Privacy Preserving Data Mining: A Novel Approach to Secure Sensitive Data Based on Association Rules

Syed Shujauddin Sameer

Abstract— The availability of data on the internet is increasing on a larger basis daily .Privacy Preservation data mining has emerged to address one of the side effects of data mining Technology. The threat to individual privacy through data mining is able to infer sensitive information from Non-sensitive information or unclassified data. There is a n urgent need to be able to infer some mechanism to avoid the projection of all the sensitive information .An approach in data mining techniques is very much essential. Alteration of data, filtering of the data, blocking of the data are Some of the approaches. Given specific rules to be hidden, the techniques involve is to hide only the given sensitive data. In this work we assume that only sensitive datais given and we analyze the approaches to secure sensitive data in the database.

Index Terms— Privacy preserving data mining, Association rules ,Sensitive data.

I. INTRODUCTION

Privacy preservation is important for data mining and other learning techniques.There is a need for different approaches required in this scenario[1].A fruitful direction for future data mining research will be the development of techniques that incorporate privacy concerns [4]. Explosive progress in networking, storage, and processor technologies has led to the creation of ultra large databases that record precedent amount of transactional information.

Privacy issues are further exacerbated, now that the World Wide Web makes it easy for the new data to be automatically collected and added to databases. Data mining, with its promise to evidently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse. The primary task in data mining is the development of models about aggregated data. Repositories of data contain sensitive information which must be protected against unauthorized access. The protection of the confidentiality of this information has been a long-term goal for the database security research community and the government statistical agencies. Consider a scenario where two parties having private databases wish to cooperate by computing a data mining algorithm on the union of their databases. Since the databases are confidential, neither party is willing to divulge any of the contents to the other.Recent advances, in data mining and machine learning algorithms, have increased the disclosure risks one may encounter when releasing data to outside parties.

Manuscript received on April 2014.

Syed Shujauddin Sameer,,Department of Computer Science ,King Khalid University, Abha, Saudi Arabia.

A key problem, and still not sufficiently investigated, is the need to balance the confidentiality of the disclosed data with the legitimate needs of the data users. Every disclosure limitation method affects, in some way, and modifies true data values and relationships.In this paper, we investigate confidentiality issues of a broad category of rules, which are called association rules. Sometimes, sensitive rules should not be disclosed to the public since, among other things, they may be used for inferencing sensitive data, or they may provide business competitors with an advantage.Many government agencies,

businesses and non-organizations in order to support their short and long term planning activities, they are searching for a way to collect, analyze and report data about individuals, households or businesses. Information systems, therefore, contain confidential information such as social security numbers, income, credit ratings, type of disease, customer purchases, etc [10]. Ideally, these effects can be quantified so that their anticipated impact on the completeness and validity of the data can guide the selection and use of the disclosure limitation method.

There have been many approaches in the regard for privacy preservation of the association rules. One of the approach is to alter the data before delivery to the data miner so that real values are obscured. One technique of this approach is to selectively modify individual values from a database to prevent the discovery of a set of rules [10,11,18,20].Another approach is to allow users access to only a subset of rules while global data mining results can still be discovered. The elicitation of knowledge that can be attained by such techniques has been the focus of the Knowledge Discovery in Databases (KDD).

II. PROBLEM DESCRIPTION

There has been extensive research in the area of statistical databases motivated by the

desire to be able to provide statistical information (sum, count, average, maximum,minimum, pth percentile, etc.)[4] without compromising sensitive information about individuals. The proposed techniques can be broadly classified into data perturbation.The perturbation family includes swapping values between records, replacing the original database by a sample from the same distribution, adding noise to the values in the database, adding noise to the results of a query, and sampling the result of a query. As we will see, it is sufficient for us to be able to reconstruct with sufficient accuracy the original distributions of the values of the confidential attributes. We adopt from the statistics literature two methods that a person may use in a system to modify the value. These methods and the level of privacy they

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

provide in the next section. Value dissociation is the third method. In this method, a value returned for a field of a record is a true value, but from the same field in some other record. Interestingly, a recent proposal to construct perturbed training sets is based on this method. Secure two party computations were first investigated by Yao, and were later generalized to multi-party
computation [9].
Let I ={ i1,i2 ,i3 ,..im } be a set of literals, called items. Given a set of
transactions D, where each transaction T is a set of items such that T C I,
an association rule is an expression X => Y where X C I,Y C I, and X ∩ Y =Φ .
The X and Y are called respectively the body (left hand side) and head (right hand side) of the rule. An example of such a rule is that 90% of customers buy hamburgers also buy Coke. The 90% here is called the confidence of the rule, which means that 90% of transaction that contains X also contains Y. The confidence is calculated as |XUY | / |X|. The support of the rule is the percentage of transactions that contain both X and Y, which is calculated as| X U 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 item sets, while the support of a rule measures the significance of the correlation between item sets. The problem of mining association rules is to find all rules that are greater than the user-specified minimum support and minimum confidence. As an example, for a given small Transactional database, a minimum support of 33% and a minimum confidence of 70%, nine association rules can be found as follows: B=>A(66%, 100%), C=>A (66%, 100%), B=>C(50%, 75%), C=>B (50%,75%),AB=>C (50%,75%),
AC=>B(50%,75%),
BC=>A(50%,100%),C=>AB(50%,75%),
B=>AC(50%, 75%).

I. Transaction Table

TID	ITEM
T1	ABC
T2	ABC
T3	ABC
T4	AB
T5	A
T6	AC

However, mining association rules usually generates a large number of rules, most of which are unnecessary for the purpose of prediction.[3] For example, given item set for prediction P = {C}, the rule set that contains only two rules C=>A (66%, 100%), C=>B (50%, 75%), will generate the same predicted item set Q = {A, B} as the nine association rules found from the above database . A predictive association rule set (or informative rule set) [15] can be informally defined as the smallest rule set that makes the same prediction as the association rule set by confidence priority The objective of data mining is to extract hidden or potentially unknown interesting rules or patterns from databases. However, the objective of privacy preserving data mining is to hide certain sensitive information so that they cannot be discovered through data mining techniques [2,5,13,16]. We assume that only sensitive items are given and there are two simple approaches to modify data in database so that sensitive predictive association rules cannot be inferred through association rule
mining. More specifically, given a transaction database D, a minimum support, a minimum confidence and a set of sensitive items X, the objective is to modify the
database D such that no predictive association rules containing X on the left hand side will be discovered. As an example, for a given database in Table 1, a minimum support of 33%, a minimum confidence of 70%, and a hidden item X = {C}, if transaction T5 is modified as AC, then the following rules that contain item C on the left hand side will be hidden: C=>B (50%, 60%), AC=>B (50%,60%), C=>AB(50%,60%).

A. *Algorithms*

In order to hide an association rule, we can either decrease its support or its confidence to be smaller than pre-specified minimum support and minimum confidence. To decrease the confidence of a rule, we can either (1) increase the support of X, i.e., the left hand side of the rule, but not support of X U Y, or (2) decrease the support of the item set XU Y. For the second case, if we only decrease the support of Y, the right hand side of the rule, it would reduce the
confidence faster than simply reducing the support of X U Y. To decrease support of an item, we will modify one item at a time in a selected transaction by changing from 1 to 0 and from 0 to 1 to increase the support. Based on these two strategies, two data mining algorithms are available for hiding sensitive predictive association rules, namely Increase Support of LHS (ISL) and
Decrease Support of RHS (DSR). The first algorithm tries to increase the support of left hand side of the rule. The second algorithm tries to decrease the support of the right hand side of the rule.
Algorithm ISL
Input: (1) a source database D,
(2) a min_support,
(3) a min_confidence,
(4) a set of hidden items X
Output: a transformed database D', where rules containing X on LHS will be hidden
1. Find large 1-item sets from D;
2. For each hidden item x€ X
3. If x is not a large 1-itemset, then X :=X-{x} ;
4. If H is empty, then EXIT;// no AR contains X in LHS
5. Find large 2-itemsets from D;
6. For each x € X {
7. For each large 2-itemset containing x {
8. Compute confidence of rule U, where U is a rule like x → h;
9. If confidence (U) < min_conf, then
10. Go to next large 2-itemset;
11. Else { //Increase Support of LHS
12. Find TL = { t in D | t does not support U} ;
13. Sort TL in ascending order by the number of Items;
14. While {confidence(U) >= min_conf and TL is not empty) {
15. Choose the first transaction t from TL;
16. Modify t to support x, the LHS(U);
17. Compute support and confidence of U;
18. Remove and save the first transaction t from TL;
19. }; // end While

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

20. }; // end if
21. If TL is empty, then {
22. Can not hide x → h;
23. Restore D;
24. Go to next large-2 item set;
25. } // end if TL is empty
26. } //end of for each large 2-itemset
27. Remove x from X;
28. } // end of for each x
29. Output updated D, as the transformed D';

B.Examples

This section shows four examples for demonstrating the two proposed algorithms in hiding sensitive predictive association rules in the association rule mining.
Example 1 Assuming that the min_supp = 33% and min_conf = 70%, the result of hiding item C and then item B using ISL algorithm is as follows. To hide item C, the rule C => B (50%, 75%) will be hidden if transaction T5 is modified from 100 to 101 using ISL (Increase Support of LHS).

II. SANITIZED DATABASE

TID	D	D1
T1	111	110
T2	111	111
T3	111	111
T4	110	110
T5	100	110
T6	101	101

III.RELATED WORK

Generation of the rules for the given database area available. Here we provide the mechanism for the analysis of the required approaches . In the algorithm approach the selection of the table is directly related to process of the association rule mining using the Apriori algorithm. Apriori is one of the algorithmns to mine rules based on the given database. The rules generated are mined as per the given support and confidence levels .The rules are displayed to the user with the columns of antecedent, consequent, support and confidence. The rules generated can also be saved in a file to be viewed later . we provide two options for selecting a database either from MSAccess or IBM DB2.The later can be used to store very large datasets which provide necessary information for analysis.

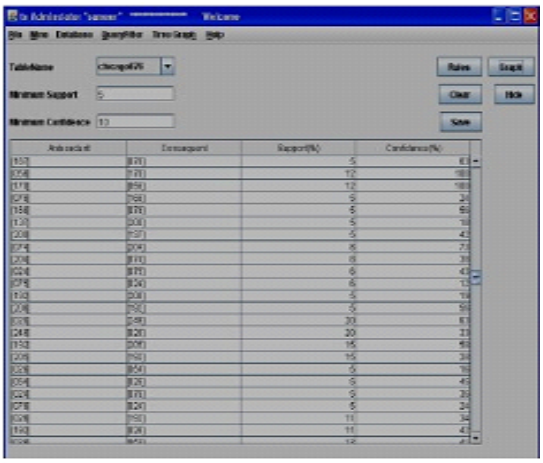

Fig a .Association Rules

A graph which displays the line chart mapping of the support and confidence values with respect to association rules is also necessary to give us a good presentation. Each association rule has a degree of support and confidence as per the user threshold values. The X axis is used to represent the rules and the Y axis is used to represent the numerical value of the support and the confidence.

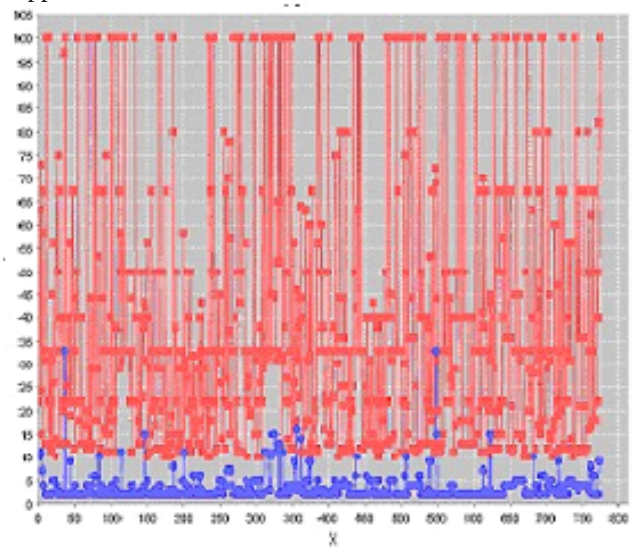

Figure 2..Line Chart of Support , Confidence with Rules

The hide module is used to hide the required association rule containing the hidden item selected by the user in the consequent. Both the algorithms i.e. ISLF, increased support of left count and DSRF, decrease support of right first can be used .The interface displays the original rules on lhs side while the changed rules are displayed on the right side .

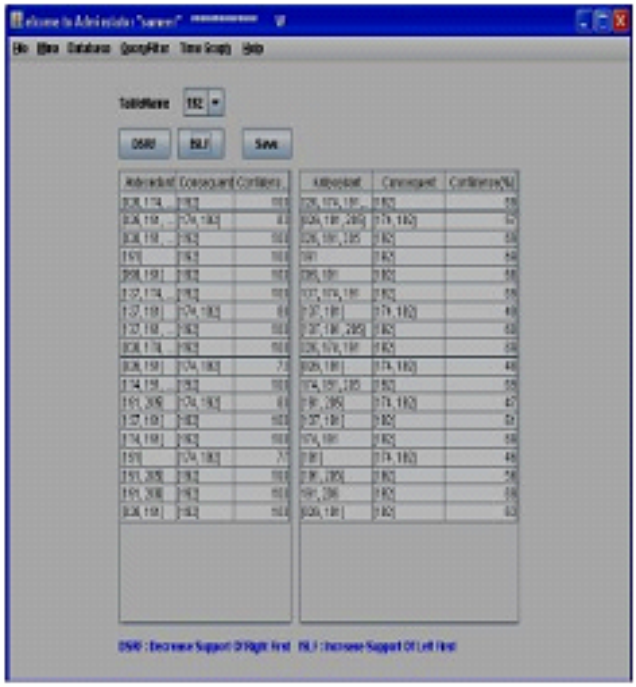

Fig 3. Algorithmns

Sanitization of the data may be required but is not a good idea for a consistent database .The number of rules would be reduced naturally. The Second approach is that os query analysis . Database vendors have made it possible to apply predictive models on relational data using SQL extensions. The predictive models can either be built naively or imported using other interchanges format. This enables us to express queries containing mining predicates. We need to optimize

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

queries containing such mining predicates. [21] .The approach is to display the results for the user without any changes to be made for the database .The database change would require time to transform the data into a new sanitized database. The process of hiding requires a new database to be created for testing the results whether a particular item is hidden or not .The filtering process removes the problem of duplication of the database by providing the user with the interface avoiding to change the database to get the required result. Here we do not require any separate table to be maintained for the process of hiding the association rule. The query processing algorithm executes the query and if the result set contains the hidden item selected by the administrator the item is hidden and the result set is displayed without containing the hidden item.The final results are displayed without the hidden item being provided. We have avoided the time required to create a new database to store the result. The user is also provided by the interface to select the transaction as per the user. The transactions are created as the user executes the query. The rules are mined as per the transactions selected by the user using the same Apriori algorithm.

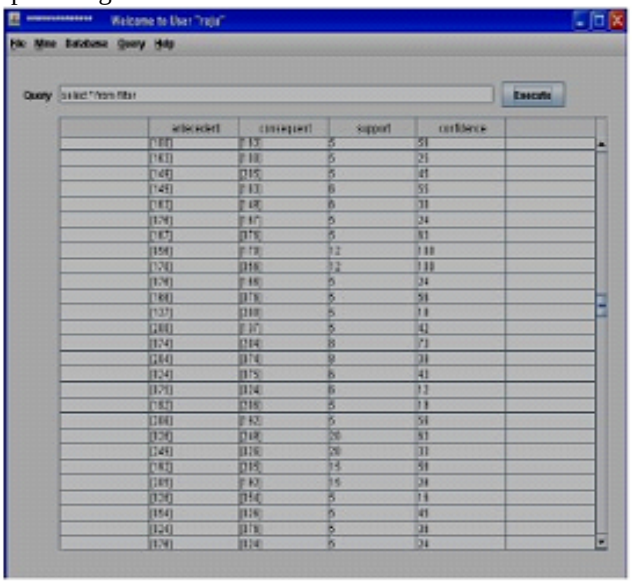

Fig 4.Query Analysis

The third approach is to block certain data from the user .The user may be required to enter the query every time the results are required.

Blocking the rules to be displayed is also used in this scenario which would work effectively as it required firstly no change to the database.The system would hide the rules based on sensitive data .The users may not be able to infer all the data and would lead to a secure database not altered for future use.

IV.CONCLUSION

We have analysed three different techniques for securing the sensitive data from the user First technique is not suitable as it requires changing the database .The query analyst requires some effort by the user. The third technique is likely to block the access to certain type of data. Keeping the data secure is the primary task of Privacy Preserving data mining.

V.FUTURE WORK

Different techniques need to be analysed as the data from the internet is increasing largely.More deep analysis has to be done by relating to different data sets using different technologies.

REFERENCES

[1] Privacy Preserving Decision Tree Learning Using Unrealized Data Sets ,Issue No 02- February 2012.Vol. 24,Knowledge and Data Engineering ,IEEE Transactions

[2] D. Agrawal and C. C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms”, In Proceedings of the 20th Symposium on Principles of Database Systems, Santa Barbara,California, USA, May 2001.

[3] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases”, In Proceedings of ACM SIGMOD International Conference on Management of Data, Washington DC, May 1993.

[4] R. Agrawal and R. Srikant, ”Privacy preserving data mining”, In ACM SIGMOD Conference on Management of Data, pages 439–450, Dallas, Texas, May 2000.

[5] Ljiljana Brankovic and Vladimir Estivill-Castro, “Privacy Issues in Knowledge Discovery and Data Mining”, Australian Institute of Computer Ethics Conference, July, 1999, Lilydale.

[6] C. Clifton and D. Marks, “Security and Privacy Implications of Data Mining”, in Workshop on Research Issues on Data Mining and knowledge Discovery, 1996.

[7] C. Clifton, “Protecting Against Data Mining Through Samples”, in Proceedings of the ThirteenthAnnual IFIP WG 11.3 Working Conference on Database Security, 1999.

[8] C. Clifton, “Using Sample Size to Limit Exposure to Data Mining”, Journal of Computer Security, 8(4), 2000.SIGMOD

[9] Chris Clifton, Murant Kantarcioglu, Xiaodong Lin and Michael Y. Zhu, “ Tools for Privacy PreservingDistributed Data Mining”, SIGKDD Explorations, 4(2), 1-7, Dec. 2002.

[10] E. Dasseni, V. Verykios, A. Elmagarmid and E. Bertino, “Hiding Association Rules by Using Confidence and Support” in Proceedings of 4th Information Hiding Workshop, 369-383, Pittsburgh,PA, 2001.

[11] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules”, In Proc. Of the 8th ACM SIGKDD Int’l Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, July 2002.

[12] Alexandre Evfimievski, “Randomization in Privacy Preserving Data Mining”, SIGKDD Explorations, 4(2), Issue 2, 43-48, Dec. 2002.

[13] Alexandre Evfimievski, Johannes Gehrke and Ramakrishnan Srikant, “Limiting Privacy Breaches in Privacy Preserving Data Mining”, PODS 2003, June 9-12, 2003, San Diego, CA.

[14] M. Kantarcioglu and C. Clifton, “Privacy-preserving distributed mining of association rules onhorizontally partitioned data”, In ACM SIGMOD Workshop on Research Issues on Data Miningand Knowledge Discovery, June 2002.

[15] Y. Lindell and B. Pinkas, “Privacy preserving data mining”, In CRYPTO, pages 36–54, 2000.

[16] D. E. O’ Leary, “Knowledge Discovery as a Threat to Database Security”, In G. Piatetsky-Shapiro and W. J.Frawley, editors, Knowledge Discovery in Databases, 507516, AAAI Press/ MIT Press, Menlo Park, CA, 1991.

[17] S. Oliveira, O. Zaiane, “Algorithms for Balancing Privacy and Knowledge Discovery i Association Rule Mining”, Proceedings of 7th International Database Engineering and ApplicationsSymposium (IDEAS03), Hong Kong, July 2003.

[18] S. Oliveira, O. Zaiane, “Protecting Sensitive Knowledge by Data Sanitization”, Proceedings of IEEE International Conference on Data Mining, November 2003.

[19] S. J. Rizvi and J. R. Haritsa, “Privacy-preserving Association rule mining”, In Proc. of the 28th Int’l Conference on Very Large Databases, August 2002.

[20] Y. Saygin, V. Verykios, and C. Clifton, “Using Unknowns to Prevent Discovery of Association Rules”, SIGMOD Record 30(4): 45-54, December 2001.

[21] Surajit Chaudhuri. “Efficient evaluation of queries with mining predicates”, In Proc. ofthe 18th Int’l Conference on Data Engineering (ICDE) 529-540,2002.

Syed Shujauddin Sameer is graduated in B.Tech (Computer Science Engineering)in 2003 from the Kbnce , VTU University, Banglore. He did his M.Tech in Computer Science Engineering , Jits ,JNTU , Hyderabad during 2009.His areas of interest includes Data Warehousing And Data Mining,Artificial Intelligence,Simulation and Modelling.He is Presently working as an Lecturer in CSE dept, King Khalid University, Abha,Saudi Arabia.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.