International Conference on Substantial Development in the field of Engineering and Management
Mahratta Chamber of Commerce, Industries and Agriculture Tilak Road, Pune (India)
29th December 2019 www.conferenceworld.in
ISBN : 978-81-943584-8-0

A Survey on Privacy Preserving Data Mining Techniques

Mr. Shrikant D. Zade, Dr.Chandikaditya Kumawat, Dr. Pradeep Chouksey

[1]Research Scholars ,CSE, Mewar University(Raj)
[2]Prof. MEWAR Uni, (Raj)
[3]Asso. Prof. ,LNCT, Bhopal

Abstract: *In the recent year, each and every activity is going to be online. Most of the data to be online stored. Since there are so many data mining algorithm such as association rule mining, clustering, classification, etc. can be apply to find out the knowledge from these data set. Simultaneously there should be risk of violating security of individual person since her/his data may be leak and misuse by third party. Recently ADHAR card data might be leaked and most of the companies use for their purposes.*

This paper provides a complete review on different PPDM techniques such as data partition, data modification, randomization, SMC techniques which could be used to prevent the data access from unauthorized users.

Keywords: privacy, k-anonymity, l-diversity, fuzzy inference system.

Introduction: Recently, DM has been analysis is a danger to privacy for the reason that due to rise of electronic data maintain by companies. This directed to enlarged concern regarding the privacy of the essential data. Recently, numerous technique have been planned to modify or to alter the data so as to maintain privacy. In this paper, we will study an outline of PPDM and focus on different PPDM technique. PPDM come across many applications which is logically theoretical to be "privacy-violating" application such as Draw Something Free – D, Words with Friends – D, GO Locker - D. The main task is to design technique or model which keeps on being efficient, without compromising security. In the paper by Aggrawal R. and Srikant R. (2000), so many methods has been discussed for bio-supervision, facial identification and point out data stealing. Most of the methods for security and privacy computation apply a number of type changes on the records to carry out privacy protection.

PPDM Technique:

The randomization method: This method use as technique for PPDM. In this technique, noise is inserted within records to cover the field value of records. Agrawal R. and Srikant R.

ISBN : 978-81-943584-8-0

(2000) , Agrawal D. and Aggarwal C. C. (2002). The noise inserted is adequately huge so that person attribute values cannot be recovered. Thus, this scheme has been devised to obtain cumulative distributions from altered data. Afterwards DM techniques are design to work with this cumulative distribution.

The *l*-diversity and *k*-anonymity: The *k*-anonymity technique is designed since the chance of indirect finding of records from civilian datasets. This is due to combination of attribute values that can be used to precisely recognize person records. In the *k*-anonymity method, decrease the granularity of information depiction with the help of technique such as aggregation, generalization, suppression and elimination. In recent years, data mining has been viewed as a threat to privacy because

of the widespread proliferation of electronic data maintained by corporations.

This has lead to increased concerns about the privacy of the underlying data.

In recent years, a number of techniques have been proposed for modifying or transforming the data in such a way so as to preserve privacy. A survey on

some of the techniques used for privacy-preserving data mining may be found in [15]. In this chapter, we will study an overview of the state-of-the-art in

privacy-preserving data mining.

Privacy-preserving data mining finds numerous applications in surveillance which are naturally supposed to be "privacy-violating" applications. The key is to design methods [16] which continue to be effective, without compromising security. In [16], a number of techniques have been discussed for bio-surveillance, facial identification, and identity theft. More detailed discussions on some of these may be found in [7, 11–16].

Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms.

This granularity is minimized satisfactorily that any records map against at least *k* another records in the dataset. The *l*-diversity technique by Li. N., Li T. and Venkatasubramanian S. (2007) was planned to hold various weakness in the *k*-anonymity technique so that identities

ISBN : 978-81-943584-8-0

to the stage of *k*-records is not the identical as protecting the equivalent susceptible value, particularly there is homogeneity of responsive values contained by a set.

This type of approach is given by Oliveira S. R. M., Zaiane O. and Saygin Y. (2004), in which the left hand side value has been changed to a changed value. Since, they are usually discussed binary transactional databases.

Following table summaries the different PPDM technique.

Data Hiding	Data Perturbation	Value Distortion	Additive Perturbation
			Multiplicative Perturbation
			Data Micro Aggregation
			Data Anonymization
			Data Swapping
			Other Randomize Tech.
		Probability Distribution	Sampling Method
			Analytical Method
	Secure Multi-Party Computation (SMC) / Cryptographic Protocol		
	Distributed Data Mining		
Rule Hiding	Association Rule Mining	Data Perturbation	
		Data Blocking	
	Classification Rule Mining	Parsimonious Downgrading	

Table 2.1 Brief Overview of Privacy Preserving Data Mining Technique

Authors noted that, the distortion and blocking processes both have various problems. New association rules are designed due to blocking or distortion. But there are some demerits such as they minimize the use of data for mining purposes. A proper proof of NP-hard of the alteration technique for hiding association rule are given by Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E. and Verykios, V. (1999). In this technique authors proposed for shifting a few of value 1 to value 0 so that the support of the equivalent susceptible rules is correctly lower. This technique was enhanced by Dasseni E., Verykios V., Elmagarmid A.and Bertino E. (2001) in which support and confidence of the appropriate rules could be lowered.

ISBN : 978-81-943584-8-0

A another set of methods for distributed PPDM is discussed by Clifton C., Kantarcioglou M., Lin X. and Zhu M.(2002) in which the security for multi-party has been discuss for number of data mining technique.

In horizontally partitioned datasets, different parties include different group of records with the equal group of attributes which are used for DM. Many of these methods use particular version of common methods discussed by Clifton C., Kantarcioglou M., Lin X. and Zhu M. (2002), Du W. and Atallah M. (2001) for a variety of problems.

The OT protocol can be used in case of vector distance in DM primitives. A typical problem of scalar dot product in distributed environment has been discussed by Ioannidis I., Grama A. and Atallah M. (2002).

In vertically partitioned datasets, many traditional primary operations such as computing the secure set size intersection or scalar product can be useful in computing the results of DM algorithms. For instance, the method by Ioannidis I., Grama A. and Atallah M. (2002) discusses gives frequent item set counting for scalar dot product computation. The process of counting can be attained by using the safe size of set intersection as given by Clifton C., Kantarcioglou M., Lin X. and Zhu M. (2002). An additional method for association rule mining given by Evfimievski A., Srikant R., Agrawal R. and Gehrke J. (2002) talk about use of scalar product for vertical bit representation of an item-set in transactions. This step is applied frequently within the framework of a roll up procedure of item-set counting. It has been shown by Evfimievski A., Srikant R., Agrawal R. and Gehrke J. (2002) that such type approach is fairly effective in practice. The process of vertically partitioned has been further process to a number of DM application for decision trees by Vaidya J. and Clifton C. (2005), Naive Bayes Classifier by Vaidya J. and Clifton C. (2004) and *k*-means clustering by Vaidya J. and Clifton C. (2003). A number of experimental results for different types of functions in vertically partitioned databases with the use of cryptographic methods are discussed by Dwork C. and Nissim K. (2004).

Data micro-aggregation is one of the famous data perturbation techniques in the field of SDBs. For a dataset with one personal attribute, uni-variate micro-aggregation sorts data

ISBN : 978-81-943584-8-0

records with the help of private entities. Multivariate micro-aggregation considers all the attributes and groups data using a clustering technique.

There are two multivariate micro-aggregation schemes proposed by researchers in the DM area. The algorithm proposed by Aggarwal C. C. and Yu P. S. (2004) presented a condensation scheme to PPDM. This scheme firstly divides the actual data into numerous group of fixed size. The anonymized data can be disclosed for DM activities and this statistical information is used to create anonymized data that has identical statistical aspects to the original dataset. This scheme preserves data covariance amongst data records. Authors proposed a *kd* -tree based perturbation, which recursively split a dataset into smaller subset in such a way that data records in each subset are more similar after every division of data record. The personal data in each subgroup are then altered using the subgroup average.

Still, two necessary questions endure have to be answered: 1) What are the theoretical lower bound and upper bound of the restoration error; and 2) What are the basic aspects that influence the correctness of the data reconstruction. The paper by Hughes D. and Shmatikov V, (2004) examine the Spectral Filtering scheme and derived an upper limit for the Frobenius norm of the reconstruction error applying matrix perturbation scheme. Authors also proposed a Singular Value Decomposition (SVD) based reconstruction process and found lower limit for the reconstruction error. Authors proved the correspondence among the SF and SVD methods, and as a result, the lower limit of SVD methods can also be studied as the lower limit of the SF methods. Huang identify the main factor that decides the correctness of data reconstruction is the association among the data attributes. Their results have discovered when the association are large and the actual data can be restoring more precisely. They further proposed two data reconstruction methods based on data correlations: used the PCA and Bayes Estimate (BE) scheme, which in essence processing literature on filtering random additive noise.

Number of another technique has been proposed for association rule hiding with some side effects are discussed by Wu Y.-H., Chiang C.-M. and Chen A. L. P. (2007). The aim of this

ISBN : 978-81-943584-8-0

scheme is to minimize the loss of non-sensitive rules, or the creation of ghost rules through the process of rule hiding.

Hence, to identify the attacks it is necessary to track incidence of these common diseases as well. The solution proposed by Polat H. and Du W. (2005) which is of selective revelation allows only limited access to the data. This method applied to the problem of classification as discussed in Liu K., Kargupta H. and Ryan J. (2006). Multiplicative perturbations can also be used for distributed PPDM.

A disparity on this scheme with the use of distance Fourier transforms, which work efficiently for a number of cases given by Mukherjee S., Chen Z. and Gangopadhyay S. (2006).

Another exciting model of personalized anonymity has been discussed by Xiao X. and Tao Y. (2006) in which a person defines the level of privacy for sensitive values. This move towards the benefit that, it allows for security of the sensitive data of persons vulnerable to dissimilar types of attacks.

The problem of utility-based PPDM was first studied properly in Kifer D. and Gehrke J. (2006). The broad idea in Kifer D. and Gehrke J. (2006) is to improve the curse of dimensionality by independently publishing minor tables containing attributes which have usefulness. A method for utility-based DM using local recoding has been proposed in Xu J., Wang W., Pei J., Wang X., Shi B. and Fu A. W. C. (2006). This method is based on different attributes have different efficacy from an application point of view.

Another indirect method for utility based anonymization for privacy-preservation algorithms are more attentive of the workload given by LeFevre K., DeWitt D. and Ramakrishnan R. (2006). In LeFevre K., DeWitt D. and Ramakrishnan R. (2006), an effective and efficient algorithm has been proposed for workload aware anonymization.

The *t*-closeness is an improvement over the concept of *l*-diversity. One feature of the *l*-diversity technique is that, it takes all values of a given attribute in an analogous way irrespective of its sharing data. This is not often the case for actual datasets, since the attribute

ISBN : 978-81-943584-8-0

values may be much skewed. Often, an opponent use previous knowledge of the universal distribution to make conclusion about sensitive values in the data. In the paper by Li N., Li T. and Venkatasubramanian S. (2007), a *t*-closeness model was proposed which uses the principle of the distance between the distribution of the sensitive attribute within an anonymized set. Additionally, the *t*-closeness methods tend to be more efficient than any other PPDM methods in the case of numeric attributes.

The new technique given by S.Zade and P.Chouksey (2017) for PPDM is fuzzy based approach. The aim of authors is to apply FIS on the data set and achieve clustering by hiding original data from the user. At the same time the task can be perform by k-anonymization technique. The main aim is to find out optimal solution as compared to the k-anonymization which has complexity of O (k log k). Their algorithm is very simple to achieve privacy in data mining.

Conclusion: The primary objective of PPDM is promoting algorithm to conceal sensitive data or over privacy. These sensitive data do not get revealed to unapproved parties or invader.

In data mining there exists a trade of between utility and privacy of data. Many PPDM techniques in existence are reviewed in the paper. Ultimately, it is concluded with the fact that there is no single PPDM technique in existence that outshines every other technique with relation to each possible criterion such as use of data, performance, difficulty, compatibility with procedures for data mining, and so on. A particular algorithm may function better when compared to another, on a specific criterion. Various algorithms may be found to function better than one another on given criterion.

Researchers are doing extensive research in ensuring that the sensitive data of a person is not revealed as well as not compromising the utility of data so that the data can be useful for many purposes.

References:

[1] Agrawal R. and Srikant R., "Privacy-Preserving Data Mining," Proceedings of the ACM SIGMOD Conference, 439-450, 2000.

ISBN : 978-81-943584-8-0

[2] Agrawal D. and Aggarwal C. C., "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," ACM PODS Conference, 247-255, 2002.

[3] Li N., Li T. and Venkatasubramanian S., "t-Closeness: Privacy beyond k-anonymity and l-diversity," ICDE Conference, 1-15, 2007.

[4] Oliveira S. R. M., Zaiane O. and Saygin Y., "Secure Association-Rule Sharing," PAKDD Conference, 1-11, 2004.

[5] Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E. And Verykios, V., "Disclosure limitation of sensitive rules," Workshop on Knowledge and Data Engineering Exchange, 1-8, 1999.

[6] Dasseni E., Verykios V., Elmagarmid A. and Bertino E., "Hiding Association Rules using Confidence and Support," 4th Information Hiding Workshop, 1-34, 2001.

[7] Clifton C., Kantarcioglou M., Lin X. and Zhu M., "Tools for privacy preserving distributed data mining," ACM SIGKDD Explorations, 4(2), 28-34, 2002.

[8] Du W. and Atallah M., "Secure Multi-party Computation: A Review and Open Problems," CERIAS Tech. Report 2001-51, Purdue University, 1-10, 2001.

[9] Ioannidis I., Grama A. and Atallah M., "A secure protocol for computing dot products in clustered and distributed environments," International Conference on Parallel Processing, 1-6, 2002.

[10] Evfimievski A., Srikant R., Agrawal R. and Gehrke J., "Privacy-Preserving Mining of Association Rules," ACM KDD Conference, 217-228, 2002.

[11] Vaidya J. and Clifton C., "Privacy-Preserving Decision Trees over vertically partitioned data," Lecture Notes in Computer Science, Vol 3654, 14.1-14.7, 2005.

[12] Vaidya J. and Clifton C., "Privacy-Preserving Naive Bayes Classifier over vertically partitioned data," SIAM Conference, 879-898, 2004.

[13] Vaidya J. and Clifton C., "Privacy-Preserving *k*-means clustering over vertically partitioned Data," ACM KDD Conference, 206-215, 2003.

[14] Dwork C. and Nissim K., "Privacy-Preserving Data Mining on Vertically Partitioned Databases," CRYPTO, 1-17, 2004.

ISBN : 978-81-943584-8-0

[15] Aggarwal C. C. and Yu P. S., "A Condensation approach to privacy preserving data mining," *EDBT Conference*, 183-199, 2004.

[16] Hughes D. and Shmatikov V., "Information Hiding, Anonymity, and Privacy: A modular Approach," Journal of Computer Security, 12(1), 3–36, 2004.

[17] Wu Y.-H., Chiang C.-M. and Chen A. L. P., "Hiding Sensitive Association Rules with Limited Side Effects," IEEE Transactions on Knowledge and Data Engineering, 19(1), 29-42, 2007.

[18] Polat H. and Du W., "Privacy-Preserving Top-N Recommendations on Horizontally Partitioned Data," Web Intelligence, 816-822, 2005.

[19] Liu K., Kargupta H. and Ryan J., "Random Projection Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining," IEEE Transactions on Knowledge and Data Engineering, 18(1), 92-106, 2006.

[20] Mukherjee S., Chen Z. and Gangopadhyay S., "A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier based transforms," VLDB Journal, 293-315, 2006.

[21] Xiao X. and Tao Y., "Anatomy: Simple and Effective Privacy Preservation," *VLDB* Conference, pp. 139-150, 2006.

[22] Kifer D. and Gehrke J., "Injecting utility into anonymized datasets," SIGMOD Conference, pp. 217-228, 2006.

[23] Xu J., Wang W., Pei J., Wang X., Shi B. and Fu A. W. C., "Utility Based Anonymization using Local Recoding," ACM KDD Conference, 785-790, 2006.

[24] LeFevre K., DeWitt D. and Ramakrishnan R., "Workload Aware Anonymization," KDD Conference, 277-286, 2006.

[25] Zade S., and Chouksey P. , "Implementation of privacy preserving in data mining by fuzzy inference system," IEEE International Conference on Smart Technology for Smart Nation,17-19 Aug 2017, Rewa University, Bengaluru.