ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

DHUI: A Divisive Apriori methodology for Fast High Utility Itemset mining

Parveentaj M, MCA, ADCA, M.Phil [1], Palanisamy A [2]

Associate Professor, Dept. of computer Applications, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore, Tamil Nadu, India[1]

M.Phil Scholar, Dept. of Computer Science, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore, Tamil Nadu, India [2]

ABSTRACT: High utility itemset mining is a challenging task in Association rule mining, which has wide applications. The state-of-the-art algorithm is High Utility Itemset (HUI) Miner. Although, this existing approach is effective, mining high-utility itemsets remains computationally expensive because HUI-Miner has to perform a costly join operation for each pattern that is generated by its search procedure. The divisive analysis is one of the types of hierarchical method of clustering; the divisive analysis is used to separate each dataset from the transaction dataset. In this proposed study a new methodology DHUI called as Divisive-Apriori methodology is proposed to find the High Utility itemset from the dynamic database. This is a "top down" approach start in one cluster and splits are performed recursively as one move down the hierarchy. Here the datasets are clustered using divisive analysis; the transactional datasets are formed into a single cluster at the end of the scanning. The implementation results show that the proposed DHUI methodology is faster than the existing system.

KEYWORDS: High utility mining, Divisive Apriori, dynamic database, item set, transaction, association rule

I. INTRODUCTION

The Itemset Share approach [9] considers multiple frequencies of an item in each transaction. Share is the percentage of a numerical total that is contributed by the items in an itemset. The authors [9] define the problem of finding share frequent itemsets and compare the share and support measures to illustrate that the share measure approach can provide useful information about the numerical values that are associated with transaction items, which is not possible using only the support measure. This method cannot rely on the downward closure property. The authors developed heuristic methods to find itemsets with share values above the minimum share threshold. Mining high utility itemsets [10] developed top-K objective-directed high utility closed patterns. The authors' definitions are different from our work. They assume the same medical treatment for different patients (different transactions) will have different levels of effectiveness. They cannot maintain the downward closure property but they develop a pruning strategy to prune low-utility itemsets based on a weaker antimonotonic condition.

The theoretical model and definitions of high utility pattern mining were given in [5]. This approach, called mining with expected utility (MEU), cannot maintain the downward closure property of Apriori and the authors of [5] used a heuristic to determine whether an itemset should be considered as a candidate itemset. Also, MEU usually overestimates, especially at the beginning stages, where the number of candidates approaches the number of all the combinations of items. This trait is impractical whenever the number of distinct items is large and the utility threshold is low. Later, the same authors proposed two new algorithms, UMining and UMining H [6], to calculate the high utility patterns. In UMining, a pruning strategy based on utility upper bound property is used. UMining H has been designed with another pruning strategy based on a heuristic method. However, some high utility itemsets may be erroneously pruned by their heuristic method. Moreover, these methods do not satisfy the downward closure property of Apriori, and therefore, overestimate too many patterns. They also suffer from excessive candidate generations and poor test methodology.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

II. LITERATURE SURVEY

Agarwal et al developed an algorithm for mining association rules between sets of items in large databases. Association rule mining is an if/then statement that helps to uncover relationships between seemingly unrelated data in a relational database or other information repository. Apriori Association rule mining technique uses a two step process. The first step is to identify all the frequent itemsets based on the support count value of the itemsets. It uses the download closure property of itemsets to remove the infrequent itemsets. The second step is the generation of association rules from the frequent itemsets using the support and confidence [1].

Han j et al developed an algorithm for mining frequent patterns without candidate generation. In this framework of frequent itemset mining, the importance of items, profit and purchased quantities of items are not considered. Frequent itemset may only contribute a small portion to the overall profit, and non-frequent itemset may contribute a large portion to the profit. Fp-growth improves the efficiency of frequent mining as it does not generate candidate itemsets during the mining process. The drawback of this approach is that it considers only the important items in the frequent pattern [2].

Y.Liu, W-K.Liao, A.Choudhary proposed a two phase algorithm which was developed to find high utility itemsets, using the download closure property of apriori. The algorithms have defined the transaction weighted utilization (twu) while maintaining the download closure property. In this paper they defined two database scans. In the first database scan, the algorithm finds all the oneelement transaction-weighted utilization itemsets and its results form the basis for two element transaction weighted utilization itemsets.In the second database scan, the algorithm finds all the two element transaction-weighted utilization itemsets and it results in three element transaction weighted utilization itemsets. The drawback of this algorithm is that it suffers from level wise candidate generation and test methodology [3].

J Hu et al developed an algorithm for frequent item set mining that identify high utility item combinations. The goal of this algorithm is to find segments of data, defined through combinations of some items (rules), which satisfy certain conditions as a group and maximize a predefined objective function. The high utility pattern mining problem considered is different from former approaches, as it conducts rule discovery with respect to individual attributes as well as with respect to the overall criterion for the mined set, attempting to find groups of such patterns that together contributes to the most to a predefined objective function [4].

Y-C. Li, J-S. Yeh and C-C. Chang proposed an isolated item discarding strategy (IIDS). In this paper, they discovered high utility itemsets and also reduced the number of candidates in every database scan. They retrieved efficient high utility itemsets using the mining algorithm called FUM and DCG+. In this technique they showed a better performance than all the previous high utility pattern mining technique. However, their algorithms still suffer with the problem of level wise generation and test problem of apriori and it require multiple database scans [5].

Liu Jian-ping, Wang Ying, Yang Fan-ding et al proposed an algorithm called tree based incremental association rule mining algorithm (Pre-Fp). It is based on a FUFP (fast update frequent pattern) mining method. The major goal of FUFP is the re-use of previously mined frequent items while moving onto incremental mining. The advantage of FUFP is that it reduces the number of candidate set in the updating procedure. In FUFP, all links are bidirectional whereas in FP-tree, links are only unidirectional. The advantage of bidirectional is that it is easy to add, remove the child node without much reconstruction. The FUFP structure is used as a input to the pre-large tree which gives positive count difference whenever small data is added to original database. It deals with few changes in database in case of inserting new transaction. In this paper the algorithm classifies the items into three categories: frequent, infrequent and pre-large. Pre-large itemsets has two supports threshold value i.e. upper and lower threshold. The drawback of this approach is that it is time consuming [6].

Ahmed CF, Tanbeer SK, Jeong BS et al developed HUC-Prune. In the existing high utility pattern mining it generate a level wise candidate generation and test methodology to maintain the candidate pattern and they need several database scans which is directly dependent on the candidate length. To overcome this, they proposed a novel tree based candidate pruning technique called HUC-tree, (high utility candidate tree) which captures the important utility information of transaction database. HUC-Prune is entirely independent of high utility candidate pattern and it requires three database scans to calculate the result for utility pattern. The drawback of this approach is that it is very difficult to maintain the algorithm for larger database scan regions [7].

Shih-Sheng Chen et al (2011) proposed a method for frequent periodic pattern using multiple minimum supports. This is an efficient approach to find frequent pattern because it is based on multiple minimum threshold support based on real time event. All the items in transaction are arranged according to their minimum item support (MIS), and it does

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

not hold download closure property, instead it uses sorted closure property based on ascending order. Then PFP (periodic frequent pattern) algorithm is applied which is same as that of FPgrowth where conditional pattern base is used to discover frequent patterns. This algorithm is more efficient in terms of memory space, thereby reducing the number of database scans [8].

Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, Young-Koo Lee, Ho-Jin Choi et al proposed a Single-pass incremental and interactive mining for finding weighted frequent patterns. The existing weighted frequent pattern (WFP) mining cannot be applied for incremental and interactive WFP mining and also for stream data mining because they are based on a static database and its require multiple database scans. To overcome this, they proposed two novel tree structures IWFPTWA (Incremental WFP tree based on weight ascending order) and IWFPTFD (Incremental WFP tree based on descending order) and two new algorithms IWFPWA and IWFPFD for incremental and interactive mining using a single database scan. IWFPFD ensures that any non-candidate item cannot appear before candidate items in any branch of IWFPTFD and thus speeds up the prefix tree. The drawback of this approach is that large memory space, time consuming and it is very difficult to support the algorithm for larger databases [9].

Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu proposed an efficient algorithm for mining high utility itemsets from transactional databases.In this paper, they discovered two algorithms named as UP-Growth and UP-Growth+ for mining high utility itemsets from transactional databases. In this technique they are totally dependent on the candidate length; it scans the database twice to construct the UP-Tree. They used efficient utility mining algorithm to generate huge number itemsets called potential high utility itemsets (PHUIs). In this technique they achieved a better performance than all previous high utility pattern mining techniques. However these algorithms still endure with the problem of search space, level wise candidate generation and wide memory usage [10].

III. PROPOSED SYSTEM

3.1 DATA PREPROCESSING

In this phase, the D-Apriori applied in the dynamic transactional database to find the high utility itemsets accesses from the dynamic transactional database. The database is pre processed in data cleaning, user identification, session identification and path completion. The clustered datasets are separated by using divisive analysis the formed transactional datasets are used in Apriori algorithm. Apriori algorithm is used for mining frequent item sets which are used for Boolean association rules generation. The wiener transformation is to convert the binary pre processes data into real data. D-Apriori algorithm is to mine the frequent occurring nodes.

3.2 Apriori Algorithm:

Apriori is designed to operate on databases containing transactions. Apriori uses a "Top down" approach, where frequent subsets are extended one item at a time and groups of candidates are tested against the data. Apriori algorithm for mining associated item sets which are used for Boolean association rules generation. Apriori algorithm is a level-wise, breadth-first algorithm which counts transactions, which is explained in the following Algorithm steps. Apriori uses an iterative approach known as a level-wise search, in which n-item sets are used to explore (n+1)-item sets. First, the set of frequent 1-itemsets is found. This set is denoted P1. P1 is used to find P2, the frequent 2-itemsets, which is used to find P3, and so on, until no more frequent n-item sets can be found. Finding of each Pn requires one full scan of the database.

3.3 Divisible Apriori Gen:

Supervised Apriori Gen(Fk−1)
1. if k = 2 {Deal with candidate 1- and 2-itemsets}
2. for each frequent 1-itemset f ∈ F1 do
3. insert f into C1. {Generate candidate 1-itemsets}
4. end for
5. (C1 class label, C1 other) = split(C1, CL).
6. for each candidate itemset c1 ∈ C1 class label do

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

7. for each candidate itemset c2 ∈ C1 other do
8. c = form c1 and c2. 9. insert c into C2. {Generate candidate 2-itemsets}
10. end for
11. end for
12. for each candidate itemset c1 ∈ C1 class label do
13. for each candidate itemset c2 ∈ C1 class label − {c1} do
14. c = form c1 and c2.
15. insert c into C2.
16. end for
17. end for
18. else
19. for each i1 in Fk−1
20. for each i2 in Fk−1
21. if (first k − 2 items of i1, i2 are same) ∧ (last item of i1, i2 differs)
22. c = form (first k − 1 items of i1) and (last item of i2).
23. insert c into Ck. 24. end if
25. end for
26. end for
27. end if
28. return Ck.

The main purpose of D-Apriori algorithm is scanning the transactional dataset for every process from the database, so the computational and scanning time is very high. So to overcome from this the D-Apriori is proposed to reduce the computation and scanning time.

IV. EXPERIMENT AND ANALYSIS

4.1 Time Comparison:

In this section, we compare the performance of the DHUI methodology with the exiting algorithms on synthetic datasets. The following table shows implemented execution time.

	CHUD (ms)	DHUI (ms)
Reading 1	35	25
Reading 2	27	24
Reading 3	38	29
Reading 4	24	20

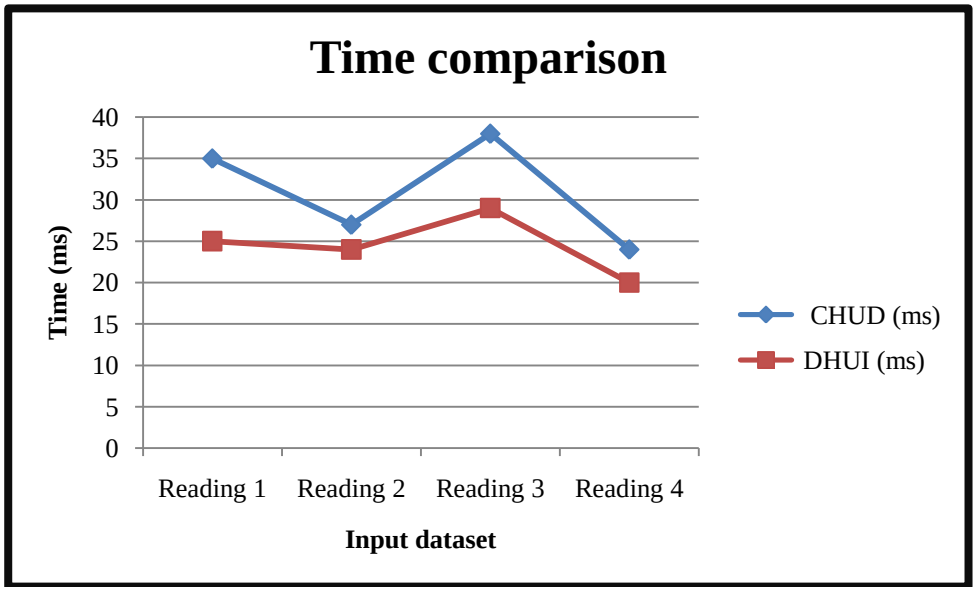
Time comparison table 4.1

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015



Time comparison chart 4.2

V. CONCLUSION

In this paper the DHUI (D-Apriori High Utility Itemset) is proposed to find the high utility dataset from the transactional database. The Divisive Apriori takes less time for computation and more efficient compare to existing system. Hence, the potential high utility itemsets can be efficiently generated from the transactional database with minimum number of scans. In the experiments, both of synthetic and real datasets are used to evaluate the performance of our methodology. The mining performance is enhanced significantly since both the search space and the number of candidates are effectively reduced by the proposed strategies. The experimental results show that DHUI methodology outperforms the state-of-the-art algorithms substantially, especially when the database contains lots of long transactions.

REFERENCES

[1] R. Agrawal, T. Mielinski, A. Swami. (1993), ―Mining association rule between sets of items in large databases‖, in: proceedings of the ACM SIGMOD international Conference on Management of data. pp: 207-216, 1993.
[2] Han J, Pei J, Yin Y, ―Mining frequent patterns without candidate generation‖ In: proc of the ACM-SIGMOD int'l conference on management of data, pp: 1-12, 2002.
[3] Y.Liu, W.K. Liao and A. Choudhary, ―A two phase algorithm for fast discovery of high utility itemset‖, Cheng, D. and Liu. H. PAKDD, LNCS. PP: 689-695, 2005.
[4] J.Hu, A. Mojsilovic, ―High utility pattern mining: A method for discovery of high utility itemssets‖, in: pattern recognition. PP: 3317-3324, 2007.
[5] Y.-C. Li,j,-s. Yeh, and C.-C. Chang, ―Isolated Items Discarding Strategy for Discovering High Utility Itemsets,‖ Data and Knowledge engg., pp: 198-217, 2008.
[6] Liu Jian-Ping, Wang Ying Fan-Ding, ‖Incremental Mining algorithm Pre-FP in Association Rule Based on FP-tree‖, Networking and Distributed Computing, International Conference, pp: 199-203, 2010.
[7] Ahmed CF,Tanbeer SK,Jeong B-S, Lee Y-K (2011) ―HUC-Prune: An Efficient Candidate Pruning Technique to mine high utility patterns‖ Appl Intell PP: 181–198, 2011.
[8] Shih-Sheng Chen, Tony Cheng-Kui Huang, Zhe-Min Lin, ―New and efficient knowledge discovery of partial periodic patterns with multiple minimum supports‖, The Journal of Systems and Software 84, pp. 1638–1651, 2011, ELSEVIER.
[9] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, Young-Koo Lee a,Ho-Jin Choi(2012) ―Single-pass incremental and interactive mining for weighted frequent patterns‖, Expert Systems with Applications 39 pp.7976– 7994, ELSEVIER 2012.
[10] Vincent S Tseng, Bai-En Shie, Cheng-Wu, Philip S,‖ Efficient algorithms for mining high utility itemsets from transactional databases‖, IEEE Transactions on knowledge and data engineering, 2013.