INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1315 - 1319
RECEIVED : 29/10/2016. PUBLISHED: 07/11/2016
November 7, 2016

PREVENTING INSIDER COLLUSION ATTACK IN DISTRIBUTION DATA MINING

[1]S.Annapooraneswari, [2]J.Lalithavani,
[1]PG Scholar,Dept Of Computer Science Engineering,Mailam Engineering College,Villupuram,
[2]Assistant Professor Of Computer Science Engineering,Mailam Engineering College,Villupuram.

Abstract:
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, ften resulting in individuals either refusing to disclose their data or providing wrong data. Nowadays, privacy preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the internet. We discuss method for Perturbation, K-Anonymization, condensation, and Distributed Privacy Preserving Data mining. In this paper, we have given a review of the state-of-the-art methods for privacy and analyze the representative technique for privacy preserving data mining and point out their merits and demerits. Finally the present problems and future directions are discussed.
Keywords: Cryptography, Distributed Privacy Preserving, k-Anonymity, Privacy-preserving, Perturbation.

1. INTRODUCTION

The need of privacy preserving data mining has become more significant in recent years because of the increasing ability to store personal data about users and the increasing sophistication of data mining algorithm to leverage this information. A number of methods such as kanonymity, classification, association rule mining, clustering have been recommended in recent years in order to perform privacy preserving data mining. Furthermore,the problem has been discussed in multiple communities such as the database, the statistical disclosure control(SDC) and the cryptography. Data mining techniques have been developed successfully to extracts knowledge in order to support a variety of domain areas marketing, weather forecasting, medical diagnosis, and national security. But it is still a challenge to mine some kinds of data without violating the data owners 'privacy .For example, how to mine patients 'private data is an ongoing problem in health care applications . As data mining become more pervasive, privacy concerns are increasing. In those cases, each organization or unit must ascertain that the privacy of the individual is not compromised or that sensitive business information is not divulged .Consider, for example, a government, or more specifically, one of its security branches interested in developing a system for determining, from passengers whose baggage has been checked, those who must be subjected to additional security measures. The data indicative of such need for further examination stems from a lot of sources like police records; airports; banks; general government statistics; and passenger information records that generally include personal information (such as name and passport number); demographic data (such as age and gender); flight information (such as departure, destination, and duration); and expenditure data (such as transfers, purchasing and bank transactions). In most countries, this information is considered as private and to avoid intentionally or unintentionally exposing confidential information about an individual, it is illegal to make such information freely available. Though many types of preserving individual information have been developed, there are ways for circumventing these methods. For example, in order to preserve privacy, passenger information records can be deidentified before the records are shared with anyone who is not permitted directly to access the relevant data. This can be done by removing from the dataset unique identity fields, such as name and passport number. Eventhough if this information is deleted, there are still other forms of information both personal and behavioral (e.g. date of birth, zip code, gender, number of children, number of calls, number of accounts) that, when connected with other available

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1315 - 1319
RECEIVED : 29/10/2016. PUBLISHED: 07/11/2016
November 7, 2016

datasets, could easily recognise subjects. To avoid these types of violations, we require various data mining algorithms for privacy preserving. We analyse recent work on these topics, presenting general frameworks that we use to compare and contrast different approaches.

2. ANONYMITY MODELS

K-anonymization techniques have been the focus of intense research in the last few years. In order to ensure anonymization of data while at the same time minimizing the information loss resulting from data modifications, several extending models are proposed, k-anonymity [1] is one of the most classic models, which technique that prevents joining attacks by generalizing and/or suppressing portions of the released microd ata so that no individual can be uniquely distinguished from a group of size k. In the k-anonymous tables, a data set is k-anonymous ($k \geq 1$) if each record in the data set is indistinguishable from at least ($k - 1$) other records within the same data set. The larger the value of k, the better the privacy is protected. k-anonymity can ensure that individuals cannot be uniquely identified by linking attacks. Let T (i.e. TABLE) is a relation storing private information about a set of individuals. The attributes in T are classified in four categories: an identifier (AI), a sensitive attribute (SA), quasi-identifier attributes (QI) and other unimportant attributes. The technology of l-diversity has some advantages than k-anonymity. Because k-anonymity dataset permits strong attacks due to lack of diversity in the sensitive attributes. In this model, an equivalence class is said to have l-diversity if there are at least l well-represented value for the sensitive attribute. Because there are semantic relationships among the attribute values, and different values have very different levels of sensitivity.

3. RELATED WORK

Several polls show that the public has an increased sense of privacy loss. Since data mining is often a key component of information systems, homeland security systems , and monitoring and surveillance systems , it gives a wrong impression that data mining is a technique for privacy intrusion. This lack of trust has become an obstacle to the benefit of the technology. For example, the potentially beneficial data mining research project, Terrorism Information Awareness (TIA), was terminated by the US Congress due to its controversial procedures of collecting, sharing, and analyzing the trails left by individuals .

QID	**Disease**	QID	Disease	**QID**	**Disease**	QID	Disease
q1	**HIV**	q1	HIV	**Q**	**HIV**	Q	HIV
q1	**non-sensitive**	q1	HIV	**Q**	**HIV**	Q	HIV
q2	**HIV**	q2	non-sensitive	**Q**	**non-sensitive**	Q	non-sensitive
q2	**non-sensitive**	q2	non-sensitive	**Q**	**non-sensitive**	Q	non-sensitive
q2	**non-sensitive**	q2	non-sensitive	**Q**	**non-sensitive**	q2	non-sensitive
q2	**non-sensitive**	q2	non-sensitive	**Q**	**non-sensitive**	q2	non-sensitive
q2	**non-sensitive**	q2	non-sensitive	**Q**	**non-sensitive**	q2	non-sensitive
(a)	**Good table**	(b)	Bad table	**(c)**	**Global**	(d)	Local

Fig.1. Diversity: Global and local recoding

Motivated by the privacy concerns on data mining tools, a research area called privacy-reserving data mining (PPDM) emerged in 2000 .The initial idea of PPDM was to extend traditional data mining techniques to work with the data modified to mask sensitive information. The key issues were how to modify the data and how to recover the data mining result from the modified data. The solutions were often tightly coupled with the data mining algorithms under consideration. In contrast, privacy-preserving data publishing (PPDP) may not necessarily tie to a specific data mining task, and the data mining task is sometimes unknown at the time of data publishing. Furthermore, some PPDP solutions emphasize preserving the data truthfulness at the record level, but PPDM solutions often do not preserve such property.

4. DISTRIBUTED PRIVACY PRESERVING DATA MINING

The key goal in most distributed methods for privacy-preserving data mining (PPDM) is to permit computation of useful aggregate statistics on the entire data set while not compromising the privacy of the individual data sets among the varoius participants. Thus, the participants may need to collaborate in getting aggregate results, but might not fully trust one another in terms of the distribution of their own data sets. For this reason, the data sets might either be horizontally partitioned or be vertically partitioned. In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of that has the identical set of attributes. In vertical partitioning, the individual entities might have different attributes (or views) of the identical set of records. Both kinds of partitioning pose different challenges to the problem of distributed privacy-preserving data mining. In horizontally partitioned data sets, totally different sites contain different sets of records with identical (or highly overlapping) set of attributes that are used for mining purposes. Several of those techniques use specialised versions of the general strategies discussed in for various problems. The work in discusses the development of a popular decision tree induction method called ID3with the usage of approximations of the best splitting attributes. Subsequently, a range of classifiers are generalized to the problem of horizontally partitioned privacy preserving mining including the Naïve Bayes Classifier and the SVM Classifier with nonlinear kernels . An extreme solution for the horizontally partitioned case is discussed in, within which privacy preserving classification is performed in a fully distributed setting, where every customer has personal access to only their own record. A number of other data mining applications have been generalized to the problem of horizontally partitioned data sets . These include the applications of association rule mining, clustering, and collaborative filtering. For the vertically partitioned case, several primitive operations like computing the scalar product or the secure set size intersection may be useful in computing the results of data mining algorithms. As an example, the methods in discuss the way to use scalar dot product computation for frequent item set counting. The method of counting can also be achieved by using the secure size of set intersection as discussed in . Another technique for association rule mining uses the secure scalar product over the vertical bit representation of item set inclusion in transactions, so as to calculate the frequency of the corresponding item sets. This step is applied repeatedly within the framework of a roll up procedure of item set counting.

5. ANALYSIS

It is more common that the attributes that constitute the quasi-identifier are themselves a subset of the attributes released. As a result, when a k-minimal solution, which we will call table T is released, it should be considered as joining other external information. Therefore, subsequent releases of generalizations of the same privately held information must consider all of the released attributes of T

a quasi-identifier to prohibit linking on T, unless of course, subsequent releases are themselves generalizations of T.

(a)

Disease	Sex	Age	ZIP Code	
1	F	29	47677	1
1	F	22	47602	2
1	M	27	47678	3
2	M	43	47905	4
2	F	52	47909	5
2	M	47	47906	6
3	M	30	47605	7
3	M	36	47673	8
3	M	32	47607	9

(b)

Group-ID	Disease	Count
1	Ovarian Cancer	2
1	Prostate Cancer	1
2	Flu	1
2	Heart Disease	2
3	Heart Disease	1
3	Flu	2

Fig.1. The quasi-identifier table

Methods to distribute ones and the methods for handling horizontally and vertically partitioned data. While all the proposed methods are only approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods.

CONCLUSION

This paper presents a survey for most of the common attacks techniques for anonymization-based PPDM & PPDP and explains their effects on Data Privacy. k-anonymity is used for security of respondents identity and decreases linking attack in the case of homogeneity attack a simple k-anonymity model fails and we need a concept which prevent from this attack solution is l-diversity. All tuples are arranged in well represented form and adversary will divert to l places or on l sensitive attributes. l-diversity limits in case of background knowledge attack because no one predicts knowledge level of an adversary. It is observe that using generalization and suppression we also apply these techniques on those attributes which doesn't need this extent of privacy and this leads to reduce the precision of publishing table. e-NSTAM.

REFERENCES

[1] P. Samarati and L. Sweeney, "Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression," Technical Report SRI-CSL-98-04, 1998.

[2] A. Machanavajjhala, J. Gehrke, et al., "l -Diversity: Privacy beyond k-Anonymity," Proceeding of ICDE, April 2006.

[3] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," Proceedings of ICDE, 2007, pp. 106-115.

[4] R. C. Wong, J. Li, A. W. Fu, et a1., "(α,k)-Anonymity: An Enhaned k-Anonymity Model for Privacy-Preserving Data Publishing," In: Proceedings of the 12th ACM SIGKDD, ACM Press, New York, 2006, pp. 754-759.

[5] M. Terrovitis, N. Mamoulis and Kalnis, "Privacy Preserving Anonymization of Set-Valued Data," VLDB, Auckland, 2008, pp. 115-125.

[6] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity," In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, June 2005, pp. 49-60.

[7] X. Ye, L. Jin and B. Li, "A Multi-Dimensional K-Anonymity Model for Hierarchical Data, Electronic Commerce and Security," 2008 International Symposium, Beijing, August 2008, pp. 327-332.

[8] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Workload-Aware Anonymization," Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, August 2006, pp. 277-286. doi:10.1145/1150402.1150435

[9] X. Xiao and Y. Tao, "M-Invariance: Towards PrivacyPreserving Re-Publication of Dynamic Datasets," In: Proceedings of SIGMOD, ACM Press, New York, 2007, pp. 689-700.

[10] Y. Bu, A. Wai-Chee Fu, et al., "Privacy-Preserving Serial Data Publishing By Role Composition," VLDB, Auckland, 2008, pp. 845-856.