Proceedings of the 2013 International Conference on
Pattern Recognition, Informatics and Mobile Engineering, February 21-22

Detection of Sensitive Items in Market Basket Database using Association Rule Mining for Privacy Preserving

S. Kasthuri
Department of Computer Science and Engineering
Alagappa University
Karaikudi, India
kasthu.s@gmail.com

T. Meyyappan
Department of Computer Science and Engineering
Alagappa University
Karaikudi, India
meyslotus@yahoo.com

Abstract— **Data mining is an essential technology to extract patterns or knowledge from large repositories of data. Association rules in market basket database represent the shopping behavior of customers. The association information may reveal trade secrets. It must be hidden before publishing. Association rule hiding in privacy preserving data mining hides sensitive rules containing sensitive items. In this paper, a new method is proposed to detect the sensitive items for hiding sensitive association rules. This proposed method finds the frequent item sets and generates the association rules. It employs the concept of representative association rules to detect sensitive items.**

Keywords*— *Data Mining; Association rule hiding; Market Basket Database; Sensitive Items; Privacy Preserving Data Mining.

I. INTRODUCTION

Data mining is a knowledge discovery process of analysing large databases to find useful patterns. It has numerous applications such as marketing, business, medical analysis, product control, engineering design, bioinformatics and scientific exploration, etc. The knowledge discovered from a database can be expressed in patterns such as decision trees, clusters or association rules. Retail shops are interested in associations between different items that customer place in the shopping basket. By discovering interesting association rules, an organization can identify underlying patterns useful in planning of new store layout, new product assortments and which product to put on promotion. The knowledge discovered can be beneficial to business competitors. In order to preserve their competitive edge some partners may hesitate to disclose sensitive information and also to prevent data mining techniques from discovering sensitive knowledge which is not even known to the database owners.

Privacy preserving data mining is a broad research area for protecting sensitive data or knowledge [2]. There have been two types of privacy concerning data mining. The first type of privacy, called output privacy, is that preventing the mining output from malicious inference attacks [4]. The second type of privacy, input privacy, is that sanitizing the raw data itself before performing mining [5]. The framework of privacy preserving data mining is shown in Fig 1.

In market basket databases, the sensitive association rule must be hidden for competency. Many algorithms have been proposed for hiding sensitive association rules. These algorithms have been given some set of sensitive items manually. The market basket database contains large number of items. Selection of sensitive items manually from market basket database takes more time.

In the proposed method, representative association rules are computed from association rules. Based on the discovered rules sensitive items are selected for privacy preserving.

The paper is organized as follows. Section 2 gives the view of the market basket analysis using association rule mining strategy. Section 3 describes association rule mining algorithms. Section 4 presents the problem statement. Section 5 analyses the theoretical background and related work. Section 6 presents the proposed algorithm for sensitive items. Section 7 shows the implementation of the proposed algorithm. Concluding remarks and future work are described in Section 8.

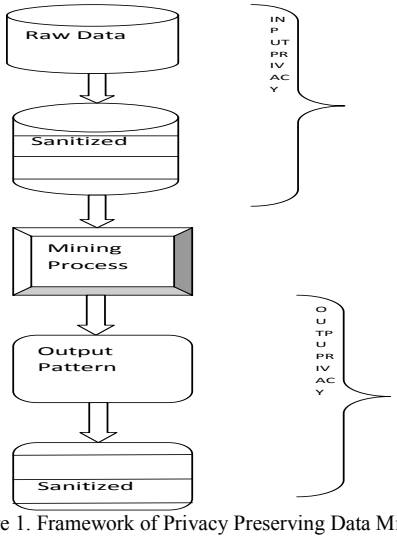

Figure 1. Framework of Privacy Preserving Data Mining

978-1-4673-5845-3/13/$31.00©2013 IEEE

II. Market Basket Analysis Using Association Rule Mining Strategy

The term market basket analysis using association rule learning in the retail business refers to the research that provides the retailer with information to understand the purchasing behavior of a customer. This association information will enable the retailer to understand the customer's needs and rewrite the store's layout accordingly, develop cross-promotional programs, or even capture new customers. An example for this was when one super market chain discovered in its analysis that customers who bought diapers often bought beer as well, have put the diapers close to beer coolers, and their sales increased dramatically.

Let I = {i1,…., in} be a set of items. Let D be a database which contains set of transactions. Each transaction t Є D is an item set such that t is a proper subset of I. A transaction t supports X, a set of items in I, if X is a proper subset of t. Assume that the items in a transaction or an item set are sorted in lexicographic order. An association rule is an implication of the form

X□Y, where X and Y are subsets of I and X∩Y= Ø.

The support of rule X□Y can be calculated by the following equation:

$$\text{Support}(X \square Y) = \frac{|X \square Y|}{|D|}$$

where |X□Y| denotes the number of transactions containing the itemset XY in the database, |D| denotes the number of the transactions in the data.

$$\text{Confidence}(X \square Y) = \frac{|X \square Y|}{|X|}$$

where |X| is number of transactions in database D that contains itemset X.

A rule X□Y is strong if support(X□Y) ≥ min_support and confidence(X□Y) ≥ min_confidence, where min_support and min_confidence are two given minimum thresholds.

III. Association Rule Mining Algorithms

The algorithms for discovering large itemsets make multiple passes over the data. In the first pass, the algorithm counts the support of individual items and determines which of them are large, i.e. have minimum support. In each subsequent pass, it starts with a seed set of itemsets found to be large in the previous pass. This seed set is used for generating new potentially large itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, it determines which of the candidate itemsets are actually large, and they become the seed for the next pass. This process continues until no new large itemsets are found. Many algorithms for generating association rules were presented over time. Some well known algorithms are Apriori, Eclat and FP-Growth. Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support.

A. Apriori Algorithm

Agrawal et al [1] has presented the first association rule mining algorithm and is named as Apriori algorithm. The algorithm is based on the a *priori hypothesis*, viz., an itemset can only be frequent if all its subsets are also frequent. This information makes possible to reduce the search space efficiently when discovering the frequent itemsets, because using this information reduces the number of candidates. The Apriori algorithm is a level-wise method, which means that it discovers the *k*-itemsets during k^{th} database scan. The algorithm counts the items in the transactions and discards the infrequent items in the first scan of the database. In this way the frequent 1-itemsets are found. From these frequent items two candidates are generated by creating all the combination of them by keeping the lexicographic order. Formally, items x and y form a candidate (x,y) when x ≤ y. During the, second database scan the support of the 2-candidates are counted. After a database reading, the counters of the candidates are checked with the minimum support threshold. If a counter exceeds the threshold, the candidate belonging to it becomes frequent, otherwise it is discarded. The 3-candidates are generated from the frequent 2-itemsets regarding the following rule.

Let be given two itemsets (i_1,i_2) and (i_3,i_4) where $i_1<i_2$ and $i_3<i_4$ as mentioned earlier. The two itemsets can form a 3-candidate if $i_1=i_3$ and (i_2,i_4) is also frequent. Fulfilling the second condition means that the a priori hypothesis is fulfilled. The resulting 3-candidate is the following: (i_1,i_2,i_4). In general two k-itemsets are joined by keeping the lexicographic order to form a $(k + 1)$-itemset if the first k-1 items of them are in common and all the $(k$-1)-subsets of the resulting candidate are frequent. The algorithm terminates if no candidates can be generated or no frequent itemsets are found.

IV. Problem Statement

Data mining represents a wide range of tools and techniques to extract useful information which can contain sensitive information from a large collection of data. Data should be manipulated or distorted in such a way that information cannot be discovered through data mining techniques. Sensitive information has to be protected against unauthorized access. Privacy preserving data mining considers the problem of data mining. Association rule hiding is one of the privacy preserving techniques for hiding sensitive association rules. For hiding sensitive association information in market basket database the sensitive items are given as input. Detection of sensitive items manually in market basket database containing thousands of transactions takes more computational time. So there is a need to find a method to detect the sensitive items.

V. Related Works

In [6], the authors hide the sensitive knowledge in association rule mining, by selecting the sensitive rules manually.

In [7] authors introduced a heuristic approach to hide restrictive association rules that requires the sensitive items. The sensitive items have to be identified manually and it has to be removed from its transactions.

In [8], the authors selected the sensitive attributes and hide the attributes by randomization process.

In [9], the authors select the sensitive item(s) using the frequent item sets and the association rules. By selecting the consequent of the rule, the sensitive items can be identified.

Instead of selecting the sensitive items manually, the proposed method helps to detect the sensitive items.

VI. Proposed Method

The proposed method first finds the number of occurrences of individual items. Using maximum number of frequent items, association rules are generated based on minimum support and minimum confidence. This approach uses the concept of representative rules. Having the frequent items on LHS of the rule, its consequents are combined in RHS. The consequents are selected as sensitive items. More number of sensitive items are selected by varying minimum support and minimum confidence, which are user specified values.

A. Representative Association Rule

The number of association rules discovered in a given database is very large. It is noted that a significant percentage of these rules are redundant and useless. A user should be presented with all of them, which are original, novel, and interesting. To address this issue, [10] introduced a notion for concise representation of association rules, called representative rules (RR).

RR is a least set of rules that allow deducing all association rules without accessing a database. In a notion of cover operator was introduced for driving a set of association rules from a given association rule. The cover of the rule $X \Rightarrow Y$, $Y \neq \phi$, is defined as follows:

$$C(X \Rightarrow Y) = \{X \cup Y \Rightarrow V \mid Z,\ V \subseteq Y \text{ and } Z \cap V = \phi \text{ and } V \neq \phi\}$$

$C(X \Rightarrow Y)$ consists of a subset of items occurring in the rule $X \Rightarrow Y$. The number of different rules in the cover of the association

$X \Rightarrow Y$ is equal to $3^m - 2^m$, $m = |Y|$.

In general, the process of generating representative rules may be decomposed in to two sub processes: frequent item-sets generations and generation of RR from frequent item-sets. Let be a frequent itemset and $\phi \neq X \subset Y$. The association rule $X \Rightarrow Z$ / X is representative rule if there is no association rule $(X \Rightarrow Z' / X)'$ where $Z \subset Z'$, and there is no association rule $(X' \Rightarrow Z / X')$ such that $X \supset X'$. Formally, a set of representative rules (RR) for a given association rules (AR) can be defined as follows:

$$RR = \{r \in AR \mid \neg \exists\, r' \in AR,\ r' \neq r \text{ and } r \in C(r')\}$$

Each rule in RR is called representative association rule and no representative rule may belong in the cover of another association rule [11], [12].

Algorithm: Detection of Sensitive Items
Input:
(1) Database D
(2) Minimum support: min_supp
(3) Minimum confidence: min_conf
Output: Sensitive item(s) for Sensitive Association rule hiding.
Algorithm:

1. Find the frequencies of individual items and store it in F.
2. Generate the association rules from D using minimum support and minimum confidence.
3. Select the single antecedent and consequent rules.
4. For each frequent item f F {
5. If F is empty then EXIT;
6. Select all the rules with min_supp containing f and store in U//f on LHS
7. Repeat {
8. Select all the rules from U with same LHS
9. Join RHS of selected rules and store in R; //make them representative rules
10. }Until (U is empty);
11. Selected RHS will be the sensitive item(s).

Maximum occurrences of items are selected by user specified value in terms of percentage. The proposed algorithm first tries to find the frequent items and the association rules are generated using the Apriori algorithm [1]. These rules are used to form representative association rules having frequent items on LHS. The selected RHS are considered as sensitive items. The selected sensitive items are used to hide sensitive rules.

VII. Experimental Results

TABLE I: Transactional Dataset

TID	*ITEMS*
T1	ABC
T2	ABCD
T3	BCE
T4	ACDE
T5	DE
T6	AB

Frequent items are selected based on the support of individual items. This is shown in Table-II.

TABLE II FREQUENT ITEMS

Items	*No. Of Occurrences*
A	4
B	4
C	4
D	3
E	3

For the Dataset given in Table - I min_supp is taken as 33% and a min_conf is taken as 70 % and frequent item F={A}.

Association rules are generated using apriori algorithm. The rules containing single antecedent and consequent are selected. All the rules containing 'A' either in RHS or LHS are chosen and represent them in representative rule format. Association rules are shown in Table III.

TABLE III: ASSOCIATION RULES

AR	*SUPP*	*CONF*
A => C	50	75
C => A	50	75
A, D => C	33.333	100
C, D => A	33.333	100
A => B	50	75
B => A	50	75
C=>B	50	75
B=>C	50	75

From these association rules, the rules containing frequent item A with single antecedent and consequent are selected. They are A=>C(50,75), C=>A(50,75), A,D=>C(33.33,100), C,D=> A(33,100), A=>B(50,75), B=>A(50, 75).

From this rules set, select the rules that can be represented in the form of representative rules. Like the rules A=>B and A=>C can be represented as

A=> BC

In the above representative rules B and C are considered as sensitive items. Similarly the frequent items B and C are taken and representative association rules are formed. From the representative association rules, sensitive item A is detected.

Hence, for the given transactional data set A, B and C are detected as sensitive items. More number of sensitive items can be selected by varying minimum support and minimum confidence of frequent items.

VIII. CONCLUSION AND FUTURE WORK

Market basket analysis is an essential activity to promote business. Buyer behavior patterns are discovered through market basket analysis. The knowledge discovered may reveal trade secrets also. The data owner may lose his business secrets. Privacy preserving data mining hides certain sensitive information, so that they cannot be discovered through data mining techniques.

Association rule hiding algorithm sanitizes the database such that no sensitive association rule is derived from it. For hiding sensitive association rules, sensitive items are selected manually. In market basket database the selection of sensitive items takes more computational time.

The proposed algorithm detects the sensitive items using a set of representative association rules. It is experimented with sample data sets. Further research work is in progress to hide the sensitive association rules using the selected sensitive items.

REFERENCES

[1] Agrawal R, Imielnski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Washington DC

[2] Agrawal R, Srikant R (2000) Privacy preserving data mining. In ACMSIGMOD Conference on Management Of Data, Dallas, Texas, pp 439–4501.

[3] Clifton C, Marks D (1996) Security and privacy implications of data mining. In: SIGMOD Workshop on Research Issues on Data Mining and knowledge Discovery.

[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] Evfimievski A (2002) Randomization in privacy preserving data mining. SIGKDD Explorations 4(2), Issue 2:43–48.

[6] Oliveira S, Zaiane O (2003 a) Algorithms for balancing privacy and knowledge discovery in association rule mining. In: Proceedings of7th International Database Engineering and Applications Symposium (IDEAS03), Hong Kong

[7] Oliveira S, Zaiane O (2003 b) Protecting sensitive knowledge by data sanitization. In: Proceedings of IEEE International Conference on Data Mining.

[8] Wenliang Du , Zhijun Zhan(2003)Using Randomized Response Techniques for Privacy Preserving Data mining.SIGKDD'03

[9] Dr. K.Duraisamy, N.Maheswari(2008) Identification of Sensitive Items in Privacy Preserving - Association Rule Mining,CCSE journal.

[10] Marzena Kryszkiewicz. "Representative Association Rules", In proceedings of PAKDD'98, Melbourne,Australia(Lecture notes in artificialIntelligence,LANI 1394, Springer-Verleg,1998, pp 198-209.

[11] Yiqun Huang, Zhengding Lu, Heping Hu, "A method of security improvement for privacy preserving association rule mining over vertically partitioned data", 9th International Database Engineering and Application Symposium, pp. 339 – 343, 2005.

[12] Saygin Y., Verykios V.S. and Elmagarmid A.K., "Privacy preserving association rule mining," IEEE Proceedings of the 12th Int'l Workshop on Research Issues in Data Engineering, pp. 151 – 158, 2002.

[13] Vassilios S. Verykios,, Ahmed K. Elmagarmid , Elina Bertino, Yucel Saygin, Elena Dasseni. "Association Rule Hiding", IEEE Transactions on knowledge and data engineering, Vol.6, NO.4, April 2004.

[14] Shyue-Liang Wang, Yu-Huei Lee, Billis S., Jafari, A. "Hiding sensitive items in privacy preserving association rule mining", IEEE International Conference on Systems, Man and Cybernetics, Volume 4, 10-13 Oct. 2004 Page(s): 3239 – 3244

[15] Aris Gkoulalas–Divanis;Vassilios S. Verykios "Association Rule Hiding For Data Mining" Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC 2010.