Effective Measurement Approach To Hide Informative Association Rule Sets

[1] V.Kaleeswari and [2] N.Abirami
[1 , 2] Assistant Professor, Computer Science & Engineering, Sree Sastha Institute of Engineering and Technology, Chennai.

Abstract

Data mining is defined as the process of extracting or mining knowledge from large amount of data and potentially useful information from data. The information and knowledge gained can be used for application ranging from multiple approaches and viewed as a result of the natural evolution of information technology. The knowledge discovered by data mining may contain sensitive information which may cause potential threats towards privacy and security. of on-sensitive knowledge is not lost. preserving data mining modifies the original data in some way, so that the private data and private knowledge remain private even after the mining process. The privacy preserving association rule mining aims at finding all the rules that are greater than the user .

2. INTRODUCTION

Data mining has been developed as an important technology tool for identifying patterns and trends from large quantities of data. Many applications of data mining have been demonstrated in marketing, business, medical analysis and engineering design etc. In these days of information security and privacy protection, such as the type of disease and details of purchase must be properly protected.

1.1 PRIVACY PRESERVING DATA MINING

The privacy preserving data mining has been response to the concerns of preserving privacy information from data mining. The main goal of privacy preserving data mining has become more important. The increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. Knowledge hiding on the other hand, is concerned with the sanitization of confidential knowledge from the data. The specific class of methods in the knowledge hiding area is known as frequent item set and association rule hiding.

The concerns of preserving privacy information are of two types in data mining. The first type of privacy, called output privacy ,is that the data is minimally altered so that the mining result will preserve certain privacy. Many techniques have been used in this type of privacy. The other type of privacy, namely input privacy, manipulates the data so that the mining result is not affected or minimally affected. These approaches require hidden rules or patterns and accept the rules required for data mining process.

The framework for privacy preservation consists of an original database, mining algorithm (Chieh-Ming Wu 2009), sanitized algorithm and hiding association rules etc. There are two stages in the framework. In the first stage, such as mining algorithm, index tree, frequent itemsets are generated before the sanitized algorithm is executed.

The second stage uses sanitized algorithm (Ali Amiri 2007) for transactional databases and also add some information with the least impact of database. In order to search for sensitive transactions in the transactional database, access the transaction IDs and items.

1.2 PRIVACY PRESERVING TECHNIQUIES

There are many approaches (Verykios 2004) which have been adopted for privacy preservation. They can be classified, based on the following dimensions:

1. Data distribution
2. Data modification
3. Data mining algorithm
4. Data or rule hiding
5. Privacy preservation

1.2.1 Data distribution

Data distribution refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to those cases where different database records reside in different places, while vertical data distribution refers to the cases where all the values for different attributes reside in different places.

1.2.2 Data modification

Data modification refers to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released to the public. In this way, high privacy protection is ensured. It is important that a data modification technique should be in consonance with the privacy policy, adopted by an organization.

Some methods of modification include:

1. Perturbation**,** which is done by the alteration of an attribute value by a new value (i.e. changing a 1-value to a 0-value, or adding noise).

2. Blocking, which is the replacement of an existing attribute value with a "?".

3. Aggregation or merging which is the combination of several values into a coarser category.

4. Swapping that refers to interchanging values of individual records.

5. Sampling which refers to releasing data for only a sample of a population.

1.2.3 Data mining algorithm

Data mining algorithm is actually something that is not known beforehand, but it facilitates the analysis and design of the data hiding algorithm. In the recent years, various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks.

1.2.4 Data or rule hiding

Rule hiding depends on whether raw data or aggregated data should be hidden. The complexity of hiding aggregated data in the form of rules is, of course, higher, and for this reason, mostly heuristics have been developed. The lessening of the amount of public information causes the data miner to produce weaker inference rules that will not allow the inference of confidential values. This process is also known as "rule confusion".

1.2.5 Privacy preservation

Privacy preservation is the most important technique which refers to the selective modification of the data. The main task is to preserve the data while transferring them from one organization to another organization.

1.3 ASSOCIATION RULE MINING

An association rule is an expression of the form X=>Y, (Verykios et al 2004) where X and Y are item sets and $X \cap Y = \{ \}$. Such a rule expresses the association that, if a transaction contains all items in X, then that transaction also contains all items in Y. The support of an association rule X =>Y in D is the support of $X \cup Y$ in D, and similarly, the frequency of the rule is the frequency of $X \cup Y$. An association rule is called frequent if its support (frequency) exceeds a given minimal support (frequency) threshold.

The confidence or accuracy of an association rule X =>Y in D is the conditional probability of having Y, contained in a transaction, given that X is contained in that transaction : Confidence (X=>Y,D) := P(Y / X) = support ($X \cup Y$, D) / support (X, D). This rule is called confident if P(Y / X) exceeds a given Minimum Confidence Threshold (MCT). Such a rule is called strong association rule or a generating itemset. For example, the association rule "bread & jam => butter" exists with confidence 45% means the probability of the transaction containing bread and jam to include external memory card is 45%. If the MCT is 30% then the rule is valid.

1.4 INFORMATIVE ASSOCIATION RULE SETS

Informative association rules are basically single target rules that there are no subset rule with higher confidence. For any non-informative single-target rule, there is an informative association rule with higher confidence. To hide rule sets, the pre-process of finding hidden rules can be integrated into the hiding process as long as the predicting items are given and automatically hiding informative association rule sets without pre mining and selection of hidden rules.

An example of such a rule is that 90% of customers who buy burgers also buy coke. The 90% here is called the confidence of the rule, which means that 90% of transaction that contains X (burgers) also contains Y (coke). The confidence is calculated as $|X \cup Y| / |X|$, where |X| is the number of transactions containing X and $|X \cup Y|$ is the number of transactions containing both X and Y. The support of the rule is the percentage of transactions that contain both X and Y. This is calculated as $|X \cup Y| / N$, where N is the number of transactions in D. In other words, the confidence of a rule measures the degree of the correlation between itemsets, while the support of a rule measures the significance of the correlation between itemsets. The problem with mining association rules is that they find all rules that are greater than the user-specified minimum support and minimum confidence.

2. RELATED WORK

2.1 INFORMATIVE ASSOCIATION RULE HIDING APPROACHES

Two basic approaches have been used in association rule hiding and concerns of information from data mining algorithms(Shyue – Liang Wang 2007).

.

The first approach hides one rule at a time and selects transactions that contain the items in a given rule. It tries to modify transaction by transaction until the confidence or support of the rule in minimum confidence or minimum support. The modification is done either by removing items from the transaction or by inserting new items to the transactions.

The second approach deals with groups of restricted patterns or association rules at a time and selects the transactions that contain the intersecting patterns of a group of restricted patterns. Depending on the disclosure threshold given by users, it sanitizes a percentage of the selected transactions in order to hide the restricted patterns. All these approaches require hidden rule or patterns in advance. The discovered rules and privacy requirements, hidden rules or patterns are then selected manually. But, for some applications. only interested in hiding certain constrained classes of association rules such as information association rule sets.

2.2 INCREASE SUPPORT OF LHS AND DECREASE SUPPORT OF RHS

In order to hide association rule X => Y, decrease its supports to be smaller than pre-specified minimum support or its confidence to be smaller than pre-specified minimum confidence. The first increases the support count of X, on the left hand side of the rule, but it does not support the count of $X \cup Y$. By decreasing the support count of the item, it becomes possible to remove one item at a time in a selected transaction by changing from 1 to 0.

For the second case, in the transactions containing both X and Y, if decrease the support of Y only, on the right hand side of the rule, it would reduce the confidence faster than reducing the support of X. By increasing the support count of an item, it will add one item by changing from 0 to 1.

For example, in a given database in Table 2.1, a minimum support of 33% and a minimum confidence of 70%, the rules is C => A (66%,100%), B => C (50%,75%), C => B (50%,75%),B => A (66%,100%), where the percentages inside the parentheses relate to supports and confidence respectively and the notations are specified Table 2.2

Table 2.1 Original Database

TID	Items
T1	*ABC*
T2	ABC
T3	AB
T4	A
T5	AC

Table 2.2 Bitmap representation

TID	Items	Size
T1	111	3
T2	111	3
T3	110	2
T4	100	1
T5	101	2

The results of hiding item C and then item B, using ISL and DSR algorithm the rule C => B (50%,75%) will be hidden, if transaction T5 is modified from 100 to 101 using ISL.

DSR is C => A (60%,100%), C => B (50%,75%),B => C (50%,75%),B => A (60%,100%) will be hidden, if transaction T6 is modified from 101 to 001,T1 is modified from 111 to 011,T1 is modified from 011 to 001, and T4 is modified from 110 to 010 using DSR. The new database D3 is shown in Table 2.3.

Table 2.3 Databases before and after hiding item C and item B using ISL and DSR

TID	**D**	**D1**	**D2**	**D3**	**D4**
T1	111	111	111	001	010
T2	111	111	111	111	111
T3	110	110	110	010	010
T4	100	101	110	100	100
T5	101	101	101	001	001

3. PROPOSED METHODOLGY

Privacy preserving data mining preserves the confidentiality of the sensitive knowledge in the form of sensitive association rule mining. The problem of data sanitization is NP-hard. It even leads to complex problems in real time due to low efficiency.

To improve efficiency, there are several methodologies such as hash-based technique, transaction reduction, partitioning, sampling, and dynamic item sets counting. In this project the efficiency can be improved by partitioning technique.

3.1 PARTITIONING METHOD

The basic idea of data partitioning approach is quite simple. The data set is initially partitioned into fragments and the whole process is divided into two main phases. In the first phase, the data mining methods are applied to process individual fragments to get a local result. In the second phase, a combinated step generates the global result. The main technical challenge lies in achieving the same computational results (as would be executed on the entire data set) but to gain some efficiency features from the adopted data partitioning approach. Hence, the main research question is how to partition the data set so that data partitioning will be beneficial.

The benefit of this method is substantial, since it allows reasonably low execution time, even when performing knowledge hiding in very large databases. The proposed modification approach addresses the efficiency and identifies the item sets in transactions and reduces the number of modified entries. To accomplish this, partitioning method is suggested to improve the efficiency.

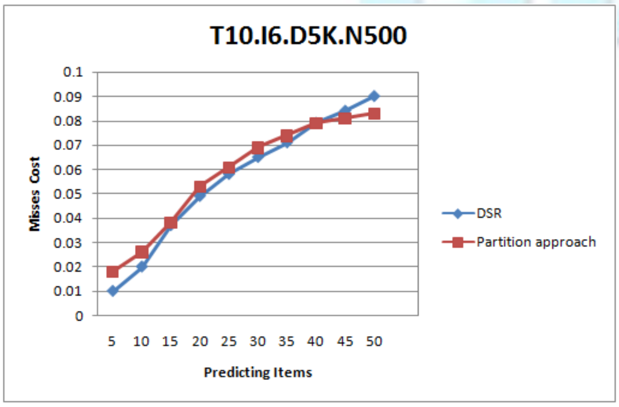

Fig 6.2 Misses Cost for T10.I6.D5K.N500

3.2 EFFECTIVENESS MEASUREMENT

Misses Cost (MC) is used to measure effectiveness of rule hiding method. It is measured in terms of the percentage of legitimate association rules that are not discovered from D’

$$\text{MC} = \frac{\#\sim SR(D') - \#\sim SR(D')}{\#\sim SR(D')}$$

Let D be original database and D’ be the modified database. #~ SR (D) denotes the number of non-sensitive association rules discovered from the database D.

REFERENCES

1. Ali Amiri (2007) ‘Dare to share: Protecting sensitive knowledge with data sanitization’, IEEE transactions on Knowlwdge and data engineering.

2. Aris Gkoulalas-Divanis, Vassilios S. Verykios (2009) ‘Exact knowledge hiding through database extension’, IEEE transactions on Knowledge and Data engineering, Vol.21, No.5.

3. Chieh-Ming Wu, Yin-Fu Huang and Jian-Ying Chen (2009) ‘Privacy Preserving Association Rules by Using Greedy Approach’, World Congress on Computer Science and Information Engineering.

4. Dasseni, E., Verykios, V.,Elmagarmid, A., &Bertino, E. (2001) ‘Hiding association rules by using confidence and support’. In Proceedings of 4th information hiding workshop,Pittsburgh,PA, pp. 369-383.

5. Li,J., Shen, H., & Topor, R. (2001) ‘Mining the smallest association rule set for predictions’.In Proceedings of the 2001 IEEE international conference on data mining,pp. 361-368.

6. Oliveira, S. Zaiane, O. (2002) ‘Privacy preserving frequent itemset Mining’, Proceedings of the IEEE ICDM Workshop on Privacy,Security and Data Mining, pp. 43–54.

7. Rizvi,S. J., & Haritsa, J. R. (2002), ‘Privacy-preserving association rule mining’.In Proceedings of the 28th International conference on very large databases,August.

8. Shyue-Liang Wang, Bhavesh Parikh, Ayat Jafari (2007) ‘Hiding informative association rule sets’ ,Expert Systems With Applications, Vol. 33, pp.316-323.

9. Verykios, V., Elmagarmid, A., Bertino, E., Saygin, Y., Dasseni, E.(2004) ‘Association rule hiding’, IEEE Transactions on Knowledge and Data Engineering,Vol. 16, pp. 434–447.

10. Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen (2007) ‘Hiding Sensitive Association Rules with Limited Side Effects,’ IEEE transactions on Knowledge and Data engineering, Vol. 19, No.2, pp. 29 – 42.