An Analysis on Preservation of Privacy in Data Mining

Madhan Subramaniam[1], Senthil R[2]

[1]Asst. Professor, Department of Information Technology, Periyar Maniammai University, TN 613403 INDIA
[2]Programmer Analyst, Cognizant Technology Solutions, TN 600097, INDIA

***Abstract*— Privacy has become a key issue for progress in data mining. Maintaining the privacy of data mining has become increasingly popular because it allows sharing of privacy-sensitive data for analysis. So people are still reluctant to share information, which often leads to people who either refuse to share information or give false information. In turn, such problems in data collection can affect the success of data mining based on sufficient amounts of accurate data to provide meaningful results. In recent years the widespread availability of personal data has made the problem of privacy preservation of data mining, an important one. Several methods have recently been proposed privacy preservation of data mining for multidimensional data records. The paper aims to repeat a number of privacy preservation of data mining technology clearly and then study the advantages and disadvantages of this technique.**

Keywords-privacy preserving; data mining

I. INTRODUCTION

With the development of data analysis and data processing agencies, industry and governments have yet to publish the micro-data (data that does not contain information on individuals), information extraction, the study of epidemics, or economic models. Although the published information is presented will provide valuable information for researchers, including sensitive information about people, whose privacy may be at risk [1].

II. *K*-ANONYMITY

When you enter the micro data for research purposes, we need to restrict the disclosure of risk to an acceptable level and maximize the usefulness of the information. To limit the risk of disclosure, Samarati et al. [2] Sweeney [3] provides for the confidentiality of k-anonymity, which requires that each record in the table stock does not stand out at least k-1 other records within the data set towards a set of quasi-identifier attributes. In order to meet the requirement of k-anonymity, they used two generalization and deletion of data anonymization. Unlike the traditional protection of personal information, such as the exchange of information and adding noise, information in K-anonymous Table through generalization and suppression remains truthful.

In particular, a table is *k*-anonymous if the QI values of each tuple are identical to those of at least *k*-1 other tuples. Table3 shows an example of 2-anonymous generalization for Table1. Even with the voter registration list, an adversary can only infer that Jerry may be the person involved in the first 2 tuples of Table3, or equivalently, the Marital Status of Jerry is discovered only with probability 50%. In general, *k*-anonymity guarantees that an individual can be associated with his real tuple with a probability at most 1/*k*. While *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. There are two attacks: the homogeneity attack and the background knowledge attack. Because the limitations of the *k*-anonymity model stem from the two assumptions [4]. First, it may be very hard for the owner of a database to determine which of the attributes are or are not available in external tables. The second limitation is that the *k*-anonymity model assumes a certain method of attack, while in real scenarios there is no reason why the attacker should not try other methods. Example 1. Table4 is the original data table [5], and Table5 is an anonymous version of it satisfying 2-anonymity.

The Marital status attribute is sensitive. Suppose Jay knows that Jerry is a 30-year old man working for 35 hours per week and Jerry's record is in the table. From Table5, John can conclude that Philip corresponds to the first equivalence class, and thus must have married. This is the homogeneity attack. For an example of the background knowledge attack, suppose that, by knowing Peter's age and hours he worked, Stanley can conclude that Peter corresponds to a record in the last equivalence class in Table5. Furthermore, suppose that Stanley knows that Peter has very low risk for married. This background knowledge enables Stanley to conclude that Peter most likely is single.

Table 1. Microdata

	Age	Marital status	Sex	Hours
1	30	Divorced	M	35
2	35	Divorced	M	40
3	27	Divorced	F	35
4	40	Divorced	M	35
5	35	Divorced	F	50
6	30	Divorced	M	40

Table 2. Voter Registration List

	Name	Age	Sex	Hours
1	Jerry	30	M	35
2	Stanley	35	M	40
3	Clara	27	F	35
4	John	40	M	35
5	Philip	35	F	50
6	Peter	30	M	40

Table 3. A 2-Anonymous Table

	Age	Sex	Hours	Marital status
1	3*	M	3*	Divorced
2	3*	M	3*	Divorced
3	3*	M	3*	Divorced
4	4*	*	4*	Married
5	3*	*	5*	Married
6	3*	*	4*	Single

Table 4. Original Data Table

	Marital status	Sex	Hours
1	Divorced	M	35
2	Divorced	M	40
3	Divorced	F	35
4	Married	M	35
5	Married	F	50
6	Single	M	40

Table 5. A 2-Anonymous Version of Table

	Marital status	Sex	Hours
1	Divorced	M	3*
2	Divorced	M	4*
3	Divorced	*	3*
4	Married	*	3*
5	Married	*	5*
6	Single	*	4*

III. THE PERTURBATION APPROACH

Perturbation approach works under the requirement that data server is not allowed to learn or restore the precise records. This limitation naturally leads to some challenges. Since the method does not reconstruct the original data values, but only distribution, new algorithms have been developed which uses these reconstructed distributions to carry out mining of the data available. This means that all the problems of personal data, such as classification, clustering, association rules, or mining, a new division algorithm is based on data mining needs to be developed. For example, Agrawal [6] develops a new algorithm which is based on the distribution of data mining classification problem, while the technical Vaidya and Clifton [7] and Rizvi and Haritsis develop methods to preserve the privacy of association rules mining. Although some sense has been developed in the mining-based dissemination of information to specific problems, such as association rules and classifications, it is clear that using distributions instead of the original records restricts different algorithmic techniques that can be used on the data [8].

cn is based on the works under an implicit assumption of treating each dimension independently. In many cases, large amounts of information relevant to data mining algorithms such as classification is hidden in the inter-attribute correlations. For example, the classification technique uses a similar technique of single-attribute split algorithm. However, other techniques such as multivariate decision tree algorithms cannot be accordingly modified to work with the perturbation approach. This is due to the independent treatment of the different attributes by the perturbation approach. This means that distribution based data mining algorithms have an inherent disadvantage of loss of implicit information available in multidimensional records.

IV. SUPRESSION TECHNIQUES

Privacy can be preserved by simply suppressing all sensitive data before any disclosure or computation occurs. Given a database, we can suppress specific attributes in particular records as dictated by our privacy policy. For a partial suppression, an exact attribute value can be replaced with a less informative value by rounding (e.g., $23.45 to $20.00), top coding (e.g., age above 70 is set to 70), generalization (e.g., address to zip code), using intervals (e.g., age 23 to 20-25, name Johnson to J-K), and so forth. Often the privacy guarantee trivially follows from the suppression policy [9]. However, the analysis may be difficult if the choice of alternative suppressions depends on the data being suppressed, or if there is dependency between disclosed and suppressed data. Suppression cannot be used if data mining requires full access to the sensitive values.

Rather than protecting the sensitive values of individual records, we may be interested in suppressing the identity (of a person) linked to a specific record. The process of altering the data set to limit identity linkage is called de-identification. A set of personal records is said to be k-anonymous if every record is indistinguishable from at least $k - 1$ other records over given quasi-identifier subsets of attributes. A subset of attributes is a quasi- identifier if its value combination may help link some record to other personal information available to an attacker, for example, the combination of age, sex, and address.

To achieve k-anonymity, quasi-identifier attributes are completely or partially suppressed. A particular suppression policy is chosen to maximize the utility of the k-anonymized data set. The attributes that are not among quasi-identifiers, even if sensitive (e.g., diagnosis), are not suppressed and may get linked to an identity. Utility maximization may create an

exploitable 6 single * 4* dependence between the suppressed data and the suppression policy. Finally, k-anonymity is difficult to enforce before all data are collected in one trusted place.

Suppression can also be used to protect from the discovery of certain statistical characteristics, such as sensitive association rules, while minimizing the distortion of other data mining results. Many related optimization problems are computationally intractable, but some heuristic algorithms were studied.

V. RANDOMIZED RESPONSE TECHNIQUES

We propose to use the Randomized Response techniques to solve the DTPD problem [10]. The basic idea of randomized response is to scramble the data in such a way that the central place cannot tell with probabilities better than a pre-defined threshold whether the data from a customer contain truthful information or false information. Although information from each individual user is scrambled, if the number of users is significantly large, the aggregate information of these users can be estimated with decent accuracy. Such property is useful for decision-tree classification since decision-tree classification is based on aggregate values of a data set, rather than individual data items.

Randomized Response (RR) techniques were developed in the statistics community for the purpose of protecting surveyer's privacy. We[11] briefly describe how RR techniques are used for single-attribute databases. And we propose a scheme to use RR techniques for multiple attribute databases.

Randomized Response technique was first introduced by Warner as a technique to solve the following survey problem: to estimate the percentage of people in a population that has attribute A, queries are sent to a group of people [12]. Since the attribute A is related to some confidential aspects of human life, respondents may decide not to reply at all or to reply with incorrect answers. Two models: Related-Question Model and Unrelated-Question Model have been proposed to solve this survey problem. In the Related-Question Model, instead of asking each respondent whether he/she has attribute A, the interviewer asks each respondent two related questions, the answers to which are opposite to each other[13].

VI. THE CONDENSATION APPROACH

We introduce a condensation approach [10], which constructs constrained clusters in the data set, and then generates pseudo-data from the statistics of these clusters [14]. We refer to the technique as condensation because of its approach of using condensed statistics of the clusters in order to generate pseudo-data. The constraints on the clusters are defined in terms of the sizes of the clusters which are chosen in a way so as to preserve *k*-anonymity. This method has a number of advantages over the perturbation model in terms of preserving privacy in an effective way. In addition, since the approach works with pseudo-data rather than with modifications of original data, this helps in better preservation of privacy than techniques which simply use modifications of the original data. Furthermore, the use of pseudo-data no longer necessitates the redesign of data mining algorithms, since they have the same format as the original data [15]. In contrast, when the data is constructed with the use of generalizations or suppressions, we need to redesign data mining algorithms to work effectively with incomplete or partially certain data. It can also be effectively used in situations with dynamic data updates such as the data stream problem.

We discuss a condensation approach for data mining. This approach uses a methodology which condenses the data into multiple groups of predefined size [16]. For each group, certain statistics are maintained. Each group has a size at least k, which is referred to as the level of that privacy preserving approach. The greater the level, the greater the amount of privacy. At the same time, a greater amount of information is lost because of the condensation of a larger number of records into a single statistical group entity. We use the statistics from each group in order to generate the corresponding pseudo-data.

VII. CONCLUSION

The increased ability to monitor and collect large amounts of data using the current hardware technology has led to an interest in developing data mining algorithms which preserve the privacy of users. With the development of data analysis and processing technique, the problem of privacy disclosure regarding person or company is inevitably exposed when releasing or sharing data to mine useful decision information and knowledge, then give the birth to the field of research on privacy preserving data mining. A number of methods have recently been proposed to preserve privacy in data mining of multidimensional data records. This paper intends to reiterate several privacy preserving data mining technologies clearly and then proceeds to analyze the advantages and disadvantages of these technologies.

REFERENCES

[1] P. Samarati, "*Protecting respondent's privacy in micro data release*", In IEEE Transaction on Knowledge and Data Engineering, 2001, pp.1010-1027.

[2] P. Samarati and L. Sweeney, "*Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*", In Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.

[3] L. Sweeney, "*k-anonymity: a model for protecting privacy*", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, 2002, pp. 557-570.

[4] WONG R C, LI J, FU A W, et a1, *"(a, k)-Anonymity and enhanced k-anonymity model for privacy-preserving data publishing* " , Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, New York, 2006,pp. 754-759.

[5] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "*K-Anonymous Data Mining: A Survey*", In Springer US, Advances in Database Systems (2008), pp. 4018-4052.

[6] Agrawal, R. and Srikant, R, "*Privacy-preserving data mining*", In Proc. SIGMOD00, 2000, pp.439-450.
[7] Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J, "*Privacy preserving mining of association rules*", In Proc. KDD02, 2002, pp.217-228.
[8] Hong, J.I. and J.A. Landay, "*An Architecture for Privacy Sensitive Ubiquitous Computing*", In Mobisys04, pp. 177-189.
[9] Lei Zhang, James Bailey, Arun S. Konagurthu and Kotagiri Ramamohanarao. "*A Fast Indexing Approach for Protein Structure Comparison*", *BMC Bioinformatics*, 11(Suppl 1):S46, 2010.
[10] Jian Wang Yongcheng Luo Yan Zhao Jiajin Le, "*A Survey on Privacy Preserving Data Mining*", In IWDTA'09, pp. 1-4.
[11] M. Kantarcioglu and C. Clifton, "*Privacy-preserving distributed mining of association rules on horizontally partitioned data*", In Proc.of DKMD'02, June 2002.
[12] H. Polat and W. Du, "*SVD-based collaborative filtering with privacy*", In The 20th ACM Symposium on Applied Computing, Track on Ecommerce Technologies, Santa Fe, New Mexico, 2005, pp. 13–17.
[13] S. Meregu and J. Ghosh, "*Privacy-preserving distributed clustering using generative models*", In Proceedings of the third IEEE International Conference on Data Mining (ICDM'03), Melbourne, Florida, 2003, pp. 211–218.
[14] Charu C. Aggarwal and Philip S. Yu, "*A condensation approach to privacy preserving data mining*", In EDBT, 2004, pp. 183–199.
[15] Ke Wang, Philip S. Yu, and Sourav Chakraborty, "*Bottom-up generalization: A data mining solution to privacy protection*", In ICDM, 2004, pp. 249–256.
[16] H. Yu, X. Jiang, and J. Vaidya, "*Privacy- preserving svm using nonlinear kernels on horizontally partitioned data*", In SAC '06:Proceedings of the 2006 ACM symposium on Applied computing,New York, USA, 2006, pp. 603–610.

AUTHOR'S PROFILE

Madhan Subramaniam received his B.E. degree in Computer Science and Engineering from A.A.M. Engineering College, Thanjavur and Master's degree in Bio-informatics from SASTRA, Thanjavur. He is working as Asst. Professor in Periyar Maniyammai University in Thanjavur, Tamil Nadu, India. His field of interest is Data Mining and Bio-informatics.

Mr. Senthil R received his B.E. degree in Computer Science and Engineering from Arunai Engineering College, Thiruvannamalai and Master's degree in Computer Science and Engineering from Motilal Nehru National Institute of Technology, Allahabad. He is working as Programmer Analyst in Cognizant Technology Solutions, Chennai, Tamil Nadu, India. His field of interest is Data Mining and Bio-informatics.