International Journal of Scientific Research in Computer Science and Engineering
Survey Paper Vol-1, Issue-3 ISSN: 2320–7639

A Survey: Preventing Discovering Association Rules For Large Data Base

Mohnish Patel[#1], Aasif Hasan[2] , Sushil Kumar[3]

[#1]*Computer Science & Engineering, RGPV, Bhopal, India*
[2]*Computer Science & Engineering, RGPV, Bhopal, India*
[3]*Computer Science & Engineering, RGPV, Bhopal, India*

Available online at www.isroset.org

Received: 17 March 2013 Revised: 28 March 2013 Accepted: 19 June 2013 Published: 30 June 2013

***Abstract*—**Data products are designed to inform public or business policy, and research or public information. Securing these products against unauthorized accesses has been a long-term goal of the database security research community and the government statistical agencies. Whether data is personal or corporate data, data mining offers the potential to reveal what other regard as sensitive (private). In some cases, it may be of mutual benefit for two parties (even competitors) to share their data for an analysis task. Sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy, as we will indicate. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the "database inference" problem.

Keywords- Association Rule mining, Data mining

I. INTRODUCTION

Data mining technology help us in extraction of useful knowledge from large data sets. The process of data collection and data dissemination may, however, result in an inherent risk of privacy threats. Some private information about individuals, businesses and organizations has to be suppressed before it is shared or published. The privacy-preserving data mining (PPDM) [13] has thus become an important issue in current years. In this paper, we propose an evolutionary privacy-preserving data mining technology to find appropriate transactions to be hidden from a database [14].

The important aspect of the research was to use text mining techniques to mine the data in a group of documents and to determine what are the common characteristics among them and then to determine other documents which contains these selected characteristics. Unregulated data can be documented as any data that is not found in a database. MP3, Images, video files could be categorized as non-textual unregulated data whereas memos, email messages, word processor documents could be categorized as textual unregulated data. During this research, we concentrate on unregulated data in textual form.

Following are the basic concepts to be take care with data mining:

A. Privacy Preserving Mining

Privacy preserving data mining [1] is a novel research area in data mining where data mining algorithms are analyzed for the side effects they incur in data privacy. The objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. There are many approaches which have been adopted for privacy preserving data mining, Shivakumar, 2003 [7].

Corresponding Author: *Mohnish Patel*

B. Data distribution

Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios are classified as horizontal data distribution and vertical data distribution, Jaideep vaidhya, 2002, [9]. Horizontal distribution refers to cases where different database records reside in different places, while vertical data distribution, refers to the cases where all values for different attributes reside in different places.

C. Data Modification

Data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Perturbation, which is accomplished by the alteration of an attribute value by a new value Aggregation or merging which is the combination of several values into a coarser category Swapping that refers to interchanging values of individual records.

D. Data Mining Algorithm

This is something that is known beforehand, but it facilitates the analysis and design of the data hiding algorithm.

E. Data or rule hiding

The complexity for hiding aggregated data in the form of rules is higher, and so, 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.

F. Privacy Preserving

Privacy preservation is technique [2][11] used for the selective modification of the data. Selective modification is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized. Heuristic-based techniques like adaptive modification that modifies only selected values that minimize the utility loss rather than all available values. Cryptography-based [8] techniques like secure multiparty[11] computation where a computation is secure if at the end of the computation.

II. Related Work

The author presents two secure protocols for securely checking whether k-anonymous database retains its anonymity, once a new record is being inserted to it. Since the implemented protocols make sure that the updated data remains K-anonymous, the results came from a user's query are also k-anonymous [15]. Thus, the patient or the data provider's privacy cannot be profaned from any query. As long as the database is updated properly using the proposed protocols, the user queries under our application domain are always privacy preserving.

The author shows the phenomenon to achieve privacy and collaborative data mining at the same time. The objective of this paper is to present technology which solve privacy-preserving collaborative data mining [16] problems over large data sets. This paper contains the following: (1) a definition of privacy for privacy-preserving collaborative data mining; (2) an solution for naive Bayesian classification with vertical collaboration; and (3) an efficiency analysis to show the performance scaling up with various factors.

Here, the author proposes a k-anonymization [17] solution for classification. This research's goal is to find a k-anonymization, not necessarily optimal for minimizing data distortion, to preserve the classification structure. The author conducted intensive experiments to calculate the impact of anonymization over the classification on future data. The quality of classification could be preserved even for highly restrictive anonymity requirements by performing experiments on real life data.

Here, the author has developed three protocols for perturbation unification. The key drawback of applying geometric data [18] perturbation in multiparty collaborative mining is to firmly unify multiple geometric perturbations that are most liked by different parties, severally.

Here the author implements three distinctive features compared to the prevailing approaches: 1) with geometric data perturbation, these protocols will work for several existing standard data mining algorithms, whereas most of alternative approaches are only designed for a specific mining algorithm; 2) each of the two major factors: data utility and privacy guarantee are well preserved, compared to other perturbation based approaches; and 3) two of the three proposed protocols even have great scalability in terms of the total number of participants, whereas several existing cryptographic approaches take into account only two or a few more participants.

The author Agrawal & Srikant and Lindell & Pinkas [19] introduce the term "privacy preserving data mining" in there research in 2000. These papers thought of two elementary issues of Privacy Preserving Data Mining, privacy preserving data collection and mining a dataset partitioned across many private enterprises. Agrawal and Srikant (2000) devised a randomization algorithm that permits an oversized variety of users to contribute their private or secure records for economical centralized data mining whereas limiting the disclosure of their values; Lindell and Pinkas (2000) unreal a scientific discipline on cryptographic protocol for decision tree construction over a dataset horizontally partitioned between two parties.

III. Privacy Preserving Techniques

Various methods are given for privacy in data mining association rules. Different methods with their advantage and disadvantage are explained below.

A. Random Data Perturbation Technique

The random value perturbation method Siva Kumar, 2003 [6] attempt to preserve privacy of the data by modifying values of the sensitive attributes. In this approach, the owner of a dataset returns a value u+v, where u is the original data, and v is a random value drawn from a certain distribution. The owner of the data provides the perturbed values u1+v1, u2+v2.... and the cumulative distribution function fV(v) of v. The reconstruction of data is done by estimate the distribution fU(u) of the original data, from the perturbed data. Where fU(u) is the value in previous step. This method gives distribution function of data. This formula is taken from bays rule.

B. ISL (Increase Support LHS) Algorithm

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. To hide an association rules the ISL Ayat Jafari, Wang 2005 [3] increases the support of the left hand side of the rule by modifying one item at a time in a selected transaction by changing its value from 0 to 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%).

C. DSR (Decrease Support RHS) Algorithm

Association rule X=>Y will be hided [4][5][12] if the support of the itemset XχY is deceased or the support of Y (the right hand side of rule) is decreased. DSR algorithm Ayat Jafari, Wang, 2005 [4] decreases the support of the right hand side of the rule by modifying one item at a time in a selected transaction by changing its value from 1 to 0.

IV. PROBLEM DOMAIN

A. This method can only reconstruct the distribution of the original data from the data perturbed by random value distortion. But it does not consider estimation of the individual values of the data-points. For avoid such problem spectral filtering technique was given.

B. If the actual dataset has a random component in it, and random noise is added to perturb it, spectral filtering method does not filter the actual data accurately. For avoid such problem random projection matrices method was given.

C. For avoid such problem other method Bottom up generalization method was given. Providing just the correlation matrix to the client party does not meet the overall objective it does not work for heterogeneous distributed data sets.

V. SOLUTION DOMAIN

In this thesis assuming that only sensitive items are given we implement two existing algorithms, ISL (Increase Support of LHS) and DSR (Decrease Support of RHS), we then propose a new approach based on these two existing algorithm. We also compare all three algorithms on the basis of number of database scans and no. of rule prune We have implemented three algorithms in which two are existing algorithm and one is proposed algorithm .We have compared these algorithms on the basis of number of hidden rule. The existing algorithm ISL and DSR algorithm scan the database for same times but this algorithm in rules. Thus the proposed algorithm is able to hide more number of rules and thus perform better privacy preserving mining 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 increase the support of *X* the left hand side of the rule, but not support of *X U Y*, or decrease the support of the item set *X U Y*. For the second case, we propose two data mining algorithms for hiding sensitive predictive association rules, namely **I**ncrease **S**upport 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

VI. APPLICATION DOMAIN

We have compared these algorithms on the basis of number of hidden rule. The existing algorithm ISL (increase support of LHS) DSR (decrease support of RHS) algorithm scan the database for same times but this algorithm hides rules. Our proposed algorithm hides rules .Thus the proposed algorithm is able to hide more number of rules and thus perform better privacy preserving mining. Which is based on two previous existing algorithms ISL and DSR we decrease support of right hand side of the rule in a rule where item to be hide is in right side. After that we increase support of left hand side of rule, where item to be hide is in left side. The experiments are pursued on both synthetic and real data sets. The synthetic data sets which we used for our experiments were generated using the procedure described in We refer readers to it for more details on the generation of data sets. We report experimental results on two synthetic data sets.

DATABASE USED

In this work will use database of “Doctor Patient Evaluation”, which will found at machine learning repository.

VII. FUTURE WORKS

As a future work a new algorithm can also be developed to regenerate the hidden rules, if we want to send all rules to authenticated user but not to unauthenticated user. The authentic user can apply the regeneration algorithm to reconvert rules. As a future work a new algorithm can also be developed to regenerate the hidden rules. For example if we want to send all rules to authenticated user but not to unauthenticated user. The authentic user can apply the regeneration algorithm to reconvert rules.

REFERENCES

[1] Alexandre Evfimievski and Tyrone Grandison, “Privacy Preserving Data Mining” at IBM Almaden Research Center, 2007.

[2] Tzung-Pei Hong, Dept. of Comput. Sci. & Inf. Eng., Nat. Univ. of Kaohsiung, Kaohsiung, Taiwan, ”Evolutionary privacy-preserving data mining”, 19-23 Sept. 2010.

[3] Adam, N. R. & Wortmann, J. C., “Security-Control Methods for Statistical Databases: A Comparative Study”, ACM Computing Surveys, Vol. 21, N. 4, pp. 515–556, 1989.

[4] ‘Shyue-Lia Wang; Yu-Huei Lee; Billis, S.; Jafari, A. Systems, Man and Cybernetics “Hiding Sensitive Items In Privacy Preserving Association Rule Mining”, 2004 IEEE International Conference on Volume 4, Issue , 10-13 Page(s): 3239 – 3244, Oct. 2004.

[5] Cornelia Gyorodi Robert Gyorodii, prof. Stefan Holban, “A Comparative Study of Association Rules Mining Algorithms",. Jan 2004.

[6] Qi Wang and Krishnamoorthy Siva Kumar, USA:”On the Privacy Preservin Properties of Random Data PerturbationTechniques” Proceedings of the Third IEEE International Conference on Data Mining, pages 43-56, 2003

[7] Krishnamoorthy Siva Kumar, “SpectralFiltering Ttechnique Method” Proceedings of the Third IEEE International Conference on Data Mining, pages 40-48, 2003.

[8] Hillol Kargupta, Kun Liu, Souptik Datta, and Jessica Ryan Krishnamoorthy Siva Kumar:" Homeland Security and Privacy Sensitive Data Mining from Multi-Party distributed Resources" The IEEE International Conference on Fuzzy Systems pages 727-764, 2003

[9] Jaideep Vaidya Chris Clifton , "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", In ACM SIGMOD Workshop on Research Issues Data Mining and Knowledge Discovery, pages 639-644, June 2002

[10] LiWu Chang and Ira S. Moskowitz, Parsimonious downgrading and decisiontrees applied to the inference problem, In Proceedings of the 1998 New SecurityParadigms Workshop, pages 82–89, 1998.

[11] Dr.Kenneth Collien, Dr Carreyand Mr.Donald Sautter, "A Perspective of Data Mining Techniques",1998.

[12] Rakesh Agrawal, Tomasz Imielinski, Arun Swami.Mining "Association Rules Between Sets Of Items In Large Databases". Proc. of the ACM SIG-MOD Conference on Management of Data 216, Washington, D.C., May 1993

[13] Alexandre Evfimievski and Tyrone Grandison, "Privacy Preserving Data Mining" at IBM Almaden Research Center, 2007.

[14] Tzung-Pei Hong, Dept. of Comput. Sci. & Inf. Eng., Nat. Univ. of Kaohsiung, Kaohsiung, Taiwan, "Evolutionary privacy-preserving data mining", 19-23 Sept. 2010.

[15] Er.M.R.Arun Venkatesh, Bharath University, Chennai, "Privacy-Preserving Updates to Anonymous and Confidential Database", June-2012.

[16] Justin Zhan from Carnegie Mellon University, USA, "Privacy- Preserving Collaborative Data Mining", 2008.

[17] Benjamin C.M. Fung, Ke Wang, and Philip S. Yu, Fellow, IEEE, "Anonymizing Classification Data for Privacy Preservation", may 2007.

[18] Keke Chen, Member, IEEE, and Ling Liu, Senior Member, IEEE, "Privacy-Preserving Multiparty Collaborative Mining with Geometric Data Perturbation", Dec-2009.

[19] Agrawal, R. & Srikant, R., "Privacy Preserving Data Mining. In Proc. of ACM SIGMOD", Conference on Management of Data (SIGMOD'00), Dallas, TX, 2000.

