ijiTECH International Journal of Innovative Technologies
WWW.IJITECH.ORG
ISSN 2321-8665
Volume.06, Issue.01,
January-June, 2018,
Pages:0124-0127

Privacy Preserving Techniques Modified Last

K. Rajesh[1], Shaik Rehmathunnisa Naga[2]

[1]PG Scholar, Dept of CSE, DJR Institute of Engineering & Technology, Andhrapradesh, India,
E-mail: rajeshkoppanathi73@gmail.com.

[2]Associate Professor, Dept of CSE, DJR Institute of Engineering & Technology, Andhrapradesh, India,
E-mail: shaikrehmathunnisa@gmail.com.

Abstract: Protection saving turns into a critical issue in the advancement advance of information mining methods. Security protecting information mining has turned out to be progressively well known in light of the fact that it permits sharing of protection touchy information for examination purposes. So individuals have turned out to be progressively unwilling to share their information, habitually bringing about people either declining to share their information or giving wrong information. Thusly, such issues in information accumulation can influence the accomplishment of information mining, which depends on adequate measures of precise information with a specific end goal to create significant outcomes. As of late, the wide accessibility of individual information has made the issue of protection safeguarding information mining an imperative one. Various techniques have as of late been proposed protection safeguarding information mining of multidimensional information records. This paper plans to repeat a few protection safeguarding information mining innovations obviously and afterward continues to investigate the benefits and deficiencies of these advancements.

Keywords: Protection Saving; Information Mining.

I. INTRODUCTION

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. 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. A hospital may release patients' diagnosis records so that researchers can study the characteristics of various diseases. The raw data, also called micro data, 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 has found considerable attention in recent years because of recent concerns on the privacy of underlying data. 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. 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. A related topic is that of privacy-preserving datamining in vertically or horizontally partitioned data.

II. EXISTING SYSTEM

With the development of data analysis and processing technique, organizations, industries and Governments are increasingly publishing micro data (i.e., data that contain un aggregated 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. For example, a hospital may release patients' diagnosis records so that researchers can study the characteristics of various diseases. The raw data, also called micro data, 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 inTable2, 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.

III. PROPOSED SYSTEM

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

Copyright @ 2018 IJIT. All rights reserved.

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. A related topic is that of privacy-preserving data mining in vertically or horizontally partitioned data as shown in Fig.1. 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. Chawlaectal. 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.

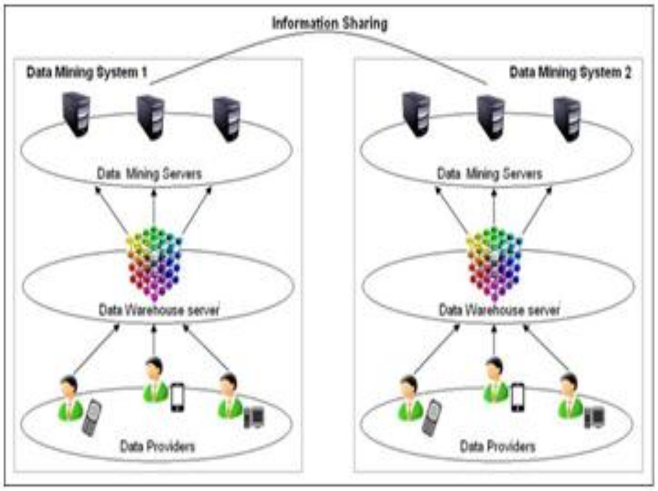


Fig.1. System Architecture.

IV. RELATED WORK

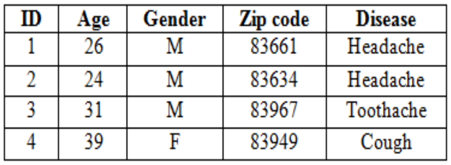
- K-Anonymity
- The Perturbation Approach
- Cryptographic Techniques
- Randomized Response Techniques
- The Condensation Approach

A. K-Anonymity

When releasing micro data for research purposes, one needs to limit disclosure risks to an acceptable level while maximizing data utility. To limit disclosure risk, Samaratietal.; Sweeney 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. shows an example of 2-anonymous generalization for. Even with the voter registration list, an adversary can only infer that Jim may be the person involved in the first 2 tuples, or equivalently, the real disease of Jim 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. Whilek-anonymity protects against identity disclosure, itdoes not provide sufficient protection against attribute disclosure. There are two attacks: the homogeneity attack and the backg round knowledge attack. Because the limitations of the k-anonymity model stem from the two assumptions. 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.

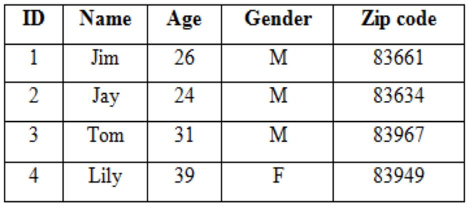
Example 1: Table4 is the original data table, and Table5is an anonymous version of it satisfying 2-anonymity. The Disease attribute is sensitive. Suppose Jay knows that Tom is a 27-year old man living in ZIP 83634 and Tom’s record is in the table. From Table5, Jay can conclude that Tom corresponds to 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 Lucycor responds to a record in the last equivalence class inTable5. 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. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work 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.

TABLE I: Micro Data



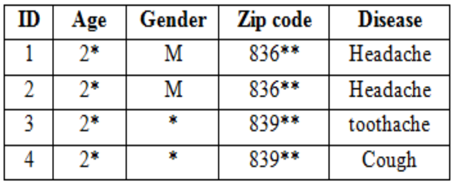
ID	Age	Gender	Zip code	Disease
1	26	M	83661	Headache
2	24	M	83634	Headache
3	31	M	83967	Toothache
4	39	F	83949	Cough

TABLE II: Voter Registration List



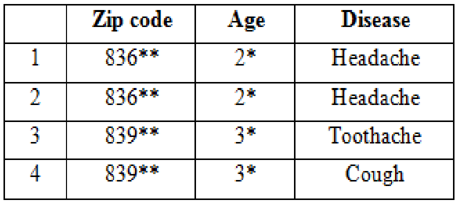
ID	Name	Age	Gender	Zip code
1	Jim	26	M	83661
2	Jay	24	M	83634
3	Tom	31	M	83967
4	Lily	39	F	83949

TABLE III: A 2-Anonymous Table



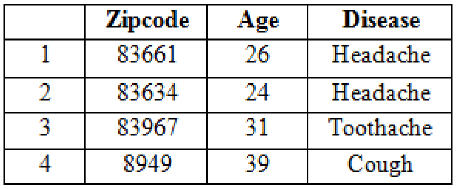
ID	Age	Gender	Zip code	Disease
1	2*	M	836**	Headache
2	2*	M	836**	Headache
3	2*	*	839**	toothache
4	2*	*	839**	Cough

TABLE IV: A 2-Anonymous Version of Table1



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

TABLE V: Original Patients Table



	Zipcode	Age	Disease
1	83661	26	Headache
2	83634	24	Headache
3	83967	31	Toothache
4	8949	39	Cough

B. 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 develops a new distribution-based data mining algorithm for the classification problem, whereas the techniques in Vaidya and Clifton and Rizviand 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 .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.

C. Cryptographic Techniques

Another branch of privacy preserving data mining which using cryptographic techniques was developed. This branch became hugely popular 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 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.

D. 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 survey's privacy. We 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. 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.

E. 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. The problem of privacy-preserving data mining has found considerable attention in recent years because of recent concerns on the privacy of underlying data. 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. 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. A related topic is that of privacy-preserving data mining in vertically or horizontally partitioned data . 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. Chawlaectal. 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 topreserve anonymity.

V. 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.

VI. REFERENCES

[1] P. Samarati, “Protecting respondent’s privacy in micro data release”,In IEEE Transaction onKnowledge and Data Engineering, 2013,pp.1010-1027.

[2] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin,and Y. Theodoridis, “State-of-the-art in privacy preserving datamining”, In Proc of ACM SIGMOD, 2014, pp. 50–57.

[3] Ackerman, M. S., Cranor, L. F., and Reagle, J, “Privacy in ecommerce:examining user cenarios and privacy preferences”, In Proc. EC99, 2014, pp. 1-8.

[4] W. Du, Y. Han, and S. Chen, “Privacy-preserving multi variatestatistical analysis: Linear regression and classification”, InProceedings of the Fourth SIAM International Conference on DataMining, 2014, pp. 222–233.113 .Authorized licensed use limited to: Gandhi Institute of Technology & Management.

[5] K. Chen and L. Liu, “Privacy preserving data classification withrotation perturbation”, InProceedings of the Fifth InternationalConference of Data Mining (ICDM’05), 2015, pp. 589592.

Author’s Profile:

K.Rajesh, completed his B.Tech in Computer Science And Engineering and pursuing M.Tech Computer Science And Engineering in DJR Institute of Engineering and Technology.

Shaik Rehmathunnisa Naga, M.Tech received her M.Tech degree and B.Tech degree in computer science and engineering. She is currently working as an Assoc Professor in DJR Institute of Engineering & Technology.