American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Investigations in Privacy Preserving Data Mining

Kshitij Pathak[a]*, Sanjay Silakari[b], Narendra S. Chaudhari[c]

[a]*Department of CSE, University Institute of Technology, RGPV Bhopal (M.P.), India*
[b]*Professor, Department of CSE, University Institute of Technology, RGPV Bhopal (M.P.), India*
[c]*Dean, Research and Development, IIT,Indore, India*
[a]*Email: er.k.pathak@gmail.com*
[b]*Email: ssilakari@yahoo.com*
[c]*Email: nsc183@gmail.com*

Abstract

Data Mining, Data Sharing and Privacy-Preserving are fast emerging as a field of the high level of the research study. A close review of the research based on Privacy Preserving Data Mining revealed the twin fold problems, first is the protection of private data (Data Hiding in Database) and second is the protection of sensitive rules (Knowledge) ingrained in data (Knowledge Hiding in the database). The first problem has its impetus on how to obtain accurate results even when private data is concealed. The second issue focuses on how to protect sensitive association rule contained in the database from being discovered, while non-sensitive association rules can still be mined with traditional data mining projects. Undoubtedly, performance is a major concern with knowledge hiding techniques. This paper focuses on the description of approaches for Knowledge Hiding in the database as well as discuss issues and challenges about the development of an integrated solution for Data Hiding in Database and Knowledge Hiding in Database. This study also highlights directions for the future studies so that suggestive pragmatic measures can be incorporated in ongoing research process on hiding sensitive association rules.

Keywords: Privacy Preserving Data Mining; Association Rule Hiding; Data Hiding in Database; Knowledge Hiding in Database.

1. Introduction

Privacy-Preserving data mining still important in the field of data mining because of the increased risk of disclosure of private information from the large datasets.

---

* Corresponding author.

Various approaches are available in the literature for preserving the privacy in the database before its release. The various area that can be view forward in the field of privacy-preserving data mining falls into 2 broad categories viz.

1) Modifying the data for protecting privacy

2) Modifying for hiding sensitive knowledge mined from various data mining applications. The first type includes the techniques like randomization, K-anonymity and l-diversity. The second type includes the techniques in which if any data mining application like classification or association rule mining is applied to the database, sensitive information which is indirectly generated through mining techniques should not be revealed. In both the categories major issues are dealing with a high dimensional dataset or large dataset in which performance of the approaches is main concerns. This paper presents various techniques that fall into these categories, discussions on their performance on the basis of accuracy as well as time complexity and enhancements that can be made in future.

2. Data hiding in database

2.1. K-Anonymity & L-Diversity, T-Closeness

References [1-3] presents the concept of K-anonymity in detail. In K-anonymity, changes are applied to the database in such a way that individual record cannot be identified directly or indirectly. In databases, the individual record can be treated as an entity which can be determined by an attribute or group of attributes. In K-anonymity changes are made to the database in such a way that combination of attributes cannot be used to identify a particular entity, it matches with at least K-entities so individual information cannot be disclosed. The techniques available for K-anonymity are generalization and suppression. In Generalization, the value of the attribute or set of an attribute for an entity is replaced by its generalized version to remove the threat of entity identification from the public database. As an example, Age of a particular person in an employee dataset is replaced by its generalized value, let say, {23-45}. This generalization of the attribute value is made in such a way that map particular values of attribute map to at least k entities. In suppression, the value of an attribute is removed to protect sensitive information either by replacing it with NULL or let say by replacing with “Not Applicable”. If suppression is applied, less attribute value will be substituted with generalized value. The techniques are treated as optimal which generates minimum k-anonymous table but finding the minimal K-anonymous problem is proved as NP-Hard [4-6].

The various challenges still left are a requirement for applications that can detect violations of K-anonymity and at the same time if the violation is detected then that can be successfully eliminated from the dataset before its release. Since this article also focuses on a combination of data hiding and knowledge hiding in the database, new techniques can be developed which while performing knowledge hiding, by defaults prepare an anonymous table to achieve k-anonymity. K-anonymity proved to effective in maintaining the privacy of individual record identification, but it does not work well when there are sensitive values exist in a group of K which is formed while anonymizing the data. So L-diversity is introduced which protect the sensitive value of an attribute by applying intra-group diversity. Various methods have been proposed in [7, 8] but these methods suffer from the

curse of dimensionality [9].

The concept of L-diversity is enhanced by T-closeness model [10]. L-diversity does not take into consideration the distribution of data. It is important because in real data attribute values are skewed. An attacker can make use of history to make assumptions of sensitive values in data, For example, an attribute corresponding to the presence of a criminal case of an employee may be sensitive if the value is positive rather than when the value is negative. In [10], a T-closeness model was proposed which call for that the "distance" among the distribution of a sensitive attribute in the original and generalized tables be at most T.

2.2. Randomization

It is a prevalent method of preserving privacy in the database. In this method, noise is either added or multiplied to records to mask the value of records. Initial work can be found in [11, 12]. Data can be reconstructed by removing noise. This is discussed in [13].

Randomization is explained by an example below: Consider a set of data values {A1, A2, A3 ...An} then it is distorted by additive strategy by adding noise generated from probability distribution {N1, N2....Nn} to produce output as {A1 + N1, A2 + N2 .....An+ Nn}. The variance of the noise added is taken large so no prediction or guessing of original values can be done. In multiplicative strategy, noise is multiplied by data values. Randomization can be extended to various data mining tasks such as classification as done in [13] and association rule mining as done in [14].

3. Association rule hiding techniques- Knowledge hiding in database

This section covers various association rule hiding techniques which fall under the research area of knowledge hiding in the database. Association rule mining is prevalent techniques which mine the interesting correlated information between the database attributes. For example, let there be market basket database which contains the list of items purchased in various transactions. One of association rule could be Milk → Butter which implies that in most of the transactions where Milk is purchased, Butter is also purchased. Data Sharing brings many advantages to corporations for sharing their data for analysis, but at the same time, they want their sensitive data should be hidden.

In association rule hiding, owners first mine the association rules by selecting two parameters namely minimum support threshold and minimum confidence threshold. After mining of association rules, owners select certain rules which have been identified as sensitive and the owner does not want them to be disclosed, so various techniques of association rule hiding will be applied to hide such sensitive, confidential information.

There are three main classes of association rule hiding viz. Border Based approaches, Heuristic approaches and exact approaches. Heuristic approaches are the fast algorithms which find quick solutions by applying certain heuristics, but they suffer from side-effects because heuristic algorithms take local decisions which are sometimes not appropriate at the global level. Border Based approaches hide sensitive association rules by modifying the border in itemset lattice formed by frequent itemset and nonfrequent itemset of the original

database. Exact Hiding approaches hide sensitive association rule by converting the problem into optimization problem or constraint satisfaction problem which can be solved by integer programming. Association rule hiding method can be characterized by hiding strategy adopted. There can be two hiding strategies viz. support based and confidence based. Support Based strategies hide sensitive association rules by deleting item in the database belongs to sensitive association rule for reducing its support below minimum support threshold. Confidence based strategies hide sensitive association rules by modifying the database to reduce the confidence below minimum confidence threshold. Another classification is based upon the types of data modification technique applied. They are data distortion based approach and data blocking based approach. Data- Distortion relies on data transformation, and exactly , the procedure is to change a selected set of 1-values to 0- values (deleting items) or 0-values to 1- values (adding items) if we consider the database as a matrix of two dimensions. The main aim of making such modifications is to reduce the support or confidence of the sensitive rules below the user pre-defined security threshold. Early data distortion techniques adopt heuristic-based sanitization strategies as Algo1a/ Algo1b/ Algo2a, Algo2b/ Algo2c[15], Naive/ MinFIA/ MaxFIA/ IGA[16], SWA[17]. Data-Blocking[18] is another popular data modification approach for association rule hiding. Instead of making data distorted, i.e., Making changes in presence or absence of item, blocking approach is executed by replacing certain data items with a question mark "?". The introduction of this special unknown value brings uncertainty to the data, making the support and confidence of association rule become too uncertain intervals respectively. The various approaches that fall under data sanitization are Nulling Out, Masking Data, Substitution, Shuffling Records, Number Variance, Gibberish Generation, Encryption/Decryption. Another categorization is based upon a number of rules considered in each iteration of hiding algorithms. They are a single rule or multiple rule strategies. In Single rule strategy, in each pass of association rule hiding algorithm only one rule is selected for hiding and database is modified to hide this sensitive association rule. In multiple rule strategies, more than one rule is considered in each iteration to be hided. Zhang [19] hides sensitive association rules by adding and removing transactions from the database. In general majority of techniques modify transaction of the database, but a number of transactions remain constant. In [19] novel approach is presented which add or remove transactions from the database to hide sensitive knowledge. Another new research area in the field of association rule hiding is data reconstruction based approach[20, 21]. In this method, hiding is performed with the help of three-phased approach. In the first phase, Knowledge is extracted from the database with the help of association rule mining algorithm. After first phase user identified generated association rules into two sets, i.e., sensitive and non-sensitive. In the second phase, knowledge sanitization is performed to hide sensitive association rule. In third phase database is reconstructed from modified knowledge. The main advantage of the reconstruction based approach is user can control hiding effects directly. Another important issue is time complexity of association rule hiding algorithms that need to be considered while designing an approach [23]. The problem of optimal Sanitization in hiding association rules is NP-Complete [24]. Reference [22] proposes a new, exact border-based approach. This method provides an optimal solution for hiding sensitive frequent itemsets. Sensitive frequent itemsets are hidden by extending the original database. In this paper database extension problem is also considered as a constraint satisfaction problem and further mapped to an equivalent binary integer programming problem. In extending the database, those transactions are picked which are underutilized and non sensitive itemsets are added to the transactions. Sometimes it is not possible to generate the optimal solution, so approach minimally relaxes constraint satisfaction problem to

provide an approximate accurate. Experimental results given in the paper highlights that approach for extending the database for hiding sensitive items provide useful solutions. Reference [25]presents the main algorithm based on border based approach. In their work, they have taken relative frequency of non-sensitive frequent itemset into consideration in combination with the frequency of frequent itemsets. Approach hides sensitive frequent itemset by modifying the transaction in the database and greedy approach is used for deciding for modifications at the local level. In [26, 27] sun proposed BBA Algorithm which is a heuristic approach based on the notion of the border of non-sensitive frequent itemsets. In this method, weights are assigned to itemsets.

[28,29] extended border based on Max-Min criteria. These strategies modified the positive border of frequent itemsets for evaluating the impact of modification. This method has better experimental result in comparison to BBA Algorithm [26, 27].

Table 1: Algorithms with their Categories for Association Rule Hiding

S No	Algorithm	Category
1	Main[25]	Border Based Approaches
2	Algo 1a [15]	Heuristic Based, Data Distortion Based
3	Algo 1b [15]	Heuristic Based, Data Distortion Based
4	Algo 1c [15]	Heuristic Based, Data Distortion Based
5	Algo 2a [15]	Heuristic Based, Data Distortion Based
6	Algo 2b [15]	Heuristic Based, Data Distortion Based
7	Naive [16]	Heuristic Based, Data Distortion Based
8	MinFIA [16]	Heuristic Based, Data Distortion Based
9	MaxFIA [16]	Heuristic Based, Data Distortion Based
10	IGA [16]	Heuristic Based, Data Distortion Based
11	SWA [17]	Heuristic Based, Data Distortion Based
12	GIH [18]	Heuristic Based, Data Blocking Based
13	CR [18]	Heuristic Based, Data Blocking Based
14	CR-2 [18]	Heuristic Based, Data Blocking Based
15	WAT-Adding [19]	Heuristic Based Approaches
16	SAT-Removing [19]	Heuristic Based Approaches
17	TAR [19]	Heuristic Based Approaches
18	BBA [26,27]	Border-Based Approaches
19	Max-Min [28,29]	Border Based Approaches
20	Menon Algorithm [30]	Exact and Heuristic Based
21	Inline Algorithm	Non-Heuristic
22	Two-Phase Iterative Algorithm	Exact Hiding

Reference [30] was the first one to consider the association rule hiding problem into two parts viz. exact part

and heuristic part. Author formulated constraint satisfaction problem for the exact part and for the heuristic part they designed an intelligent sanitization approach to improve the performance. Reference [31]proposed a non-heuristic approach in which problem is modified as constraint satisfaction problem similar to [30], but here binary integer programming (BIP) is used for finding a solution. Reference [32] extends work of [31] by forming a two-phase iterative algorithm. The summary of all the algorithms of this field, reviewed by us is presented in a tabular form, see Table 1.

4. Research Directions in the Field of Privacy Preserving Data Mining

In privacy preserving data sharing perspective, there is a need to hide sensitive data as well as sensitive knowledge (i.e., Sensitive association rule). [20]Hiding sensitive data is referred as Data Hiding in Database (DHD) whereas hiding sensitive association rule is referred as Knowledge Hiding in Database (KHD). Data Hiding in Database (DHD) and Knowledge Hiding in Database (KHD) techniques are always investigated separately. An approach is still required to integrate both DHD and KHD techniques. Development of such an approach is significant and provides confidence to data owners for data sharing. This work proposes to investigate the applications and improvements in the field of privacy preserving data mining. Performance can be enhanced while performing privacy preservation during data mining process. The following proposals can be taken forward during the tenure of the research work:

(1) Development of an integrated approach for protecting private data as well as to hide sensitive association rules.

(2) A sensitive rule can be hided either by decreasing confidence or support of the rule. This has to be achieved by making minimum changes in the database as well as with limited or no side-effects.

(3) Develop an approach using unknowns for knowledge hiding in the database. Evidence has shown that the use of unknowns in several real-life scenarios is much more preferable than the use of conventional distortion techniques. This is true because distortion techniques fail to provide a distinction between the real values in the dataset and the ones that were distorted by the hiding algorithm in order to allow for its proper sanitization.

(4) The time complexity of data mining process may be reduced by integrated DHD and KHD approach.

(5) Performance can be improved by using a cluster of computing machines.

(6) The formalism of an approach to protect private data and sensitive knowledge which gives better performance and more secured results and provide data owner a trust to release database for data sharing. Proposed to develop a formalism to cross verify the experimental result for a different approach.

(7) A hybrid approach can be formed by combining the concept of DSR (Decrease support of RHS) and ISL (Increase support of LHS) to hide rules.

(8) Association rule hiding methodologies modifies the original database in a way that among the three goals given below, at least one of the following goals must be achieved: All the rules that are considered as sensitive from user's perspective must be successfully hidden, i.e., No sensitive rule can be mined. All the association rules which are non-sensitive from user's perspective can be still mined from the modified database. No ghost rule must be generated from the modified database. (Ghost rule are the rules which are derived from modified database but was not derived from original database).

5. Conclusion

This paper discusses various approaches available for privacy preserving data mining. For data hiding in the database, various approaches like randomization, K-anonymity, L-diversity etc. are discussed. For knowledge hiding in database different approaches like heuristic based, border based etc are discussed. The limitations of the privacy-preserving data mining technique is none of the algorithms outperform all others on all the measures. The performance of the privacy-preserving data mining techniques is measured regarding performance, the effectiveness of the algorithm, balancing privacy and utility etc. There exists algorithm better in one specific criterion but does not out outperform other algorithms on particular measures. It is also well known that getting an optimal sanitization in association rule hiding is a NP-complete problem. This paper also highlights the various research directions that can be taken further by researchers to provide an efficient solution for database security administrators to preserve privacy in large databases.

References

[1] Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. In Proc. of the 31th VLDB Conference, Trondheim, Norway, September 2005.

[2] Pierangela Samarati. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027, November 2001.

[3] Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In Proc. of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, page 188, Seattle, WA, 1998.

[4] Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Anonymizing tables. In Proc. of the 10th International Conference on Database Theory (ICDT'05), Edinburgh, Scotland, January 2005.

[5] Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Approximation algorithms for k-anonymity. Journal of Privacy Technology, November 2005.

[6] Adam Meyerson and Ryan Williams On the complexity of optimal k- anonymity. In Proc. of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Paris, France, June

2004.

[7] Machanavajjhala A., Gehrke J., Kifer D., and Venkitasubramaniam M.: l-Diversity: Privacy Beyond k-Anonymity. ICDE, 2006.

[8] Xiao X., Tao Y. Anatomy: Simple and Effective Privacy Preservation. VLDB Conference, pp. 139-150, 2006.

[9] Aggarwal C. C. On k-anonymity and the curse of dimensionality. VLDB Conference, 2005.

[10] Li N., Li T., Venkatasubramanian S: t-Closeness: Privacy beyond k-anonymity and l-diversity. ICDE Conference, 2007.

[11] Warner S. L. Randomized Response: A survey technique for eliminating evasive answer bias. Journal of American Statistical Association, 60(309):63–69, March 1965.

[12] Liew C. K., Choi U. J., Liew C. J. A data distortion by probability distribution. ACM TODS, 10(3):395–411, 1985.

[13] Agrawal R., Srikant R. Privacy-Preserving Data Mining. Proceedings of the ACM SIGMOD Conference, 2000.

[14] Evfimievski A., Srikant R., Agrawal R., Gehrke J.: Privacy-Preserving Mining of Association Rules. ACM KDD Conference, 2002.

[15] V.S. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, ―Association rule hiding,IEEE Trans. Knowledge and Data Engineering, vol. KDE-16, no. 4, pp. 434-447, 2004.

[16] S.R.M. Oliveira and O.R. Zaïane, ―Privacy preserving frequent itemset mining,in Proc. 2 nd IEEE-ICDM Workshop on Privacy, Security and Data Mining, Australian Computer Society, 2002, pp. 43-54.

[17] Oliveira, S.R.M. and Zaïane, O.R. Protecting sensitive knowledge by data sanitization. In: Proc. of the 3Prd IEEE Int'l Conf. on Data Mining (ICDM'03). IEEE Computer Society, USA, 2003. 613-616.

[18] Saygin, Y., Verykios, V.S., and Clifton, C. Using unknowns to prevent discovery of association rules. SIGMOD Record, 2001, 30(4):45-54

[19] Zhang, Xiaoming. Knowledge Hiding in Data Mining by Transaction Adding and Removing. In Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International, vol. 1, pp. 233-240. IEEE, 2007.

[20] Guo, Yuhong. Reconstruction-based association rule hiding. Proceedings of SIGMOD2007 Ph. D.

Workshop on Innovative Database Research. Vol. 2007. 2007.

[21] Chen, X., Orlowska, M., and Li, X. A new framework for privacy preserving data sharing. In: Proc. of the 4th IEEE ICDM Workshop: Privacy and Security Aspects of Data Mining. IEEE Computer Society, 2004. 47-56.

[22] Gkoulalas-Divanis, A.; Verykios, V.S. "Exact Knowledge Hiding through Database Extension", Knowledge and Data Engineering, IEEE Transactions on, On page(s): 699 - 713 Volume: 21, Issue: 5, May 2009

[23] Pathak, K.; Chaudhari, N.S.; Tiwari, A., "Privacy preserving association rule mining by introducing concept of impact factor," Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference on , vol., no., pp.1458,1461, 18-20 July 2012 doi: 10.1109/ICIEA.2012.6360953

[24] Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., and Verykios, V.S. Disclosure limitation of sensitive rules. In: Scheuermann P, ed. Proc. of the IEEE Knowledge and Data Exchange Workshop (KDEX'99). ,1999. PP. 45-52.

[25] Sun, X., & Philip, S. Y. (2007). Hiding Sensitive Frequent Itemsets by a Border-Based Approach. JCSE, 1(1), 74-94.

[26] X. Sun and P. S. Yu. A border–based approach for hiding sensitive frequent itemsets. In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM), pages 426–433, 2005.

[27] X. Sun and P. S. Yu., Hiding sensitive frequent itemsets by a border–based approach. Computing science and engineering, 1(1):74–94, 2007.

[28] G. V. Moustakides and V. S. Verykios. A max–min approach for hiding frequent itemsets. In Workshops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), pages 502–506, 2006.

[29] G. V. Moustakides and V. S. Verykios. A maxmin approach for hiding frequent itemsets. Data and Knowledge Engineering, 65(1):75–89, 2008.

[30] S. Menon, S. Sarkar, and S. Mukherjee. Maximizing accuracy of shared databases when concealing sensitive patterns. Information Systems Research, 16(3):256–270, 2005.

[31] A. Gkoulalas-Divanis and V. S. Verykios. An integer programming approach for frequent itemset hiding. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM), pages 748–757, 2006.

[32] A. Gkoulalas-Divanis and V. S. Verykios. Hiding sensitive knowledge without side effects. Knowledge and Information Systems, 20(3):263–299, 2009.