IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

An Approach for Hiding the Sensitive High Utility Itemsets without the Information Loss

Ms. M. Kanchana[1], Mrs. C. Senbagavalli Msc, MPhil[2]

M.Phil Scholar, Department of Computer Science, Kovai Kalaimagal College of Arts and Science, Coimbatore, India[1]

Assistant Professor, Dept of Computer Science, Kovai Kalaimagal College of Arts and Science, Coimbatore, India[2]

Abstract: The Association Rule Mining is the traditional mining technique which identifies the frequent itemsets from the databases and this technique generates the rules by considering the each items. The traditional association rule mining fails to obtain the infrequent itemsets with higher profit. Since association rule mining technique treats all the items in the database equally by considering only the presence of items within the transaction. The above problem can be solved using the Utility Mining technique. The Utility Mining technique identifies the product combinations with high profit but low frequency itemsets in the transactional database. Hiding High Utility Itemsets (HUIs) is the main challenges faced in the utility mining. The proposed MHIS algorithm computes the sensitive itemsets by utilizing the user defined threshold value. In order to hide the sensitive itemsets, the frequency value of the itemsets is changed. If the utility values of the items are same, the algorithm selects the accurate items and then the frequency values of the selected items are modified. The proposed algorithm MHIS reduces the computational complexity as well as improves the hiding performance of the itemsets.

Keywords: Hiding High Utility Itemsets (HUIs), MHIS, Utility Mining technique, Association Rule Mining (ARM).

I. INTRODUCTION

Data mining can be defined as an activity that extracts some knowledge contained in large transaction databases. Conventional data mining techniques have focused largely on finding the items that are more frequent in the transaction databases, which is also called frequent itemset mining. These data mining techniques were based on support confidence model. Different types of techniques were followed in the data mining such as Association rule mining, High utility itemsets mining and utility mining etc. Association Rule Mining (ARM) is the concept involved in the frequent itemset mining. ARM is the most widely used traditional technique in the data mining and in knowledge discovery and it has a enormous application in the business, inventory predictions, supply chain management, product marketing etc.., the ARM identifies the frequent itemsets from the transactional database and it generates the rules for the itemsets by considering the each item a single value.

The main objective of the ARM is to find the frequently occurring itemsets in the database. It finds all the itemsets frequency that is beyond the minimum support threshold and then the ARM generates the rules for the co-occurrence of the frequent itemsets. The frequent itemset mining concentrates only on the itemsets which occurs frequently in the transaction database neither concentrating on the profit of the itemsets. For example, considering the itemsets such as 'Printers' and the 'scanners', the frequency or the occurrence of the printer in the transactional database be 8 and the occurrence of the scanners be 5 then the ARM determines only the itemsets which are high in the frequency, ie. Obviously the printers. The profit of the 8 printers may be of 30% and the profit of the 5 scanners be 50% but this profit is not considered in the frequent itemset mining only the higher frequency of the itemsets are considered. The limitation of frequent itemset mining lead researchers towards utility based mining approach, which allows a user to conveniently express his or her perspectives concerning the usefulness of itemsets as utility and then find itemsets with high utility values higher than given threshold.

During mining process we should not identify either frequent or rare itemsets but identify itemsets which are more useful to us. Our aim should be in identifying itemsets which have higher utilities in the database, no matter whether these itemsets are frequent itemsets or not. This leads to a new approach in data mining which is based on the concept of utility called as utility mining. High utility itemset mining refers to the discovery of high utility itemsets. The main objective of high utility itemset mining is to identify the itemsets that have utility values above given utility threshold.

II. DECREASING UTILITY VALUE

A tradeoff between the privacy protection and knowledge discovery in the sharing process is an important issue (Jieh-Shan Yeh, Po-Chiang Hsu 2010). The study focuses on privacy preserving utility mining (PPUM) and presents two novel algorithms HHUIF Hiding High Utility Itemsets First (HHUIF) and Maximum Sensitive Itemsets Conflict First (MSICF) to achieve the goal of hiding sensitive itemsets so that the adversaries cannot mine them from the modified database. The main goal of HHUIF is to decrease the utility value of each sensitive itemset by modifying the quantity values of items contained in the sensitive itemset.
The goal of MSICF is to reduce the number of modified items from the original database; it selects an appropriate

DOI 10.17148/IJARCCE.2015.41231

item which has the maximum conflict count among items in the sensitive itemsets. These algorithms reduce the impact on the source database of privacy preserving utility mining. Two algorithms for privacy preserving utility mining are Hiding High Utility Item First Algorithm (HHUIF) and Maximum Sensitive Itemsets Conflict First Algorithm (MSICF).

III. SANITIZATION PROCESS

In general, the sanitization process for PPUM consists of the following three steps such as apply utility mining algorithm on collected database to obtain all high utility itemsets, to identify sensitive itemsets based on business requirements and to apply sanitizing algorithm to generate the sanitized database. The main goal of the HHUIF algorithm is to decrease the utility value of each sensitive itemset by modifying the quantity values of items contained in the sensitive itemset. To decreases the utility value of each sensitive itemset S, HHUIF modifies the item quantity value with the highest utility value in some transaction containing S.

The process repeats until the utility values of all sensitive itemsets are below the minimum utility threshold. This algorithm first calculates the difference between the utility of an itemset and the minimum utility threshold value which is said to be the amount of utility value that is to be reduced. HHUIF sanitizes the transactions containing the sensitive itemsets repeatedly until the difference value is less than or equal to 0. Then the item which is having the maximal utility value for a given sensitive item is selected. The quantity of the item in the transaction is modified if the utility value of the item is less than the difference value. To reduce the number of modified items from the original database, MSICF selects an appropriate item which has the maximum conflict count among items in the sensitive itemsets. Then, it modifies the transaction. HSICF sanitizes the transactions containing the sensitive itemsets repeatedly until difference value is less than or equal to 0. Compared with HHUIF, MSICF selects an appropriate item which has the maximum conflict count among items in the sensitive itemsets in order to reduce the number of items modified from the original database.

IV.GA-BASED APPROACH TO HIDE HUI

Genetic algorithm (Chun-Wei Lin et al, 2014) is used in many research issues since they could provide feasible solutions in a limited amount of time. A GA based privacy preserving utility mining method is proposed to find appropriate transactions to be inserted into the database for hiding sensitive high utility itemsets. Privacy preserving factors such as quantities and profits are considered. The operations in the GA such as selection, crossover and mutation are performed and there occurs the changes in the original database. But privacy preserving utility mining using GA does not change the original database. The side effects like hiding failures, missing costs and artificial costs are considered during the hiding process. The proposed algorithm provides less information loss and protects high risk information in the database.

A GA-based privacy preserving utility mining approach is thus designed to find the appropriate transactions from original database to be inserted in the database. It firstly finds a recommendable size of transactions as the number of genes in a chromosome, which can be determined by the maximal inserted utility value and the total utility value in the original database. The designed algorithm is to extract the transactions from the original database as the optimal solution for a chromosome to be inserted into the database. A lower utility threshold is also obtained to find the pre large transaction-weighted utilization itemsets, thus avoiding the rescanning time of original database. For the proposed GA-based algorithm, a maximal utility value for insertion will be obtained to derive the lower utility threshold. Since the lower utility threshold is obtained according to the maximal insertion utility, it can be considered as an overestimated threshold from all of chromosomes. For each chromosome, a precise threshold value will be determined according to the insertion utility, which is called sliding count for filtering the unpromising pre large transaction-weighted utilization itemsets to reduce the evaluation cost.

V.DISCOVERING HIGH-UTILITY ITEMSETS

An algorithm is proposed (Jianying ,Mojsilovic ,2007) for the frequent item set mining that identifies the high utility itemset combinations. The goal is to find the segments of data, which is defined through the combinations of few items. The author tackles the problem of mining frequent combinations of items with the highest contribution to a particular business objective. The task is considered as a optimization problem and it is solved using the specialized partition trees called High-Yield Partition Trees. Binary partition tree is used to solve a problem of identifying significant defining patterns, where each tree node represents a particular pattern and includes all of its subscribing transactions. We can get a clearer understanding of the requirements of high-utility mining by formulating it as an optimization problem. Let us assume that S is a set of all possible patterns that can be derived from the items I. We are interested in finding, for a given size v, a subset of S containing v patterns, PS ⊂ S, which provides high yield and satisfies the following condition such as All patterns in PS are defining patterns, All patterns in PS are SDPs, All patterns in PS are non-overlapping. Because of the non-overlapping property, identifying SDPs of interest is equivalent to identifying clusters of transactions where transactions in each cluster all satisfy a particular pattern. We call these trees High-Yield Partition Trees, or HYP trees in short.

VI.TREE PRUNING

When the recursive splitting procedure is completed, all insignificant leaves are removed. The remaining leaves are all SDP nodes and together form a pattern set that has the highest possible yield for the chosen item selection criterion. However, such set may very well contain more patterns than the desired size, thus a pruning procedure is required to “trim” the tree to the desired size. The

DOI 10.17148/IJARCCE.2015.41231

procedure needs to be designed so that the number of SDP nodes is reduced one at a time, and maximum yield is preserved at each newly reduced size. For this purpose, we identify a class of nodes called fringe nodes. The pruning procedure is carried out iteratively. During each iteration, all fringe nodes and the associated pruning operations are identified. The operation with minimum incurred in pattern contribution loss is selected and carried out. Any newly "exposed" non-SDP leaves are then recursively removed so that at the end of each iteration all leaves are again SDP nodes.

The iteration stops when the desired number of patterns is reached. Despite the wide used of data mining techniques in client segmentation and market analysis applications, so far there have been no algorithms that allow for the discovery of high-utility frequent item sets, i.e. the ones that contribute the most to a predefined utility, objective function or performance metric. We have presented a novel algorithm for frequent item set mining to identifying such combinations of items. The algorithm is fairly general and can easily account for different item attributes and objective functions. The algorithm has been tested on real world data with promising results. It is currently being used as a decision support tool by the IBM Market Intelligence to investigate impact and cohesion of different product lines, and architect new market strategies.

VII.HIDING THE SENSITIVE HIGH UTILITY MINING

The proposed algorithm is applied to the sensitive high utility itemsets in order to obtain the sanitized database. The sanitized database is the modified database which hides the sensitive high utility itemset. In proposed scenario user introduce the concept of modified hiding high utility itemset algorithm for achieving the privacy as well as low hiding failure. To accomplish the hiding process, this method finds the sensitive itemsets and modifies the frequency of the high valued utility items. However, the performance of this method lacks if the utility value of the items are the same. The items with the same utility value decrease the hiding performance of the sensitive itemsets and also it has introduced computational complexity due to the frequency modification in each item.

To solve this problem, A modified HHUIF algorithm with Item Selector (MHIS) algorithm is proposed. The proposed MHIS algorithm is a modified version of existing HHUIF algorithm. The MHIS algorithm computes the sensitive itemsets by utilizing the user defined utility threshold value. In order to hide the sensitive itemsets, the frequency value of the items is changed. If the utility values of the items are the same, the MHIS algorithm selects the accurate items and then the frequency values of the selected items are modified. The proposed MHIS reduces the computation complexity as well as improves the hiding performance of the itemsets. The algorithm is implemented and the resultant itemsets are compared against the itemsets that are obtained from the conventional privacy preserving utility mining algorithms.

VIII.CONCLUSION

Hiding the sensitive high utility itemsets from the unauthorized users plays a major role in the utility mining. The challenges faced in the utility mining like preserving the privacy by hiding the highly sensitive high utility itemsets are analyzed with different algorithms. The above problem of loss in the original information is addressed in the proposed MHIS algorithm. The proposed technique first presents a privacy preserving utility mining (PPUM) model and builds up an MHIS algorithm to reduce the impact on the source database of privacy preserving utility mining. This algorithm modifies the database transactions containing sensitive itemsets to minimize the utility value below the given threshold while preventing reconstruction of the original database from the sanitized one. The experimental results proved that the performance of proposed MHIS algorithm was better than the conventional HHUIF algorithm.

REFERENCES

1. Al-Ahmadi M.S,(2008)," Privacy-preserving data mining for horizontally-distributed datasets using EGADP", Journal of Communications of the IBIMA, vol.5, no.2, pp.7-15, 2008.
2. Cheng Wei Wu, Bai-En Shie, Philip S. Yu, Vincent S. Tseng (2012)," Mining Top-K High Utility Itemsets", KDD'12, Beijing, china .
3. Chun-Wei-Lin, Guo-Cheng Lan, Tzung-pei Hong (2012), "An incremental mining algorithm for high utility itemsets", in Expert system with applications ELSEVIER.
4. Chung-jung Chu, Vincent S.Tseng, Tyne Liang (2008)," An efficient algorithm for mining temporal high utility itemsets from data streams", in science direct, the journal of system and software, ELSEVIER.
5. Chun-Wei Lin, Tzung-Pei Hong, Jia-Wei Wong, Guo-cheng Lan, wen-yang Lin (2014)," A GA-Based Approach to Hide Sensitive High Utility Itemsets", in Hindawi publishing corporation, the scientific world journal.
6. Chun-Wei Lin, Tzung-Pei Hong, Hung-Chuan Hsu (2014)," Research article- Reducing side effects of hiding sensitive itemsets in privacy preserving data mining", in Hindawi publishing corporation The scientific world journal.
7. V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati (2008), " K-Anonymous Data Mining: A survey in Springer US, Advances in Database Systems .
8. Dongwon Lee, Sung-Hyuk Park, Songchun Moon (2012)," Utility-based association rule mining: A marketing solution for cross-selling", in expert systems with applications ELSEVIER.
9. A. Evmievski, J. Gehrke and R. Srikant (2003)," Limiting privacy breaches in privacy preserving data mining, Proc. of the 22nd ACM SIGMOD-SIGACT-SIGART
10. Guo-cheng Lan, Tzung-pei Hong, Jen-peng Huang, Vincent S.Tseng (2013)," On-shelf utility mining with negative item values", in expert systems with applications, ELSEVIER.
11. Guo-Cheng Lan, Tzung-Pei Hong, Vincent S. Tseng (2010)," Discovery of high utility itemsets from on-shelf time periods of products",in expert systems with application, ELSEVIER .
12. Hong Yao, Howard J.Hamilton (2005)," Mining itemset utilities from transaction database", in ELSEVIER.
13. Jianying Hu, Aleksandra Mojsilovic (2007)," High-utility pattern mining: A method for discovery of high-utility item sets", in pattern recognition ELSEVIER.

DOI 10.17148/IJARCCE.2015.41231