification based On Association-Rule Mining Techniques: A General Survey and Empirical Comparative Evaluation”, Ubiquitous Computing and Communication Journal.

[5] Syed Ibrahim. S.P, Chandran K.R, Abinaya.M.S (2011), “Compact Weighted Associative Classification” in the IEEE International Conference on Recent Trends in Information Technology (ICRTIT 2011), MIT, Anna , pp. 1099 – 1104.

[6] Li, W., Han, J., and Pei, J. (2001),” CMAR: Accurate and Efficient Classification based on Multiple-Class association rule”, In ICDM’01, pp. 369-376.

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

[8] Infrequent Weighted Item set Mining using Frequent Pattern Growth, IEEE Transactions On Knowledge and Data Engineering, Vol. 26, No. 4, 1041-4347 2014.

7. AUTHOR’S PROFILE

K. Sujatha is pursuing her Ph.D. in Krishna University,Machilipatnam, A.P. She is interested doing research in data mining. Present she is carrying her research in Infrequent Pattern Mining. She has Nine international journal publications with high impact factors and indexing. She has two national journal publications. She attended for Two AICTE sponsored 2-week workshops and attended for a number of FDPs. She is member in Indian Association of Engineers (IAE).She has a total of 10 years experience in teaching. She is working as an Associate Professor in CSE Department, Malineni Lakshmaiah Engineering College, Singaraya konda, Prakasam District. A.P.

Prof. K. Rajasekhara Rao is director of Sri Prakash College of Engineering, Thuni, East Godavari District, Andhra Pradesh. He hold several key positions in K.L.University, as Dean (Administration) & Principal, K L College of Engineering (Autonomous). Having more than 26 years of teaching and research experience, Prof. Rao is actively engaged in the research related to Embedded Systems, Software Engineering and Knowledge Management. He had obtained Ph.D in Computer Science & Engineering from Acharya Nagarjuna University (ANU), Guntur, Andhra Pradesh and produced 58 publications in various International/National Journals and Conferences. Prof.KRR was awarded with “Patron Award” for his outstanding contribution, by India’s prestigious professional society Computer Society of India (CSI) for the year 2011 in Ahemadabad. He has been adjudged as best teacher and has been honored with “Best Teacher Award”, seven times. **Dr. Rajasekhar** is a Fellow of IETE, Life Member’s of IE, ISTE, ISCA & CSI (Computer Society of India). Dr.Rajasekhar is nominated as sectional committee member for Engineering Sciences of 100th Annual Convention of Indian Science Congress Association. He has been the past Chairman of the Koneru Chapter of CSI.

Dr. Y. K. Sundara Krishna qualified in Ph.D. in Computer Science and Engineering from Osmania University, Hyderabad. Now, he is working as Professor in the Department of Computer Science, Krishna University, Machilipatnam and presently holding several key positions in Krishna University. His research interests are Mobile Computing, Service Oriented Architecture and Geographical Information Systems and having practical work experience in the areas of Computing Systems including Developing Simulators for Distributed Dynamic Cellular Computing Systems, Applications of Embedded and Win32 clients, Maintenance of Multi-user System Software. He has about 24 international publications and about 2 national publications. Has attended about 5 international conferences and 25 national conferences.