International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 6, Nov-Dec-2018, ISSN (Online): 2395-566X

Data mining Approach for High utility Mining as Outlier Detection: A Survey

Rashmi Rohitas **Prof. Ruchi Dronawat**
rashmi.rohitas0001@gmail.com dron.ruchi@gmail.com
Department of Computer Science & Engineering
Sagar Institute of Research and Technology
Bhopal, M.P., India

Abstract- **Data mining is the process of identifying patterns in data sets by applying appropriate methods with cluster of machine learning techniques. In recent decades, high utility item set (HUI) mining has become the emerging research area, which focuses on frequency and also on utility related with the item sets. Each item set has a value like profit or user's interest, called as the utility of that item set. HUIs are present in customer transaction databases which yield a high profit. The target of HUI is to discover the item sets that have utility value higher than the threshold value. The issues faced in HUIs are dealing with negative item values and number of database scans, mining in XML database, candidate sets and distributed computing network. This paper presents a survey of various algorithms and their restrictions in mining HUIs and the performance analysis of the surveyed algorithms.**

Keywords- **Data Mining, association rule mining, high utility item set (HUI), utility mining, negative item values, distributed computing.**

I. INTRODUCTION

Data mining is an algorithmic method that accepts information or data as input and generates significant patterns, like, classification rule set, association rules, itemsets, or summaries, as output. This led to the discovery of more comprehensive algorithms to refine the ideas of the already existing applications in other fields. The goals of data mining are.

1. Prediction - determining the behavior of certain attributes,
2. Identification – identifying patterns in itemsets,
3. Classification – partitioning of data into various classes,
4. Optimization – improving the use of limited resources like memory space, execution time or materials.

Data Mining is closely associated with Knowledge Discovery in Database (KDD). The purpose of KDD is discovering significant and useful information in extremely larger itemsets in the database. Two elementary issues in KDD are frequent itemsets mining (FIM) and association rule mining (ARM). Conventional Data Mining algorithms were focused on discovering correlation between items that appear commonly in the database. But it absolutely was inappropriate for the practical applications wherever factors such as gains of items and purchase amount got to be examined. High utility itemset mining is an area of research, which is applicable to a wide collection of applications which include stock market prediction, retail market data analysis and recommended systems [1]. This high utility itemset (HUI) mining mines a transaction database that consists of transactions, where unit profit of every item is also taken into account, in addition to purchase quantities [2]. Before going ahead, some basic preliminaries have to be focused.

1. Association Rule Mining (Arm)

Association Rule Mining is the process of finding frequent patterns, associations or correlations in item sets from transactional databases. This mining also helps in finding the association between the item sets, which are present in large databases [3]. One of the best examples of ARM is market analysis, in which, the frequently associated items are discovered in a transaction.

An example of frequently associated items are (bread, jam) and (butter), that is, when a customer purchases bread and jam, then he is 80% likely to buy butter along with bread and jam. ARM can be divided into the following two steps, (1) Generation of frequent item sets, and (2) Generation of association rules. The major challenging task in ARM is identifying the frequent item sets. The task of generating association rules is quite uncomplicated; the focus of most of the researchers is on generating the frequent item sets.

II. FREQUENT ITEMSET MINING (FIM)

FIM is the method of discovering the itemsets that seem often unitedly in the transactions [5]. The focus of FIM is

to witness the entire set of itemsets those appears frequently in the transactional databases. It helps in finding descriptive patterns that exceeds a threshold. This mining ignores the factors such as, utility, profit, cost and quantity related with the item in the itemset [6]. The occurrence of the itemset exclusively may not be sufficient, as it depicts only the number of transactions that holds the itemsets.

Nevertheless, utility of an itemset like, quantity, weight and profit are also significant for treating the real-world decision making problems that involves maximizing the utility in an organisation. To treat this problem, a new algorithm, weighted association rule mining was proposed by Shankar, et al [8].

The conventional method may find a huge sum of frequent itemsets which are low-valued and hence loses the information on priceless itemsets which have low marketing frequencies [4]. Therefore, it cannot be able to satisfy the user's requirement, who desires to locate itemsets with high utilities, in other words, high gain. To cover these kinds of consequences [9] and [10], a utility based itemset mining came into existence.

III. HIGH UTILITY ITEMSET MINING (HUIM)

The fundamental motivation behind Utility Mining algorithms is to distinguish the itemsets with highest utilities, over a specific threshold, by taking into account the profit, quantity, cost or other user preferences [7]. If the support of an itemset surpasses a minimum support threshold specified by a user, that itemset is taken into account as frequent. The utility of an itemset is portrayed as the outer utility increased by the interior utility.

A HUI has its utility esteem is higher than a client indicated least edge esteem; vice versa is considered as a low-utility itemset [11]. Evidently, this process couldn't endure the problems of a search space, when databases hold lots of lengthy transactions or there is a fixed low minimum utility threshold.

HUIM is necessary for many applications like market analysis, streaming analysis, biomedicine and mobile computing. Recently hinted compact tree structure called UP-Tree preserves the data of transactions and item sets, helps the performance of mining and avoids scanning of master database repeatedly. This paper concentrates on deep view of the existing utility mining algorithms and list of methods that has been used to acquire the outcomes efficiently and also an analysis is performed to improvise the state-of-the-art algorithms to refine the results.

IV. LITERATURE SURVEY

This section briefly discusses about the manuscripts that deal with certain problems in utility mining and their respective solutions. Wu et al (2018) proposed the recent method of association rule mining for itemsets that involve low-frequency itemsets along with high frequency itemsets. For the low frequency itemsets in order to gain either the utility or the frequency or both, the method of mixing the utility along with frequency aided for developing different association rule [31]. In case of different business strategies, it is proved that the different association rule should be used for best result. To achieve the candidate itemsets, Single phase FP-Growth algorithm is applied for calculating frequency, utility, support, and confidence measure to achieve the association rule.

Index structure is used for generating the utility values, which is used in FP-Growth for achieving the candidate itemsets depending on support values. The performance of the algorithm can be further increased for computing the utility itemsets in future. Using the advanced algorithms, it is aimed that without using the candidate generation, the high utility itemset to be generated, so that it will minimize the time taken for generating the needed association rule for all type of itemsets. Generally the HUI deals many data sets consisting of different values but, Kannimuthu & Premalatha (2014) proposed GA approach to handle the enormous number of different items and transaction [23].

This proposed algorithm will be optimum on important issues, such as, time complexity and space complexity. GA has been placed in main role in data mining applications because of their ability to handle enormous search efficiently. To resolve the leading problems in utility mining like space complexity and database-dependent minimum utility threshold, GA-based method is used for mining HUIs from the database, which also contains negative item values in itemsets.

Most of the utility mining problems deal with client server computing so; Kannimuthu et al (2012) aimed at introducing a new procedure for mining the HUI in distributed environment, called Knowledge as a Service (KaaS) [20]. KaaS makes use of iFUM algorithm to help in providing data independency and reducing the data integration cost in the distributed computing network.

The investigation of the utility mining problem presented by Kannimuthu & Premalatha (2014) is the infusion of HUIs from the data base that remains as the major task in mining. In case of XML data, it is very challenging to perform the mining because of its high complexity [24]. This paper proposed a distributed approach for mining the HUIs in databases that are represented in XML

format. It is already known that KaaS is used for reducing the integration cost in centralized environment. Similarly, to reduce the data integration cost in distributed environment, KaaS integrated with HUI-MINERXML algorithm has been developed. The algorithm proposed is tested with XML database obtained from IBM synthetic dataset (T10I4D10K),

Mushroom and Kosarakand Accidents by varying the minUtil threshold. A challenge in utility mining is encountering high utility itemsets accompanied by negative values from prominent databases. This issue has been investigated through Kannimuthu et al (2015). Various strategies of mining like data structures, pruning strategies and many utility measures of utility mining algorithm have been talked about in this paper [27]. For mining HUIs with negative values, UP-GNIV algorithm has been suggested. To eliminate the negative utility values, RNU and PNI filtering strategies have been used in the databases. For generating the PHUIs efficiently, UP-Tree algorithm has been used. IBM synthetic dataset is used to evaluate the proposed algorithm.

Following generation data mining technologies are mainly focusing on treating of disseminated data sources that are implemented as web services. Kannimuthu et al (2012) proposed a work that uses the process of parallel mining in distributed database to generate the candidate itemsets that are parallel at different slave sites, for mining the HUIs in distributed databases [20]. This algorithm has comparatively less execution time, when compared with centralized version of utility mining algorithms.

Many of the utility mining algorithms operate fine on infrequent and short utility patterns. But they went wrong in extricating the patterns efficiently in dense and extended patterns. Erwin et al (2007) analyzed the problem of mining with dense and long pattern data sets and developed an algorithm, CTU-Mine algorithm, such that the anti-monotone problem is eliminated while using utility based pattern [13]. The proposed algorithm has been compared with existing Two-phase algorithm with same value of transaction weighted utility.

The result shows that the proposed algorithm works efficiently than the existing ones. Most of the approaches used for mining the association rules will inherently consider that the utilities of the itemset will be always equivalent. However, Yao et al (2004) presumed that utilities of the item set may diverge and analyzed different transactional databases and theoretically found a solution to increase the efficiency in utility mining, by identifying the utility bound property as well as the support bound property(by analyzing the utility relationship among the itemsets) [12]. Based on these properties a mathematical model has been designed for utility based mining. Even though there are innumerable significant perspectives based on HUI itemset, there occurs a problem in producing large number of candidate item set when database consist of lots of long transaction. To handle this issue, Tseng et al (2010) presented a resourceful algorithm named UP-Growth for mining HUI itemset miming from transactional databases. UP-Tree data structure has been proffered for preserving the information of HUIs [18].

Hence, the potential HUIs can be effectively prompted from the UP-Tree by scanning the database only twice. In the demonstrations, synthetic and real datasets have been employed to judge the performance of the algorithm. The mining production has been improved extensively because the proposed strategies effectively reduced the search space and the number of candidates. The UP-Growth algorithm outperforms when there are lots of long transactions present in the database.

The discovery of temporal high utility patterns in data streams has been a major challenging task. However the main limitation of FIM is it does not eliminate the non-periodic itemset, so that the efficiency of mining will be decreased which lead to decline in the profit of the retails. Viger et al (2016) proposed an algorithm for filtering the non-periodic pattern in the database. A built-in disadvantage of the conventional high-utility item set mining procedure is that they are incompatible in discovering the recurring customer behaviour on purchase [29].

For example, some products can be bought by the customer frequently for every week or month. This limitation can be overcome by the use of periodic high-utility item set mining. This type of mining algorithm detects the class of items that are bought by the customers periodically for generating high gain. This algorithm helps in filtering a high amount of non-periodic patterns for discovering the needed periodic high-utility itemsets only.

The detection of High utility pattern in the data stream owing to immense application on numerous domains was a demanding task. Liu & Qu (2012) discloses that Utility-list data structure provides the pruning information for HUI-Miner along with the item sets utility information. A very immense number of candidate itemsets are being processed by conventional algorithms during the mining process. Candidate generation process has been skipped while using HUI-Miner for mining high utility item sets [21]. This obviates the utility computation and costly generation of candidates. This HUI-Miner procedure has improved production concerning memory consumption and running time. FIM has a concern that it presumes all

items has same significance and generates the itemset which produces low profit. To resolve this issue, HUIM [26] algorithm has been proposed, which resolves the issue to some extent, but the utility of an itemset is neither monotonic nor anti-monotonic. So to make the algorithm more efficient, Viger et al (2014) proposed an algorithm called FHM (Fast High-Utility Miner), which consolidates the strategy EUCP (Estimated Utility Co-occurrence Pruning)that is used to scale down the number of joins during the mining of high-utility item sets by utilizing the utility-list data structure. This pruning approach is experimentally proved to diminish the search space by 95% and is faster than HUI-Miner algorithm by six times.

HUIs mining continues to be very time consuming, though it is a foremost mining algorithm for countless applications. Zida et al (2016) brought in EFIM to revamp the efficiency of mining in terms of execution timeand memory consumed. This algorithm depends on two upper-bounds for pruning the search space, namely, local utility and sub-tree utility [29].

These upper-bounds are calculated using an array-based utility algorithm called as Fast Utility Counting. This approach calculates the upper bound in linear time and space. High-utility Database Projection (HDP) and High-utility Transaction Margin (HTM) are the two techniques that are used for database projection and transaction merging in order to reduce the cost of database scans. EFIM algorithm consumes comparatively less memory than other algorithms such as, FHM [ref 2], UP-Growth+, HUP-Miner, HUP. EFIM out performs on both dense and sparse datasets.

V. CONCLUSION

The frequent itemset mining was discovered on the proposition that the itemset that appear more customarily in the transactional databases are considered to be in the top of user's wish list. Still, in reality, the mining of itemsets, by taking into account only the frequency of itemset, is a challenging task, as it has been proved in many applications that the itemsets that bestow to the most are not inevitably the frequent itemsets.

Utility mining attempts to bridge this gap by making use of utility factors such as profit, quantity or cost based on user's view. This paper delivered a quick analysis of the various algorithms involved in mining of high utility itemsets from transactional databases. Most of the algorithms were focused on reducing the number of scans done on candidate itemset generation and finding the high utility itemsets with negative values. In future, we will be developing an algorithm for high utility itemset mining.

REFERENCE

[1] Chongsheng Zhang, George Almpanidis, WanwanWang, ChangchangLiu: An empirical evaluation of high utility itemset mining algorithms, Expert Systems With Applications-101, 91–115 (2018)
[2] Pillai J. and Vyas O.P. : Overview of itemset Utiltiy Mining and its Applications, August International Journal of Computer Applications (0975 - 8887) Volume 5 – No. 11, (2010)
[3] Ruowu Zhong, Huiping Wang, Institute of Computer, Shaoguan University, Shaoguan, Guangdong Province,: Research of Commonly Used Association Rules Mining Algorithm in Data Mining, International Conference on Internet Computing and Information Services 978-1-4577-1561-7, (2011)
[4] Jayant Kayastha, Prof. N. R. Wankhade: A Survey Paper on Frequent Itemset Mining Techniques, Volume 6, Issue 12, ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering (2016)
[5] 5. Shekhar Patel B Madhushree: A Survey on Discovering High Utility Itemset Mining from Transactional Database, Information and Knowledge Management www.iiste.org, ISSN 2224-5758 (Paper) ISSN 2224-896X (Online) Vol.5 (2015)
[6] Ms. YogitaKhot, Mrs. Manasi Kulkarni : Survey on High Utility Itemset Mining from Large Transaction Databases, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 4 (2014)
[7] Sudip Bhattacharya, Deepty Dubey: High utility itemset mining, International Journal of Emerging Technology and advanced Engineering, ISSN 2250-2459, Volume 2, Issue 8 (2012)
[8] Shankar S, Dr. Pursothaman T, Jayanthi S: Novel Algorithm for mining High Utility Itemsets, International Conference on Computing, Communication and Networking (2008)
[9] J. Han, J. Pei, and Y. Yin: Mining frequent patterns without candidate generation, ACM SIGMOD Int. Conf. Manag. Data pp. 1–12 (2000)
[10] V. S. Tseng, C. Wu, B. Shie, and P. S. Yu: UP-Growth: An efficient algorithm for high utility itemset mining, ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 253–262 (2010)
[11] Komal Surawase: Efficient Discovery of High Utility Itemset CPGCON-15, Ingle M.D (2015)
[12] Yao, Hong, Howard J. Hamilton and Cory J. Butz: A Foundational Approach to Mining Itemset Utilities from Databases, SDM (2004).
[13] A. Erwin, R. P. Gopalan and N. R. Achuthan: CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach, 7th IEEE International Conference on Computer and Information Technology (CIT), Aizu-Wakamatsu,

Fukushima, pp. 71-76. doi: 10.1109/CIT.2007.120 (2007)

[14] Jianying Hu, Aleksandra Mojsilovic: High-utility pattern mining: A method for discovery of highutility item sets: Pattern Recognition, Volume 40, Issue 11,Pages 3317-3324,ISSN 0031-3203, (2007).

[15] B. Vo, H. Nguyen and B. Le: Mining High Utility Itemsets from Vertical Distributed Databases, 2009 IEEE-RIVF International Conference on Computing and Communication Technologies, Da Nang, pp. 1-4. (2009).

[16] C. F. Ahmed, S. K. Tanbeer, B. Jeong and Y. Lee: Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases, in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 12, pp. 1708-1721, Dec. (2009).

[17] Fournier-Viger P. FHN: Efficient Mining of High-Utility Itemsets with Negative Unit Profits. In: Luo X., Yu J.X., Li Z. (eds) Advanced Data Mining and Applications. ADMA 2014. Lecture Notes in Computer Science, vol 8933. Springer, Cham(2014)

[18] Vincent S. Tseng, Cheng-Wei Wu, Bai-EnShie, and Philip S. Yu. 2010: UP-Growth: an efficient algorithm for high utility itemset mining: In Proceedings of the 16th ACM SIGKDD International conference on Knowledge discovery and data mining (KDD '10). ACM, New York, NY, USA, 253- 262.(2010).

[19] Subramanian, Kannimuthu, Premalatha Kandhasamy and Sriram Subramanian: A Novel Approach to Extract High Utility Itemsets from Distributed Databases. Computing And Informatics, Vol 31, No 6+, (2012).

[20] S. Kannimuthu, K. Premalatha and S. Shankar: Investigation of high utility itemset mining in service oriented computing: Deployment of knowledge as a service in E-commerce, Fourth International Conference on Advanced Computing (ICoAC), Chennai, pp. 1-8.(2012).

[21] Mengchi Liu and Junfeng Qu: Mining high utility itemsets without candidate generation, In Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12). ACM, New York, NY, USA, 55-64. (2012)

[22] Cheng-Wei Wu, Yu-Feng Lin, Philip S. Yu, and Vincent S. Tseng: Mining high utility episodes in complex event sequences : In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13), Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, and Jingrui He (Eds.). ACM, New York, NY, USA, 536-544.(2013)

[23] S. Kannimuthu, K. Premalatha: Discovery of High Utility Itemsets Using Genetic Algorithm with Ranked Mutation, Applied Artificial Intelligence, 28:4, 337-359 (2014).

[24] Kannimuthu, S. ,Premalatha, K: A Distributed Approach to Extract High Utility Itemsets from XML Data, World Academy of Science, Engineering and Technology, International Science Index 87, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 8(3), 498 - 506 (2014)

[25] Fournier-Viger, Philippeand Wu, Cheng-Weiand Zida, Souleymaneand Tseng, Vincent S: FHM: Faster High-Utility Itemset Mining Using Estimated Utility Co-occurrence Pruning, Foundations of Intelligent Systems",Springer International Publishing (2014)

[26] Fournier Viger, Philippe: FHN: Efficient Mining of High-Utility Itemsets with Negative Unit Profits. 8933. 16-29. 10.1007/978-3-319-14717-8_2 (2014).

[27] Subramanian, Kannimuthu and Kandhasamy, Premalatha: UP-GNIV: an expeditious high utility pattern mining algorithm for itemsets with negative utility values: International Journal of Information Technology and Management,14,26-42 (2015)

[28] Philippe Fournier-Viger and SouleymaneZida.: FOSHU: faster on-shelf high utility itemset mining -- with or without negative unit profit, In Proceedings of the 30th Annual ACM Symposium on Applied Computing (SAC '15). ACM, New York, NY, USA, 857-864 (2015)

[29] Fournier-Viger P., Lin J.CW., Duong QH., Dam TL. (2016) PHM: Mining Periodic High-Utility Itemsets. In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2016. Lecture Notes in Computer Science, vol 9728. Springer, Cham (2016).