

Adding Dummy Items To Hide Sensitive Association Rules

Ruchi.P.Kanekar¹, Prof. Rachel Dhanaraj²

¹(Computer Engineering Department, Goa College Of Engineering, India)

²(Computer Engineering Department, Goa College Of Engineering, India)

Abstract: Association rule and data mining techniques is most commonly used nowadays by every organizations and individual to extract the knowledge or the useful information from large amount of data. Knowledge and the information gives a gain to the organization .The main objective of the association rule hiding algorithms is to hide sensitive information so that they cannot be discovered through association rule mining algorithm, but at the same time not losing the great benefit of association rule mining and try to hide more and more rules and retrieve the original database without losing the integrity of the database when association rule mining algorithm is inverted. However the proposed system will hide the sensitive rule by using the dummy items and generate different database which is known as modified database so that no one can get the original database.

Keywords: Association rule, Data Mining, Frequent Itemset, Knowledge Discovery in Database(KDD), Sensitive Rule.

I. Introduction

Data mining is the technique which provides functionality for discovering any pattern for any kind of database, which includes classification of data and prediction of data and also help to find out frequent pattern on any database. In data mining it includes various databases such as relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files.

Data mining technique are used to discover hidden information from large databases. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Various data mining techniques such as, decision trees, association rules, and neural networks are already proposed and become the point of attention for several years.

Association Rules have proven to be beneficial in inducing knowledge from the dataset and helping in the crucial decision making in all the fields. However, it also possesses threat to the privacy of data. By making use of Association rule mining or other data mining techniques business firms or individuals can easily gain knowledge about each other which can be used to increase the business, resulting into gain for one and loss for other.

To understand the concept there is the simple example that one company want to collaborate with other company for business relationship and in future certain problem occur between the company and they break the business relation and don't want to work together .So this time one company has the database of other company it could easily retrieve the association rule and the transaction of customer purchase. To avoid the risks in future one company should hide the sensitive rule and modify the database and handover the database to other company now this time it becomes unavailable for other company to find out the sensitive rule and the items which is more frequently purchased.

II. Motivation

Data mining is consist of several data, which performs an operation on different kinds of data it doesn't require a particular form of data to perform an operation ,machine learning and data mining both performs classification, clustering and association techniques on data. Nowadays, machine learning and data mining is mostly used in every fields of business .Association rule can be used on centralized and decentralized data. Centralized means the data is been present at one location and Decentralized means the data is been present at several different locations. For business gain association rule mining is done so that the organization can get success , in order to avoid the leakage of sensitive rule here a hiding approach is used .

III. Literature Survey

In [2] which includes a Association rule mining and also discuss about various Association rule hiding approaches, here approaches are first is Heuristic based approaches which is further divided into Data distortion technique ,this technique changes a selected set of 1-values to 0-values(delete items) or 0-values to 1- values (add items) and Data blocking technique, this technique adds uncertainty in the database by replacing 0"s and 1"s by unknowns ("?) in selected transaction instead of inserting or deleting items .Second is Cryptography based approaches in which data is present on several sites and each sites encrypt the original database before sending to the Admin . Third is Border based approaches, in this sensitive association rules are hidden by modifying the borders in the lattice of the frequent and the infrequent item set of the original database. Fourth is an exact approach, in this approach formulates the hiding process as a constraints satisfaction problem (CSP) or an optimization problem which is solved by binary integer programming (BIP) and last is Reconstruction based approaches or Data reconstruction approaches place the original data aside and start from sanitizing the so-called "knowledge base".

In [3] has discuss about Privacy Preserving Data Mining Methods and also classified the privacy preserving methods which include four methods such as data distribution, purposes of hiding, data mining algorithms, and privacy preserving techniques. Data distribution includes Centralized-DB and Distributed-DB, Hiding Purpose includes Data Hiding and Rule hiding which comes under Centralized-DB and Data Hiding comes under Decentralized –DB. Data mining algorithms include Classification, Clustering and Association Rules. Privacy Preserving Techniques include Generalization, Data Distortion, Data Sanitation, Blocking and Cryptography Techniques. Some technique of Privacy Preserving comes under Centralized –DB and some in Distributed-DB. Generalization transforms and replaces each record value with a corresponding generalized value.

In [4]a privacy is preserved on distributed database by using a cryptographic technique, here a method for mixed partitioned in which data is first partitioned vertically and then horizontally and also shows another mixed method in which data is partitioned horizontally and then vertically partitioned. In distributed environment, database is a collection of multiple, logically interrelated databases distributed over a computer network and are distributed among number of sites. As the database is distributed, different users can access it without interfering with one another.

In the horizontally partitioned distributed database model, there is n number of sites and every site owner has local autonomy over their database and one special site called Trusted Party (TP) who has special privileges to perform certain tasks. Sign based secure sum cryptography method to find global association rules by preserving the privacy.

In [5] the author has proposed the PPDM by using the Bayesian Network which includes steps described as follow:

1. Read XML Document, here the document is the transactions in XML format.
 2. Form transactional itemset and binary table from the inputted document. Transactional Symbolized Items are a group of symbolized items that forms a transaction based on XML document items. Binary Table of Transaction is a table containing 1's and 0's to represent the presence or absence of an item in a transaction.
 3. Apply Apriori algorithm on the transactional itemset from D to generate association rules. Apriori algorithm is for effectively generating XML association rules after preprocessing in step 2.
 4. From the binary table of transactional itemset, use K2 to generate a Bayesian Network. BN and K2 Algorithm produce a useful graphical model that trains and displays interesting relationship among nodes in a probabilistic manner.
 5. Item# column is read to identify Mode using Conditional Probability Table (CPT). This table contains items and their conditional probabilities according to their dependency in Bayesian Network.
 6. Modify the transactional itemset based on Mode that is obtained in step 5. From CPT, the most frequent item(s) are identified for the modification of transactions. This kind of frequent item identification is called Mode.
 7. Apply Apriori algorithm again on the modified transactional itemsets. Then output the results in XARs.
- In [9], a method is introduced to extract task-oriented information from biological texts.

In [10] to hide the association rule , a hybrid algorithm is proposed which is based on two previous existing algorithm ISL and DSR ,K-Mean Neural gas Cluster Algorithm with Number of cluster in this algorithm, first we decrease support of right hand side of the rule in a rule where item to be hide is in right side. Here a real time database is used such as Doctor Patient Evaluation . It shows the comparison between the Execution Time shows a comparison between K-mean + ISL + DSR ,Neural gas + ISL + DSR, ISL ,DSR. Neural gas is a simple algorithm for finding optimal data representation based on feature vectors.

IV. Related Theory

Data mining is often viewed as a step in the larger process of Knowledge Discovery in Databases, or KDD for short, which consists of a series of steps: cleaning, integration, selection, transformation, data mining, evaluation and finally presentation [1].

Data mining is used to build predictive and descriptive models. A predictive model is used to explicitly predict values. As an example, based on the customers who have responded to an offer, the model can predict what other customers are most likely to respond to the same offer. Descriptive models on the other hand describe patterns in existing data. It can provide valuable information, such as identifying different customer segments [1].

4.1 Transaction data

A transaction consists of a transaction id and a set of the items that are part of the transaction. Usually there will also be additional information, such as the date of the sale, and information about the customer and salesperson. But the dataset used for association rule will include the transaction and multiple items in each transaction.

4.2 Star Schema

The star schema (also called star-join schema, data cube, or multi-dimensional schema) is the simplest style of data warehouse schema. The star schema consists of one or more fact tables referencing any number of dimension tables. The star schema is an important special case of the snowflake schema, and is more effective for handling simpler queries.

Star Schema has certain advantages such as every dimension will have a primary key, a dimension table will not have any parent table and there is no relation between any two dimension tables.

4.3 Association rules Mining

Association rules are if/then statements that help to uncover relationships between unrelated data in a database, relational database or other information repository.

For example, if the customer buys bread then he/she may be likely to buy butter. If the customer buys laptop then he/she may also buy pen drive.

There are two basic criteria that association rules uses first is support and second is confidence. It identifies the relationships and rules generated by analyzing data for frequently used if/then patterns. Association rules are usually needed to satisfy a user-specified minimum support and a user -specified minimum confidence at the same time.

Association rule generation is usually split up into two separate steps:

1. First, minimum support is applied to find all frequent item sets in a database.
2. Second, these frequent item sets and the minimum confidence constraint are used to form rules.

Patterns mined from transaction data often come in the form of association rules, which have the following definition.

Let $I = \{i_1, i_2, \dots, i_m\}$ be the set of all items. Let D be the set of all transactions, where every transaction T is a set of items $T \subseteq I$. An association rule is of the form $X \rightarrow Y$ where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$.

The support for a set of items T is the probability of that itemset appearing in a transaction: $P(T)$ The support for an association rule $X \Rightarrow Y$ is the probability $P(X \cup Y \subseteq I)$ To compute the support, the number of transactions containing all the items in both X and Y is divided by the total number of transactions.

The confidence of an association rule is defined as $P(Y | X)$. It is computed by dividing the number of transactions containing Y , by the number of transactions containing X .

There are various algorithm to find out the association rule but the basic algorithms to find the frequent itemset is as follows:

i. Apriori Algorithm

Apriori[7] is the most classical and important algorithm for mining frequent itemsets and this algorithm makes multiple passes over the database. It has the most important property that "All non empty itemsets of a frequent itemset must be frequent".

Disadvantages:

- It needs to generate a huge number of candidate sets.
- It needs to repeatedly scan the database and check a large set of candidates by pattern matching.
- It is costly to go over each transaction in the database to determine the support of the candidate itemsets.

Advantages:

- Easy to implement.
- Use large item set property.

ii. Eclat Algorithm

Eclat[7] uses a vertical database layout i.e. instead of explicitly listing all transactions; each item is stored together with its cover (also called tidlist) and uses the intersection based approach to compute the support of an itemset.

Disadvantage:

- It takes more time when the transaction id list is large.

Advantages:

- No need to scan database each time.

iii. FP Growth Algorithm

FP Growth[7] is another important frequent pattern mining method, which generates frequent itemset without candidate generation. It constructs conditional frequent pattern tree and conditional pattern base from database which satisfy the minimum support.

It suffers from certain disadvantages:

- FP tree may not fit in main memory
- Execution time is large due to complex compact data structure

Advantages:

- Database is scanned only two times.

a. Sensitive Association Rule Hiding

As the word Sensitive data means” a data is important and handle with care”, Data mining is the process of identifying patterns from large amount of data. Association rule hiding is one of the techniques of privacy preserving data mining to protect the sensitive association rules generated by association rule mining. Data mining that finds sensitive rules on any forms of databases. Association which is one of the technique in data mining which is used to find out the association rule and sensitive rule which should be hide by using a hiding technique. A dummy item is used to hide the sensitive rules.

In order to hide the knowledge from a database, the rules that help in discovering this knowledge have to be hid. These rules that help in the discovering the knowledge and help in taking decisions are called as Sensitive Association Rule.

V. Proposed Methodology

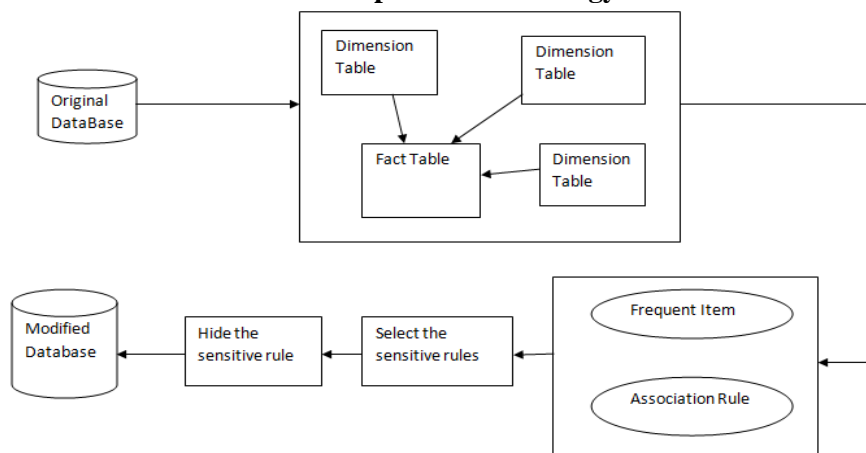


Figure 1. Overall flow diagram

The above diagram shows the flow of the proposed system. These steps are specified below:

1. **Original Database:** The database which include the transactions and a transactions which include different items that customer purchased.

2. Star Scheme representation : Star Schema which includes fact table and dimension table ,in star schema there is one or two fact table and multiple dimensional tables .In fact table it has only the dimensional table id and in dimensional table it include the details of particular table.

3. Frequent Itemset and association rule : In this support is used to find out the frequent itemset and confidence is used to find the sensitive association rule

4. Support and confidence : This is the most important aspect ,this help to find out the sensitive rules. Support and Confidence is specified by the user.

5. Modified database: Here this database which include all the item as well the dummy item so that no one can get the original database.

VI. Algorithm

1. Create a database which includes multiple tables such as dummy tables, product table etc.
2. Represent the star schema which includes the fact table and dimensional table.
3. Take the user specified minimum support and minimum confidence.
4. Calculate the frequent itemset using the support.
5. Calculate the association rule using the confidence.
6. Select the sensitive rules, based on the confidence greater than equal to minimum confidence.
7. Use the dummy items to hide the rules.
8. Insert the dummy items in the rules.
9. And then modify the database.
10. Display the database which is modified.

VII. Conclusion

There are various algorithms which is been studied to find out the frequent itemset and generate the association rules. Here the dummy item is added to hide the sensitive rules. It uses the transaction of customer purchase to find out which items are purchased most often. By using this approach organizations and individuals can easily share their database with each others without the fear of sensitive information getting revealed and also the database remains secure.

References

- [1]. Anton Flank , **Multirelational Association Rule Mining**, 7th September 2004.
- [2]. Suma B, **Association Rule Hiding Methodologies: A Survey**, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2 Issue 6, June– 2013.
- [3]. K. Srinivasa Ra , B. Srinivasa Rao , **An Insight in to Privacy Preserving Data Mining Methods** , The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 3, July-August 2013 .
- [4]. N V Muthu lakshmi, Dr. K Sandhya Rani, **Privacy Preserving Association rule mining for Horizontally Partitioned Distributed Database**, International Journal of Computer Science and Information Technologies, Vol. 3 (1) , 2012, 3176 – 3182.
- [5]. Khalid Iqbal, Sohail Asghar, Simon Fong, **A PPDM Model Using Bayesian Network for Hiding Sensitive XML Association Rules**, IEEE-2011.
- [6]. R.D.Chintamani, I.F.Shaikh A.D.Waghmare **Hiding Sensitive Association Rules on Stars** Emerging Trends in Computer Science and Information Technology -2012.
- [7]. Siddhrajsinh Solanki, Neha Soni, **A Survey on Frequent Pattern Mining Methods Apriori, Eclat, FP growth** *International Journal of Computer Techniques*-ISSN :2394-2231.
- [8]. Aakansha Saxena, Sohail, **Gadhiya A Survey on Frequent Pattern Mining Methods Apriori, Eclat, FP growth** 2014 IJEDR, Volume 2, Issue 1, ISSN: 2321-9939.
- [9]. K Dhanasekaran, R Rajeswari, **A method for extracting task-oriented information from biological text sources**, International Journal of data mining and bioinformatics, Vol.12 (4), pp.387-399, 2015
- [10]. Mohnish Patel, Prashant Richhariya, Anurag Shrivastava, **A Novel Approach for Data Mining Clustering Technique using NeuralGas Algorithm**, 2014 Fourth International Conference on Advanced Computing & Communication Technologies.