An assessment of Identity Security in Data Mining

Kirubhakar Gurusamy, Venkatesh Chakrapani

Abstract-Privacy preserving becomes an important issue in the development progress of data mining techniques. Privacy preserving data mining has become increasingly popular because it allows sharing of privacy-sensitive data for analysis purposes. So people have become increasingly unwilling to share their data. This frequently results in individuals either refusing to share their data or providing incorrect data. In turn, such problems in data collection can affect the success of data mining, which relies on sufficient amounts of accurate data in order to produce meaningful results. In recent years, the wide availability of personal data has made the problem of privacy preserving data mining an important one. A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records. This paper intends to reiterate several privacy preserving data mining technologies clearly and then proceeds to analyze the merits and shortcomings of these technologies..

Index Terms— privacy preserving; data mining.

I. INTRODUCTION

With the development of data analysis and processing technique, organizations, industries and governments are increasingly publishing microdata (i.e., data that contain unaggregated information about individuals) for data mining purposes, studying disease outbreaks or economic patterns. While the released datasets provide valuable information to researchers, they also contain sensitive information about individuals whose privacy may be at risk [1].

For example, a hospital may release patients' diagnosis records so that researchers can study the characteristics of various diseases. The raw data, also called microdata, contains the identities (e.g. names) of individuals, which are not released to protect their privacy. However, there may exist other attributes that can be used, in combination with an external database, to recover the personal identities. Now we assume that the hospital publishes the data in Table1, which does not explicitly indicate the names of patients. However, if an adversary has access to the voter registration list in Table2, he can easily discover the identities of all patients by joining the two tables on {Age, Sex, Zipcode}. These three attributes are, therefore, the quasi-identifier (QI) attributes.

The problem of privacy-preserving data mining [2] has found considerable attention in recent years because of recent concerns on the privacy of underlying data.

Various privacy preservation techniques fall under

- K-Anonymity
- The Perturbation approach
- Cryptographic techniques
- Randomized Response techniques
- The Condensation approach

Manuscript Received on June, 2013.

Kirubhakar Gurusamy, Research Scholar, Surya Engineering College, Erode,

Venkatesh Chakrapani, Dean, Faculty of Engineering, Erode Builder Educational Trust's Group of Institutions, Kangayam.

Many recent papers on privacy have focused on the perturbation model and its variants. Methods for inference attacks in the context of the perturbation model have been discussed in [3].

A number of papers have also appeared on the k-anonymity model recently. Other related works discuss the method of top-down specialization for privacy preservation, and workload-aware methods for anonymization [4]. A related topic is that of privacy-preserving data mining in vertically or horizontally partitioned data [5]. In this case, we determine aggregate characteristics of the data which are distributed across multiple sites without exchanging explicit information about individual records. The key in many of these approaches is to reduce the communication costs as much as possible while retaining privacy. Chawla et al. [6] discuss transformation based methods to preserve the anonymity of the data. This is different from our technique which uses group-based pseudo-data generation in order to preserve anonymity.

Microdata

ID	Attributes			
	Age	**Sex**	**Zip code**	**Disease**
1	26	M	83661	Headache
2	24	M	83634	Headache
3	31	M	83967	Toothache
4	39	F	83949	Cough

Table II. Voter Registration List

ID	Attributes			
	Name	**Age**	**Sex**	**Zip code**
1	Jim	26	M	83661
2	Jay	24	M	83634
3	Tom	31	M	83967
4	Lily	39	F	83949

Table III. A 2-anonymous Table

ID	Attributes			
	Age	**Sex**	**Zip code**	**Disease**
1	2*	M	836**	Headache
2	2*	M	836**	Headache
3	3*	*	839**	Toothache
4	3*	*	839**	Cough

II. K-ANONYMITY

When releasing microdata for research purposes, one needs to limit disclosure risks to an acceptable level while maximizing data utility. To limit disclosure risk, Samarati et al.[7]; Sweeney [8] introduced the k-anonymity privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least k-1 other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the

k-anonymity requirement, they used both generalization and suppression for data anonymization. Unlike traditional privacy protection techniques such as data swapping and adding noise, information in a 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 Jim may be the person involved in the first 2 tuples of Table3, or equivalently, the real disease of Jim is discovered only with probability 50%. In general, kanonymity 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 [9]. 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, and Table5 is an anonymous version of it satisfying 2-anonymity. The Disease attribute is sensitive. Suppose Jay knows that Tom is a 27-year old man living in ZIP 83634 and Tom's record is in the table. From Table5, Jay can conclude that Tom corresponds to the first equivalence class, and thus must have headache. This is the homogeneity attack. For an example of the background knowledge attack, suppose that, by knowing Lucy's age and zip code, Jay can conclude that Lucy corresponds to a record in the last equivalence class in Table5. Furthermore, suppose that Jay knows that Lucy has very low risk for cough. This background knowledge enables Jay to conclude that Lucy most likely has toothache.

Table IV. Original patients table

ID	*Zip code*	*Age*	*Disease*
1	83661	26	Headache
2	83634	24	Headache
3	83967	31	Toothache
4	83949	39	Cough

Table V. A 2-anonymous version of Table1

ID	*Zip code*	*Age*	*Disease*
1	836**	2*	Headache
2	836**	2*	Headache
3	839**	3*	Toothache
4	839**	3*	Cough

III. THE PERTURBATION APPROACH

The perturbation approach works under the need that the data server is not allowed to learn or recover precise records. This restriction naturally leads to some challenges. Since the method does not reconstruct the original data values but only distributions, new algorithms need to be developed which use these reconstructed distributions in order to perform mining of the underlying data. This means that for each individual data problem such as classification, clustering, or association rule mining, a new distribution based data mining algorithm needs to be developed. For example, Agrawal [10] develops a new distribution-based data mining algorithm for the classification problem, whereas the techniques in Vaidya and Clifton [11] and Rizvi and Haritsa develop methods for privacy-preserving association rule mining. While some clever approaches have been developed for distribution-based mining of data for particular problems such as association rules and classification, it is clear that using distributions instead of original records restricts the range of algorithmic techniques that can be used on the data [12].

In the perturbation approach, the distribution of each data dimension is reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. In many cases, a lot of relevant information for data mining algorithms such as classification is hidden in inter-attribute correlations. For example, the classification technique uses a distribution-based analogue 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 because of 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. CRYPTOGRAPHIC TECHNIQUES

Another branch of privacy preserving data mining which using cryptographic techniques was developed. This branch became hugely popular [13] for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it.
Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work [14] has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

V. RANDOMIZED RESPONSE TECHNIQUES

We propose to use the Randomized Response techniques to solve the DTPD problem. 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 surveyee's privacy. We [15] 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 [16]. 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 [17].

VI. THE CONDENSATION APPROACH

We introduce a condensation approach, which constructs constrained clusters in the data set, and then generates pseudo-data from the statistics of these clusters [18]. 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 [19]. 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 [20]. 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 increasing ability to track and collect large amounts of data with the use of current hardware technology has lead to an interest in the development of data mining algorithms which preserve user privacy. With the development of data analysis and processing technique, the privacy disclosure problem about individual or company is inevitably exposed when releasing or sharing data to mine useful decision information and knowledge, then give the birth to the research field on privacy preserving data mining. A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records. This paper intends to reiterate several privacy preserving data mining technologies clearly and then proceeds to analyze the merits and shortcomings of these technologies.

REFERENCES

1. P. Samarati, “Protecting respondent’s privacy in micro data release”, In IEEE Transaction on Knowledge and Data Engineering, 2001.
2. V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, “State-of-the-art in privacy preserving data mining”, In Proc of ACM SIGMOD, 2004.
3. Ackerman, M. S., Cranor, L. F., and Reagle, J, “Privacy in ecommerce: examining user scenarios and privacy preferences”, In Proc. EC99, 1999.
4. W. Du, Y. Han, and S. Chen, “Privacy-preserving multivariatestatistical analysis: Linear regression and classification”, In Proceedings of the Fourth SIAM International Conference on Data Mining, 2004.
5. K. Chen and L. Liu, “Privacy preserving data classification with rotation perturbation”, In Proceedings of the Fifth International Conference of Data Mining (ICDM’05), 2005.
6. K. Liu, H. Kargupta, and J. Ryan, “Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining”, IEEE Transactions on Knowledge and Data Engineering (TKDE), January 2006.
7. 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.
8. L. Sweeney, “*k*-anonymity: a model for protecting privacy”,International Journal on Uncertainty, Fuzziness and Knowledgebased Systems, 2002.
9. WONG R C, LI J, FU A W, eta1, “(α, k)-Anonymity : an 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.
10. Agrawal, R. and Srikant, R, “Privacy-preserving data mining”, In Proc. SIGMOD00, 2000.
11. Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J, “Privacy preserving mining of association rules”, In Proc. KDD02, 2002.
12. Hong, J.I. and J.A. Landay, “An Architecture for Privacy Sensitive Ubiquitous Computing”, In Mobisys04.
13. S. Laur, H. Lipmaa, and T. Mielik¨ainen, “Cryptographically private support vector machines”, In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.
14. Ke Wang, Benjamin C. M. Fung, and Philip S. Yu, “Template-based privacy preservation in classification problems”, In ICDM, 2005.
15. M. Kantarcioglu and C. Clifton, “Privacy-preserving distributed mining of association rules on horizontally partitioned data”, In Proc. of DKMD’02, June 2002.
16. 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.
17. 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.
18. Charu C. Aggarwal and Philip S. Yu, “A condensation approach to privacy preserving data mining”, In EDBT, 2004.
19. Ke Wang, Philip S. Yu, and Sourav Chakraborty, “Bottom-up generalization: A data mining solution to privacy protection”, In ICDM, 2004.
20. 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.