IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-9 E-ISSN: 2347-2693

On Privacy Preserving Data Mining Techniques: Merits and Demerits

Mohana Chelvan P[1*], Perumal K[2]

[1]Dept. of Computer Science, Hindustan College of Arts and Science, Chennai, India
[2]Dept. of Computer Applications, Madurai Kamaraj University, Madurai, India

**Corresponding Author: pmohanselvan@rediff.com, Tel: +91-9790646139*

Available online at: www.ijcseonline.org

Received: 07/Aug/2017, Revised: 18/Aug/2017, Accepted: 11/Sep/2017, Published: 30/Sep/2017

Abstract—Data mining is the process that extracts previously not known valid and actionable information from large archived data to make crucial business and strategic decisions. In recent years, privacy preserving data mining techniques has been studied and more research has been done in this area due to proliferation of internet in everyday life along with huge availability of personal data. Huge volume of microdata is produced on every minute due to e-governance and e-commerce which contains private data about individuals and businesses. The data has been modified in some way to preserve the privacy of individuals. The main goal of privacy preserving data mining is hiding an individual's sensitive identity and at the same time maintains the usability of data. This paper will give an overview about these rapidly changing techniques and their advancements.

Keywords—privacy-preserving data mining, k-anonymity, l-diversity, t-closeness, slicing

I. INTRODUCTION

Privacy preserving in data mining refers to the area of data miming that seeks to safeguard privacy-sensitive information from unsolicited or unsanctioned disclosure and hence protecting individual data records and their privacy. This technique provides individual privacy while at the same time allowing extraction of useful knowledge from data. There are several methods which can be used to enable privacy preserving data mining. In this paper we discuss the privacy preserving data mining techniques.

The paper is organised as follows, Section II contains randomization methods, Section III contains personalised privacy preservation, Section IV contains distributed privacy preserving data mining methods, Section V contains privacy preservation of application results, Section VI contains limitation of privacy: the curse of dimensionality, Section VII contains genomic privacy and section VIII contains conclusion.

II. RANDOMIZATION METHODS

In the randomization method, a calculated amount of noise is added to the data in order to mask the attribute values of the records. The amount of noise is so high by which the original values cannot be recovered from perturbed data. The technique is to derive aggregate distribution from perturbed data.

A. Data Swapping

The noise addition or multiplication is not the only technique which can be used to perturb the data. A related method is that of data swapping, in which the values across different records are swapped in order to perform the privacy-preservation [1]. The advantage of this technique is that the lower order marginal totals of the data are not perturbed at all and are completely preserved. Therefore certain kinds of aggregate computations can be exactly performed without violating the privacy of the data. The value of a record to be perturbed independently of the other records and is the general principle in randomization and is not followed in this technique. Therefore, this technique can be used in combination with other frameworks such as *k*-anonymity, as long as the swapping process is designed to preserve the definitions of privacy for that model.

B. Group Based Anonymization

The noise can be added to a given record is independent of the behaviour of other data records and so randomization method is a simple technique which can be easily implemented at data collection time. In this technique the outlier records can often be difficult to mask and so this is also a weakness. It is desirable to have a technique in which the level of inaccuracy depends upon the behaviour of the locality of that given record and the privacy-preservation does not need to be performed at data-collection time. It does not consider the possibility that publicly available records can be used to identify the identity of the owners of that record and is another key weakness of the randomization framework. In [2], it has been shown that the privacy getting heavily compromised in high-dimensional

cases due to the use of publicly available records. The outlier records can be easily distinguished from other records in their locality. Therefore, constructing groups of anonymous records which are transformed in a group-specific way is a broad approach to many privacy transformations.

C. The K-Anonymity Model

The k-anonymity property will be possessed by the anonymous data. In this method the published records cannot be re-identified and at the same time the data can practically useful. A release of data will have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release [3]. There will be two methods to achieve k-anonymity as follows:

- Suppression: In this method, some values are replaced with values like"*". All or some values of common are replaced by "*".
- Generalization: In this method, individual values of attributes are replaced by broader category.

This model is developed due to the possibility of indirect identification of records from public databases. It is possible by the use of combination of records attributes to identify individual record. In this method there will be reduction of the granularity of data representation by generalization and suppression. This granularity is reduced sufficiently that any given record will be mapped into at least k other records of data.

D. The L-Diversity Model

For preserving privacy in data sets, the *l*-diversity is the method of group based anonymization that is used by reducing the granularity of a data representation. There will be a trade off between the results in some loss of effectiveness of data management or mining algorithms and to gain some privacy. This model is an extension of the *k*-anonymity model. Here it reduces the granularity of data representation using techniques including generalization and suppression such that any given record maps onto at least *k* other records in the data. Especially when the sensitive values within a group exhibit homogeneity, the *l*-diversity model has the advantage by which some of the weaknesses in the *k*-anonymity model where protected identities to the level of *k*-individuals is not equivalent to protecting the corresponding sensitive values that were generalized or suppressed. The *l*-diversity model will put in the support of intra-group diversity for sensitive attributes in the anonymization method.

The *k*-anonymity is susceptible to many attacks in which availability background knowledge to an attacker will make the attacks become even more effective. The types of such attacks are as follows.

- Homogeneity Attack: In this type of attack, there will be the values for a sensitive data within a set of *k* records are identical. In such cases, the sensitive value for the set of *k* records may be exactly predicted.
- Background Knowledge Attack: One or more quasi-identifier attributes are associated with the sensitive attribute and this will reduce the set of probable values for the sensitive attribute. For example, to narrow the range of values for a sensitive attribute of a patient's disease will be used by knowing that heart attacks occur at a reduced rate in Japanese patients.

The *l*-diversity method was created to further *k*-anonymity by additionally maintaining the diversity of sensitive fields as the sensitive attributes may be inferred for *k*-anonymity data [4].

If there are at least l "well-represented" values for the sensitive attribute, then an equivalence class is said to have *l*-diversity. If every equivalence class of the table has *l*-diversity, then the table is said to have *l*-diversity.

E. The T-Closeness Model

The *t*-closeness is an enhancement of *l*-diversity group based anonymization that is used to preserve privacy in data sets by plummeting the granularity of a data representation. In this technique there will be a trade off that result in some loss of effectiveness of data management or mining algorithms in order to gain some privacy. The *t*-closeness model is the extension of the *l*-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute.

The attribute values will be semantically similar or very much skewed in the case of real world data sets; the value distributions may cause difficulty in creating feasible *l*-diverse representations. The *l*-diversity technique will be useful in that it may hinder an attacker leveraging the global distribution of an attribute's data values in order to infer information about sensitive data values. In real data sets, not every value may exhibit equal sensitivity. The *l*-diversity may be difficult and unnecessary to achieve when protecting against attribute disclosure [3]. Here sensitive information leaks may occur because it does not recognize that values may be the semantically close while *l*-diversity requirement ensures "diversity" of sensitive values in each group. For example, an attacker could deduce a stomach disease applies to an individual if a sample containing the individual only listed three different stomach diseases.

The distribution of values for *l*-diverse data results in the inference of sensitive attributes and by additionally

maintaining the distribution of sensitive fields, the *t*-closeness method was created to further *l*-diversity. Privacy beyond *k*-anonymity and *l*-diversity defines *t*-closeness.

In the t-closeness principle, an equivalence class is said to have *t*-closeness if the distance between the distribution of a sensitive attribute in the class and the distribution of the attribute in the whole table is no more than a threshold *t*. Here, a table is said to have *t*-closeness if all equivalence classes have *t*-closeness.

F. The Slicing Model

For privacy preserving microdata publishing, there will be a number of anonymization techniques including generalization and bucketization has been used. Recently it has been found that in the case of high-dimensional data, generalization loses considerable amount of information. If there have been no clear partition between quasi-identifying attributes and sensitive attributes, the bucketization method does not apply for that data and also does not thwart membership disclosure. A new technique called as slicing gives better solution for the above problems by partitioning the data both horizontally and vertically. From the research experiments it have been proved that slicing preserves better data utility than generalization and can be used for membership disclosure protection and also the slicing technique can handle high-dimensional data [5]. In the case of slicing technique, the highly correlated group of attributes are preserved for better data utility and uncorrelated group of data are sliced both horizontally and vertically to preserve privacy of the published microdata.

III. PERSONALIZED PRIVACY-PRESERVATION

Not all individuals or entities are equally concerned about their privacy. For example, a corporation may have very different constraints on the privacy of its records as compared to an individual. This leads to the natural problem that there is a need to treat the records in a given data set very differently for anonymization purposes. From a technical point of view, this means that the value of *k* for anonymization is not fixed but may vary with the record. A condensation based approach [6] has been proposed for privacy-preserving data mining in the presence of variable constraints on the privacy of the data records. This technique constructs groups of non-homogeneous size from the data, such that it is guaranteed that each record lies in a group whose size is at least equal to its anonymity level. Subsequently, pseudo-data are generated from each group so as to create a synthetic data set with the same aggregate distribution as the original data. Another interesting model of personalized anonymity is discussed in which a person can specify the level of privacy for his or her sensitive values. This technique assumes that an individual can specify a node of the domain generalization hierarchy in order to decide the level of anonymity that he can work with. This approach allows for direct protection of the sensitive values of individuals than a vanilla *k*-anonymity method which is susceptible to different kinds of attacks.

IV. DISTRIBUTED PRIVACY-PRESERVING DATA MINING

The computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants is the aim of most distributed methods for privacy preserving data mining. Thus, the participants may not fully trust each other in terms of the distribution of their own data sets but may wish to collaborate in obtaining aggregate results. The data sets may either be horizontally partitioned or be vertically partitioned for the purpose. The individual records are spread out across multiple entities, each of which has the same set of attributes in the case of horizontally partitioned data sets. The individual entities may have different attributes (or views) of the same set of records for vertical partitioning. Both kinds of partitioning pose different challenges to the problem of distributed privacy preserving data mining. There will be a relation between distributed privacy preserving data mining and cryptography for determining secure multi-party computations. For computing functions over inputs provided by multiple recipients without actually sharing the inputs with one another and the broad approach to cryptographic methods tends to be used.

V. PRIVACY-PRESERVATION OF APPLICATION RESULTS

In many cases, the output of applications can be used by an adversary in order to make significant inferences about the behaviour of the underlying data. There are a number of miscellaneous methods for privacy preserving data mining which tend to preserve the privacy of the end results of applications such as association rule mining and query processing. There will be increasingly sophisticated methods for adversaries to make inferences about the behaviour of the underlying data provided due to the advances in data mining methods, the problem is related to that of disclosure control [6] in statistical databases. As the association rules may represent sensitive information for target-marketing purposes, which need to be protected from inference where the commercial data needs to be shared.

There will be number of issues of disclosure control for a number of applications such as association rule mining, classification, and query processing. From the end results of data mining and management applications, here is a need to prevent adversaries from making inferences.

A. Association Rule Hiding

In recent years, there will be tremendous advances in the ability to perform association rule mining effectively. Such rules often encode important target marketing information about a business [7]. Two broad approaches are used for association rule hiding:

- Distortion: In distortion, the entry for a given transaction is modified to a different value. Since, there will be typically dealing with binary transactional data sets, the entry value is flipped.
- Blocking: In blocking, the entry is not modified, but is left incomplete. Thus, unknown entry values are used to prevent discovery of association rules.

It has been noted that both the distortion and blocking processes have a number of side effects on the non-sensitive rules in the data. Some of the non-sensitive rules may be lost along with sensitive rules, and new ghost rules may be created because of the distortion or blocking process. Since they reduce the utility of the data for mining purposes, such side effects are undesirable.

VI. LIMITATIONS OF PRIVACY: THE CURSE OF DIMENSIONALITY

In the presence of public information, many privacy preserving data mining methods are inherently limited by the curse of dimensionality. For example, the technique in [8] analyzes the *k*-anonymity method in the presence of increasing dimensionality. When adversaries may have considerable background information, the curse of dimensionality becomes especially important. The boundary between pseudo-identifiers and sensitive attributes may become blurred as a result of the curse of dimensionality. This is generally true, since adversaries may have greater information about them than what is publicly available and may be familiar with the subject of interest. This is also the motivation for techniques such as *l*-diversity [4] in which background knowledge can be used to make further privacy attacks. The work in [8] concludes that in order to maintain privacy; a large number of the attributes may need to be suppressed which leads to the data lose its utility for the purpose of data mining algorithms. The broad intuition behind the result in [8] is that when attributes are generalized into wide ranges, the combination of a large number of generalized attributes is so sparsely populated. It has been noted that the problem of high dimensionality seems to be a fundamental one for privacy preservation, and it is unlikely that more effective methods can be found in order to preserve privacy when background information about a large number of features is available to even a subset of selected individuals. Indirect examples of such violations occur with the use of trail identifications [9] [10], where information from multiple sources can be compiled to create a high dimensional feature representation which violates privacy.

VII. GENOMIC PRIVACY

Recent years have seen tremendous advances in the science of DNA sequencing and forensic analysis with the use of DNA. As a result, the databases of collected DNA are growing very fast in the both the medical and law enforcement communities. DNA data contains almost uniquely identifying information about an individual and so it is considered extremely sensitive. As in the case of multidimensional data, simple removal of directly identifying data such as social security number is not sufficient to prevent re-identification. In [11], it has been shown that software called CleanGene can determine the identifiability of DNA entries independent of any other demographic or other identifiable information. The software relies on publicly available medical data and knowledge of particular diseases in order to assign identifications to DNA entries. It has been shown in [11] that 98-100% of the individuals are identifiable using this approach. The identification is done by taking the DNA sequence of an individual and then constructing a genetic profile corresponding to the sex, genetic diseases, the location where the DNA was collected etc. This genetic profile has been shown in [11] to be quite effective in identifying the individual to a much smaller group. One way to protect the anonymity of such sequences is with the use of generalization lattices which are constructed in such a way that an entry in the modified database cannot be distinguished from at least $(k - 1)$ other entities. Another approach discussed is construction of synthetic data which preserves the aggregate characteristics of the original data, but preserves the privacy of the original records.

Another method for compromising the privacy of genomic data is that of trail re-identification, in which the uniqueness of patient visits patterns [9] [10], is exploited in order to make identifications. The premise of this work is that patients often visit and leave behind genomic data at various distributed locations and hospitals. The hospitals usually separate out the clinical data from the genomic data and make the genomic data available for research purposes. While the data is seemingly anonymous, the visit location pattern of the patients is encoded in the site from which the data is released. It has been shown in [9] [10] that this information may be combined with publicly available data in order to perform unique re-identifications.

VIII. CONCLUSION

In this paper, there will be analyses of various privacy preserving data mining techniques. A variety of data modification techniques such as randomization and *k*-anonymity based techniques have been discussed. Also some of the fundamental limitations of the problem of privacy-preservation in the presence of increased amounts of

public information and background knowledge have been discussed. The level of uncertainty or resistance to data mining algorithms, data utility, its performance are some of the measures for the success of privacy preserving data mining algorithm. Some of the privacy preserving algorithm outperforms on some possible criteria. Rather, an algorithm may perform better than another on one but not on all criteria.

REFERENCES

[1] Malin B., Sweeney L., *"Determining the identifiability of DNA database entries"*, Journal of the American Medical Informatics Association, pp. 537–541, November 2000.

[2] Fienberg S., McIntyre J., *"Data Swapping: Variations on a Theme by Dalenius and Reiss",* Technical Report, National Institute of Statistical Sciences, pp. 14–29, 2003.

[3] Aggarwal C. C., *"On Randomization, Public Information and the Curse of Dimensionality",* ICDE Conference, pp. 136-145, 2007.

[4] Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V., *"Disclosure limitation of sensitive rules"*, Workshop on Knowledge and Data Engineering Exchange, 1999, DOI: 10.1109/KDEX.1999.836532.

[5] S. Rathod, B.J. Doddegowda , "*m-Privacy Preserving Data Analysis And Data Publising*", International Journal of Computer Sciences and Engineering, Vol.2, Issue.6, pp.54-58, 2014.

[6] Machanavajjhala A., Gehrke J., Kifer D., and Venkitasubramaniam M., *"l-Diversity: Privacy Beyond k-Anonymity",* ICDE, 2006, DOI: 10.1109/ICDE.2006.1.

[7] Malin B, Sweeney L., *"Re-identification of DNA through an automated linkage process",* Journal of the American Medical Informatics Association, pp. 423–427, 2001.

[8] Aggarwal C. C., Yu P. S., *"On Variable Constraints in Privacy-Preserving Data Mining",* SIAM Conference, pp. 115-125, 2005.

[9] Aggarwal C. C., *"On k-anonymity and the curse of dimensionality",* VLDB Conference, pp. 901–909, 2005.

[10] Li N., Li T., Venkatasubramanian S, *"t-Closeness: Privacy beyond k-anonymity and l-diversity",* ICDE Conference, 2007, DOI: 10.1109/ICDE.2007.367856.

[11] Malin B., *"Why methods for genomic data privacy fail and what we can do to fix it"*, AAAS Annual Meeting, Seattle, WA, 2004.

Authors Profile

Mr. P. Mohana Chelvan working as an Assistant Professor in Department of Computer Science at Hindustan College of Arts and Science, Chennai, India since 2015. His educational qualifications are MCA, NIELIT C Level (IT), MPhil. (CS) and UGC NET. He is currently Ph.D. research scholar in Computer Science in Madurai Kamaraj University, Madurai, India in the area of privacy preserving data mining.

Dr. K. Perumal working as an Associate Professor in Department of Computer Applications at Madurai Kamaraj University, Madurai, India since 1990. He awarded his Ph.D. degree in Computer Science from Madurai Kamaraj University in the area of digital image processing. He has contributed more than 50 papers in the International Journals and Conferences and also editor of proceedings for National Conference on Contemporary Developments in Information and Communication Technologies. He has guiding 9 scholars. His research interest includes Data Mining, Big Data and image processing especially in medical image processing.

