International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 01 | Jan-2021 ISSN: 2582-3930

Utility Mining Over Data Streams Using the Concept of Preserving Privacy

Aswathy V Shaji[1], Jijiya R P[2]

[1]Department of Computer science, Sree Narayana College, Cherthala, kerala, India

*[2]Department of Computer science, Trikaripur College of Arts and Science, Kasargod, Kerala, India

---

Abstract -Utility pattern mining extracts useful knowledge from databases with utility issues. The concept of utility is to consider the quantity and profit of each item. This approach can solve the problem that does not deal with market data in general frequent pattern mining. That is frequent pattern mining simply analyzes the ratio of each itemset to the whole database whereas utility pattern mining considers not only the frequency for each itemset but also their different importance. Therefore results of utility pattern mining can include more important but private, sensitive information for this reason it is more essential to preserve such information from malicious users by applying privacy-preserving utility mining (PPUM) approaches that specializes in utility pattern mining rather than traditional pattern mining. For preserving privacy concerns this system uses a perturbation algorithm that uses a Tree structure and Tables. The perturbation process is one of the PPDM techniques for effectively hiding sensitive utility items or patterns by removing certain items from original databases. This paper Implements a perturbation algorithm for privacy-preserving utility mining in data streams. It is a two-step approach. First, create a new tree structure called PUTT-tree and two tables and then Propose a perturbation algorithm for creating perturbed database.

***Key Words*::**Privacy Preserving Data Mining(PPDM), Privacy Preserving Utility Mining, High Utility itemset Mining based onUT-Tree , Utility Tail Tree

1.INTRODUCTION

The traditional association rule mining, one of the most important methodologies indata mining, discovers all itemsets, which support values, are greater than a given threshold. There are lots of algorithms proposed for discovering the frequent itemsets in literature. The Apriori algorithm is considered as the most famous one. In order to measure how useful an itemset isin the database, utility mining is proposed . It overcomes the shortcomings of traditional association rule mining, which ignores the sale quantity and price (or profitability) among items in a transaction. Utility pattern mining one of the interesting pattern mining techniques, which can analyze business relationships in market data including utility issues since such utilitypattern mining approaches can deal with non-binary data such as quantities of products and consider relative importance of items such as theirprofits.

Recently, integrating utility constraints into data mining tasks has drawn much attention among the researchers.Several researchers have proposed many algorithms and techniques for mining high utility item sets.Moreover,researchers from the data mining area have highly utilized the qualitative aspects of attributes such as significance, utility than considering only the quantitativeones(e.g.numberofappearancesinadatabase,etc.,)for thereasonthatqualitative properties are required in order to completely use the attributes present in the dataset. Mining high utility item sets improves the standard frequent item set mining framework because it utilizes the intuitively defined utility rather than statistics-based support measure. Utility mining is widely used in many practical applications. Naturally, utility is a measure of how useful (i.e.gainful)anitemsetis.Thelocaltransactionutilityandexternalutilityareemployedto evaluate the utility of an item or item set. The local transaction utility of an item is defined based on the information stored in a transaction,such as the number of the item,,sold in the transaction, whereas the external utility of an item is based on the information from resources besides transactions, like a profit table. In some business environments, the data mining may need to be processed among databases.

Although the data may be distributed among several sites, the sites are not allowed to reveal its database to another site. For instance, some insurance companies have their own databases which contain their insured ones information. For mutual benefit, these companies decided to work with insurance fraud detection by distributed data mining. The data mining model must be high accurate to identify fraud, because a fault leads to huge loss of income or great amounts of pay. Also, insurance companies cannot share the data about their clients withother companies, because of the restriction laws (and having a high competitive edge). They can share information about the fraudulent insurance records, but not their data. Each company have made an effort to share their black-box models to find out more interesting rules on the entire shared information than that on their own database, and can defend the private records that other companies may find. Privacy considerations may prevent thisapproach.

Privacy Preserving DataMining

Privacy preservation is becoming more and more a serious problem for future progress of data mining techniques with great potential access to datasets having private, sensitive, or confidentialinformation.Themajorchallengeforexistingdatamining algorithmsisextracting accurate data mining results while still maintaining privacy of datasets. Due to the increasing concern on privacy, a new category of data mining called privacy preserving data mining (PPDM) has been introduced. However, the privacy-preserving data mining has turned into a major problem in recent years because of the huge amount of private data which is tracked by several business applications. In many situations, the users are reluctant to provide personal information unless the privacy of sensitive information is assured. PPDM was first introduced by Agrawal and Srikant in 2000. PPDM algorithms are developed by integrating privacy protectionmechanismtoconcealsensitivedatabeforeexecutingdataminingalgorithms.Then several different branches with different goals have been developed. Privacy preservingclassification techniques prohibit a miner from building a classifier which is capable of forecasting the personaldata.

The main consideration in privacy preserving data mining is the sensitive nature of raw data. The data miner, while mining for comprehensive statistical information about the data, should not be able to access data in its original form with all the sensitive information. This necessitates more robust techniques in privacy preserving data mining that intentionally alter the data to conceal sensitive information as well as protect the inherent statistics of the data which is vital for mining purpose.

Privacy preserving Utility Mining

Many questions against privacy concerns in the business and market environments have been increased after utility pattern mining is proposed. These privacy concerns say that malevolentpeopleabuseprivateinformationdiscoveredbyutility miningfortheirownprofits.PPUM approacheshavebeendevelopedtoefficientlypreventtheseprivacy concernsinmarketdatabases by integrating PPDM and utility pattern mining methods. In the PPDM field, there are vari- ous techniques that can be combined with utility pattern mining, and they are classified as input and output privacy. Input privacy preserving methods change the contents of databases before conducting data mining operations for the databases. This approach includes the following techniques: perturbation, data swapping, randomization, locking-based technique, and k-anonymity. The other PPDM research field is output privacy preserving methods suchas secure multiparty computation (SMC). These approaches are based on a major premise that malevolent people do not realize how to process privacy preserving tasks. Such limitation is also similar to the case of cryptographicalgorithms. preference (2) Data distribution changes constantly with time (3) The amount of data is enormous (4) Data flows in and out with fast speed (5) Immediate response is required. Thesecharacteristics create a great challenge to data mining. Traditional data mining algorithms are designedforstaticdatabases.Ifthedatachanges,itwouldbenecessarytorescanthedatabase, whichleadstolongcomputationtimeandinabilitytopromptlyrespondtotheuser.Therefore,

traditionalalgorithmsarenotsuitablefordatastreamsanddatastrea msminghasrecently become a very important and popular researchissue.

2.DESIGN AND ARCHITECTURE

Privacy preserving against mining algorithms is a new research area that examines the side effects of data mining techniques that obtained from the privacy diffusion of persons and organizations[6] . The objective of PPDM algorithms is to mine the significant knowledge from huge amounts of data while preserving sensitive personal information at the same time. Recent research made in this area has given much effort to establish a trade-off between the right to privacy and the need of knowledge discovery. It is often the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria. The utility of the data, at the end of the privacy preserving process, is an important problem, because in order for sensitive information to be hidden, the database is essentially modified by the insertion of forged information or by the blocking of data values. Also, the measure employed to calculate the information loss depends on a particular data mining technique with respect to which a privacy algorithm is performed. Issues regarding privacy-preserving data mininghave emerged globally. The recent development in PPDM techniques is evident. Assume that, we have a server and multiple clients, where each client has a set of data. The clients require the server to collect statistical information about the relationship among items in order to provide recommendationstothecustomers.But,theclientsdonotwantthe servertoknowaboutsome sensitive patterns. Sensitive pattern is the frequent itemset that have a sensitive knowledge. So, when a client sends its database to the server, some sensitive patterns are concealed from its database based on some particular privacy policies. Hence, the server only can collect the statistical information from the modified database.

Byconsideringthesefactsthispaperpresentedanefficient approachforminingofprivacy preservinghighutilityitemsetsfromtheperturbeddatastream.Th isapproachisexecutedwith the aid of three majorsteps:

- Find sensitive itemsets from the originaldatastreams.
- Transformation of Original datastream into perturbeddatastream.
- Construction of sensitive utility PUTT-Tree using perturbeddatastream.
- Mining of sensitive utility item sets from the sensitive utilityPUTT-tree

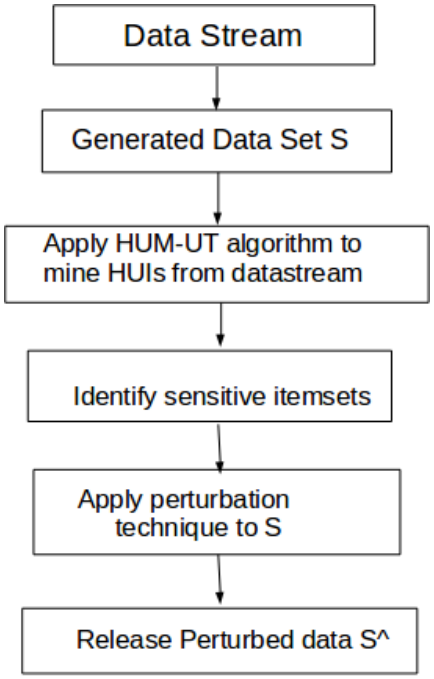

Figure. 1: Perturbation Process

Mining High Utility itemsets from datastream

ThissystemusesaminingalgorithmcalledHUM-UT(HighUtilityitemsetsMiningbased on UT-Tree) to find high utility itemsets from transactional data streams. Here adatastructure called UT-Tree(Utility on Tail Tree)is used for maintaining utility information of transaction item sets to avoid mulitipledb scans. The UT-Tree is created with one db scan and contains a fixed number of transaction itemsets;utility information is stored on tail nodes only .Based on this data structure and the sliding window approch The HUM UT algorithm mines high utility itemsets from the UT-Tree without additional database scan.

The tree main steps in HUM UT algorithm are:

- Construction of aUT-Tree
- Removing obsolete data and updating newdata
- Mining high utility itemsets from the currentwindow

Main steps of HUIsare:

- calculate the minimum utilityvalue
- Create a header table for the globaltree.
- Add an attached list to each leafnode
- Calculate twu and utility value of each item in the headertable
- Create a prefix-tree and a header table for the base-itemset.
- Process theprefix

The output of this algorithm is high utility items. Along with this HUIs the original datastream is given to the PUTT system to produce perturbed datastream.

Data Stream Perturbation process

Data perturbation is a popular technique for privacy-preserving data mining. The ma jor challenge of data perturbation is balancing privacy protection and data quality, which are normally considered as a pair of contradictive factors. In this approach, the distribution of each data dimension reconstructed independently.ie perturbation process change the contents of databases before conducting data mining operations for thedatabases.

It has 3 main steps

- Constructing the PUTT-tree
- Constructing SI-table andII-Table
- Perturbing the Generated DataStream UsingPUTT-Tree

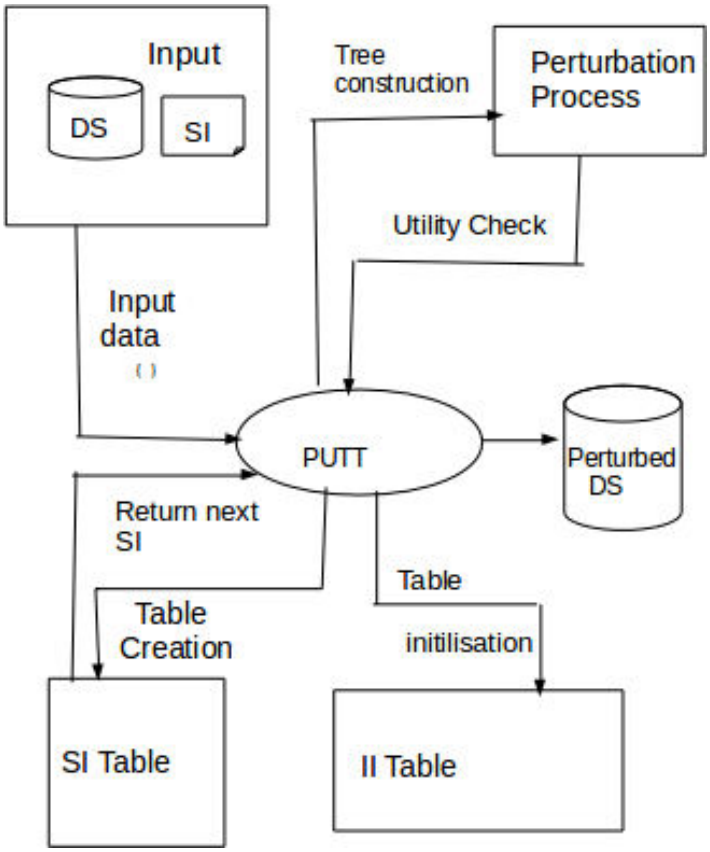

Figure 2: overall PUTT Process

3.CONCLUSION

This approach present an idea for fast perturbation algorithm based on a tree structure ,whichmorequicklyperform a data streamperturbationprocessesforpreventingsensitiveinformation from being exposed. This approach is based on three main steps: (i) Transformation of Original data stream into perturbed data stream.(ii)Construction of sensitive utilityFPUTT-Tree using sanitized database and (iii)Mining of sensitive utility item sets from the sensitive utility FPUTT-tree. The effectiveness of the approach is based on the algorithm used in the mining process. Here we use HUM UT algorithm. The proposed system helps to prevent data from malevolentpeople.

REFERENCES

[1] UnilYun,JiwonKim , "A fast perturbation algorithm using tree structure for PPUM",Expert Systems with Applications 42 (2015) 11491165, Science Direct septem- ber2014.

[2] L.Feng,LeWangandBoJin,"UT-Tree:Efficientminingofhighutilityitemsetsfromdata streams", Intelligent Data Analysis17(2013)585-602.

[3] Rakesh Agarwal,RamakrishnanSrikant "Privacy-Preserving Data Mining" ,In Proceed- ings of the 2000 ACM SIGMOD international conference on management of data (pp. 439450). Dallas, TX,USA.

[4] Yeh, J., Hsu, P, " HHUIF and MSICF: Novel algorithms for privacy preserving utility mining.",Expert Systems with Applications, 37(7),47794786.

[5] MortezaZihayat, Aijun An, "Mining top-k high utility patterns over data streams",Information Sciences 285 (2014) 138161, Science Direct february 2014..

[6] Gangin Lee, UnilYun a, Keun Ho Ryu "Sliding window based weighted maximal frequent pattern mining over data streams",Expert Systems with Applications 41 (2014) 694708, Science Direct september2013.

[7] Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun, "Privacy-Preserving Classification of Data Streams",Tamkang Journal of Science and Engineering, Vol. 12, No. 3, pp. 321- 330(2009)

[8] ArisGkoulalas-DivanisVassiliosS. Verykios, "Hiding sensitive knowledge without side effects",KnowlInfSyst (2009) 20:263299,Springer-Verlag London Limited2008.

[9] GrigoriosLoukides , ArisGkoulalas-Divanis , "Utility-preserving transaction data anonymization with low information loss ",IBM Research Zurich, Zurich,Switzerland-
December 2011

[10] Vincent S. Tseng, Bai-EnShie, Cheng-Wei Wu, and Philip S. Yu, Fellow, IEEE, "Effi- cient Algorithms for Mining High Utility Itemsets from Transactional Databases ",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25,NO. 8,AUGUST2013.

[11] Mohammad NaderiDehkordi, KambizBadie, Ahmad KhademZadeh, "A NovelMethod for Privacy Preserving in Association Rule Mining",Journal on Software, vol. 4, no. 6, pp: 555- 562,2009.