A Survey on Privacy Preserving Data Mining Approaches and Techniques

Maheyzah Md Siraj, Nurul Adibah Rahmat, Mazura Mat Din
School of Computing,
Faculty of Engineering,
Universiti Teknologi Malaysia,
Johor, Malaysia.
+607-5532206
maheyzah@utm.my

ABSTRACT

In recent years, the importance of the Internet in our personal as well as our professional lives cannot be overstated as can be observed from the immense increase of its users. It therefore comes as no surprise that a lot of businesses are being carried out over the internet. It brings along privacy threats to the data and information of an organization. Data mining is the processing of analyze larger data in order to discover patterns and analyze hidden data concurring to distinctive sights for categorize into convenient information which is collected and assembled in common areas and other information necessities to eventually cut costs and increase revenue. In fact, the data mining has emerged as a significant technology for gaining knowledge from vast quantities of data. However, there was been growing concern that use of this technology is violating individual privacy. This tool aims to find useful patterns from large amount of data using by mining algorithms and approaches. The analysis of privacy preserving data mining (PPDM) algorithms should consider the effects of these algorithms in mining the results as well as in preserving privacy. Therefore, the success of privacy preserving data mining algorithms is measured in term of its performances, data utility, level of uncertainty, data anonymization, data randomization and so on based on data mining techniques and approaches are presented in this paper to analyze.

CCS Concept

• Theory of computation→Data structures design and analysis

Keywords

Data Mining, privacy preserving, knowledge, PPDM (Privacy Preserving Data Mining

1. INTRODUCTION

Data mining is one of the core processes in knowledge discovery of databases. That means the task of creating aggregate models from the available data and hence it requires the extraction of potentially useful information from huge collections of data with a range of application areas from many purposes to precise data for the same. There are many various techniques and approaches that will be recognized for privacy preserving techniques or algorithms. Many organizations include private and government sectors stored their related information in their database. All information stored is shared within their related party in the business operational process. Information is limitless and it is increased from time to time. The organization needs to manage the big amount of information in the database due to the gradually increasing volume of the information. In addition, the security concern rose during the data processing steps and thus the data transmission over the network must be safe and secured to avoid attacks at all the time. In fact, the privacy must protect the mining aspects include association rules, classification rules, randomization, anonymization, and clustering. Therefore, PPDM is an allow to diffusion of respondent data while conserving respondent privacy. Privacy Preserving Data Mining discourses the offending of emerging exact techniques about accumulated data without access to accurate information in individual data record.

2. BACKGROUND

Data mining is one of the core processes in knowledge to extract larger database. The extraction of potentially useful information from large collections of data with a variety of application areas which are data stored either in databases, data warehouse, machine learning or other information repositories. Data mining [1] with its assurance to efficiently discover valuable, non-obvious information from large database, is particularly vulnerable to misapply. So, there might be a dissension among data mining and privacy. Therefore, data mining is one of the business intelligence techniques because it can extract valuable knowledge from huge databases and find hidden relations between data. The completer and more accurate the data, the better the data mining results. Privacy [2] refers to sensitive information for using data mining to extract. The more specific term which is defined as the right of an individual to keep his personal information from being revealed. Privacy is an essential respect to protect sensitive data from revealed and leaked to untrusted person or to public domain easily. In other hand, the general privacy issues are secondary use of the personal information, handling misinformation and crumbled access to personal information. There were concern while allowing access to different classes of the data set such as business and medical dataset for mining [3] In medical dataset especially, a person's specific data and disease must be not revealed into public domain. Nowadays several known in PPDM techniques and approach exist and these are comprehensively studied [4].

ICSCA '19, February 19–21, 2019, Penang, Malaysia
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6573-4/19/02…$15.00
https://doi.org/10.1145/3316615.3316632

2.1 Privacy Preserving

Privacy is becoming as an important various data mining application to secure and prevent from malicious attacked stole the data without permission. There is some type on that such as healthcare, security, financial, business and other types of sensitive data. There are many issues of privacy preserving surrounding that take where is more effectiveness on model or algorithm to secure the data from malicious within different issues have in order to privacy requirement. The bigger company or business within a many account number would like a far complex level of privacy preservation a customer within fewer high level of privacy protections. Therefore, the result of resolves can be used by an adversary in order to make significant inferences about the behavior of the underlying data.

2.2 Data Mining

Data mining is one of the business intelligence techniques and approaches because it can extract costly knowledge from huge databases and find hidden relations between data [5]. This technique is to find out previously unknown relationship and new patterns in big data which even can predict for future decisions by the use of some helpful algorithms and techniques like clustering, classification, association, regression and others. The completer and more accurate the data, the better data mining results. In privacy application method is most of the issues are with integrity of the data, ensuring that the data is confidential and available when it is needed. Also, by monitoring the user data for a long-time attacker can extract private information about the user and can use this sensitive data for his/her personal gain.

2.3 PPDM

Privacy Preserving Data Mining (PPDM) is a technique used to ensure that the privacy of a certain individual is not embraced while the process of data mining [6]. PPDM is a whole process during collect and result of data mining. This process must be applied in three steps which are data acquisition, application of data mining algorithm and interpretation of information. Based on [7], utmost privacy preserving data mining methods or algorithms was applied that bring down a transformation that give the potency of the primary data. Hence, there is regular exchange between solitude and exactness, even this swapping is affected by the particular algorithm. The main problem is to remain all-out effectiveness of the data without compromise the essential confidential constraint. The problem is scheming efficacy-based algorithm to service effectively with assured types of data mining was addressed. The most distributed methods for privacy preserving data mining is to allow computation of useful quantity statistics over the entire of data set without compromising the privacy of the individual data sets. Therefore, the problem of distributed privacy-preserving data mining overlapping closely with a field in cryptography for determining secure multi-party computations.

3. PPDM TECHNIQUES

There are mainly two methodologies in data mining that in turn aided researches come up with various techniques for PPDM which are:

- protect the sensitive data itself in the mining process.
- protect the sensitive data mining results (i.e. extracted knowledge) that were produced by the application of the data mining.

However, there are several exertions on this approach to solve those issues. This section discusses on the techniques or algorithms that have been implemented in the existing works. These include additive noise, random projection, random perturbation, association rules, clustering and classification. Based on these methodologies, PPDM techniques can be broadly categories into:

3.1 Additive Noise

Additive noise is one of the techniques to solve the issue of privacy disclosure. The technique [8] which added random noise to preserve the privacy of attribute values. The process of this additive noise technique includes the de-identification of the original data by adding the noise into it which can achieve the goal of preserving the privacy of the dataset. This technique is used to modify the original dataset by transforming the confidentiality of attributes in order to achieve the goal of privacy preserving. By using this technique, it is no longer used to estimate the original values of individual records. The implementation of the additive noise is the simplest as it is able to preserve the privacy from being disclosed.

3.2 Random Projection

Random projection method is used to overcome privacy disclosure during the preservation process. The random projection method recently becomes a powerful method in the dimensionality reduction. It can process either noisy or noiseless dataset. The original data is projected into the random low-dimensional subspace to generate results and is comparable to the conventional dimensionality reduction techniques. The originality of the data can be preserved well by using this random projection method. The computation of the random projection is significantly cheaper compared to principle component analysis [9]. The complexity of the random projection computation is easier and cheaper. The random projection can preserve the sensitive information with a small space and the error rate is reasonable by introducing classification.

3.3 Random Perturbation

The random perturbation technique is one of the methods used in randomly modifying the original data by the randomized process. The random value perturbation method can hide the sensitivity of the data in order to achieve the goal of privacy preserving. The technique distorts the sensitive values of the attribute but it is still able to estimate the underlying distribution information. The data swapping algorithm is a method of random perturbation. It was first introduced by Tore Dalenius and Steven Reiss (1978) to preserve the confidentiality of the dataset. The objective of the data swapping is to retain the amount of the information and randomly swapping the data values to preserve the sensitive data [10]. It can remove the relationship of the respondents and the records. It is suitable to be used on sensitive attributes without disturbing the non-sensitive attributes.

3.4 Association Rules

Associate rules algorithm in data mining is the method which scans the dataset of the records and calculates the support and confidence rules. Thus, the algorithm only retrieves the rules which support the higher confidence than the minimum support specified by the owners and confidence threshold. This method can be used to prevent the sensitive information from disclosure. Heuristic approach is fast, efficient and scalable method which can hide the sensitive associate rules. This approach is based on the data distortion technique and blocking technique [9].

3.5 Soft Computing

Soft computing technique refers to the adoption of the system which has high Machine Intelligent Quotient (MIQ). Those techniques are used to conduct different kinds of challenges in data mining. The examples of the soft computing techniques are Fuzzy logic, neural networks, genetic algorithms and rough set. There are some soft computing algorithms that have been used in PPDM such as clustering, classification and optimization [11].

3.6 Clustering

Artificial Neural Network (ANN) is one of the Clustering techniques in soft computing. It has its own nature of distributed data storage, self-organization learning and parallel processing. In this technique, it contains three phases which are model, learning algorithm and activation function. Based on the previous proposed method, they aim to hide the selected attribute from others which can solve the privacy preserving collaborative data mining problem. They modify the data so that the value is similar to the original. Thus, they target to hide the sensitive attribute values. The neural work in data mining consist of three main phases, which are data preparation, rules extracting and rules assessment [12].

3.7 Classification

Referring to previous works, there are some researches that introduce a classification data mining method based on the decision tree. The data mining result which produces by building a decision tree classifier which the original values of the perturbed are very different from the original values. The distribution of the data values is different to the original data values. It proves that the original data is damaged. So, the researchers proposed a reconstruction procedure to estimate the distribution of the original values in order to solve this problem. The researchers are able to build a classifier which can compare the accuracy of classifier built with the original data by using that to reconstruct the distribution. The decision tree method is one of the famous classifications to produce a better result of data mining.

3.8 Optimization

There are some researchers who were working on the optimization algorithm in the data mining in order to preserve the privacy in data mining. There was a group of researchers who proposed the Particle Swarm Optimization (PSO) algorithm to generate an optimal generalization feature set. PSO algorithm is a technique which is based on the heuristic search in population. The researchers found that there is possibility to reduce the data loss by optimizing an aggregated value for whole features and it can solve the Evolutionary Algorithm (EA) optimization problems.

3.9 Normalization

Normalization technique is one of the techniques which is used in the data pre-processing step. The normalization process is one of the data preprocessing methods to improve the effectiveness of the analysis. The normalization method involves several processes in order to exact the normalized data from the dataset. Normalized data have provided better overall database and reduce the redundant data in the database. It has improved the data consistency of the data in the dataset. Among the benefits of the normalization of the dataset is that the normalized data can provide a better security for the data and this is considered to be the most important role. The normalization technique can also increase the consistency and the accuracy of the data in the dataset [13].

4. PPDM APPROACHES

PPDM is introducing in order to overcome the issues that arise in data mining process. The PPDM approaches had been identified and used in related works to preserve sensitive data. Based on previous studies, PPDM techniques can be categorized into five main categories. Below is the discussion on the related techniques of the approaches and related works in PPDM which include the Anonymization approach, Condensation approach, Cryptography approach, Perturbation approach and Randomization approach [14].

4.1 Anonymization Approach

Data publishing cannot be avoided because it is useful for the research and implementation purpose. The disclosure risk of the sensitive information should be minimized to an acceptable level as it can be used to increase the data utility to achieve a balance of the goals of PPDM. Anonymization approach refers to the hidden identity or sensitive data of the information by assuming them for a preservation of the analysis. There are four types of attributes which present the basic form of data in a table. The attributes included explicit identifiers, quasi-identifiers, sensitive attribute and non-sensitive attribute. By implementing this approach, the explicit identifiers which contain the identifiers of the clear record should be removed to achieve privacy preservation but there is a danger of privacy intrusion where the quasi-identifiers which contain the potential identified record combined with the publicly available data are defined as linking attacks [15].

4.2 Condensation Approaches

Condensation approach constructs constrained clusters in dataset and then generates pseudo data from the statistics and they provide an extra protection to avoid the adversarial attack on synthetic data. The problem of classification and aggregation of the data behavior preserved using this condensation method is effective and useful for various kinds of problems in data mining. The pseudo data is also an advantage which does not modify the original data as compared to other approaches. The pseudo data can be processed even when there is no redesign of the data mining algorithm because it is in the same format as the original data. However, it also comes with some side effects such as information loss and it requires the act of condensing a large volume of data into a statistical group.

4.3 Cryptography Approach

Cryptography is used to preserve the privacy of data mining. During the sharing resources within multiparty in business, it requires the preservation of the privacy of the data. Sometimes, they have to reveal necessary information to others but they do not trust the parties involved. Thus, this cryptography method can provide a secured communication to prevent the sensitive information from disclosure. The types of the distribute data can be categorized into two collaborators which include vertically-partitioned and horizontally- partitioned. The individual entities may have different attributes in the same set of a dataset in the vertically- partitioned data but the individual entities are spread out across multiple entities which have the same set of attributes. During the preservation of the privacy, this technique is not allowed to perform when there are more parties are involved in the data sharing process and the data mining result may be compromised during the information- sharing process via the medium of the communication. Thus, this approach does not produce a solution to the PPDM problems where the results of the data are not protected and secured during the computation process [16].

4.4 Perturbation Approach

The perturbation approach had been used in statistical disclosure control and it has the ability to preserve the statistical information. The original values are not significantly similar because they are replaced with some synthetic data values from the perturbed data in the statistical information. This can control the statistical from disclosure because it has an inherent attribute of efficiency, simplicity features and can prevent statistical information disclosure. Thus, the attacker is unable to recover sensitive information from the data which have been published and cannot carry out any sensitive linkage. This is because the perturbed data are totally different from the real-world record. So, the individual record which is in the form of perturbed format, become meaningless to humans and the statistical characteristic can be preserved. However, the implementation with additive noise in the data, synthetic data generation and data swapping can be integrated into this approach. The decision tree is one of the algorithms which cannot be modified in this approach. This is because the approach treats each attribute independently. Another method of synthetic data generation is protecting the owner's privacy and important information by controlling the statistical disclosure [17].

4.5 Randomization Approach

Randomization method has been traditionally used in the context of altering data by probability distribution for methods such as surveys which have an evasive answer bias because of privacy concerns [18]. This approach refers to a statistical technique which is introduced to solve a survey problem. The data will scramble anytime so the data central should be in secret better than pre-defined threshold, where the data is original or encrypted. By using this randomization approach in the privacy preserving data mining, it comes with some benefits such as it is very simple and does not require knowledge of the distribution of other records. This method can be implemented since the data gathering period. In addition, it does not require a trusted server which contains all the original records to perform the anonymization process. There are some various kinds of methods introduced in this approach in order to achieve the privacy preserving data mining. Classification, perturbed transaction, association rule mining, decision tree and soft computing-based method are some of the examples of the method [19].

5. CONCLUSION

The knowledge in data mining process comes from data, and data contain personal information about individuals. In fact, the knowledge is coming from data so that the major objective of privacy preserving data mining is developing algorithm to hide or provide privacy to certain sensitive information so that they cannot be revealed to unauthorized parties or intruder. Some approaches and techniques in PPDM are proposed, however, privacy protection technology needs further research because of the complexity of the privacy problem. As a conclusion, there is no single PPDM technique and an approach in existence that outshines every other technique with relation to each possible criteria such as use of data, performance, difficulty, compatibility with procedures for data mining, and so on.

6. ACKNOWLEDGMENTS

We would like to thank Ministry of Higher Education (MoHE) and Universiti Teknologi Malaysia for funding this work through Research University Grant with vot number (15J10).

7. REFERENCES

[1] Shrada Patel, Ronak Patel, "A Review on Privacy Preserving Data Mining", IJSDR (International Journal for Scientific Research & Developmental) vol 3: 2321-0613, 2016.

[2] M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.

[3] Bhargav Sundararajan, Deeprthi Peri, Nita Radhakrishnan, Mehul Awasthi, "An Extensive Survey of Privacy Preserving Data Mining Techniques", IJCSN International Journal of Computer Science and Network, V olume 6, Issues 5, October 2017.

[4] Vinoth Kumar J, Santhi V, "A Brief Survey on Privacy Techniques in Data Mining", IOSRH Journal of Computer Engineering (IOSR-JCE), volume 18, Issue 4, July-August 2016.

[5] Matwin, S., "Privacy Preserving Data Mining Techniques: Survey and Challenges", in Discrimination and Privacy in the Information Society: 209-221. Springer Berlin Heidelberg, 2013.

[6] Alaa H Hamami, Suhad Abu Shehab, "An Approach for Privacy Preserving and Knowledge In Data Mining Application", Journal of Emerging Trends in Computing and Information Sciences, vol 4: ISSN 2079-8407, January 2013.

[7] R. Agarawal and R. Srikant. "Privacy Preserving Data Mining", ACM SIGMOD Conference on Management of Data, pp:439-450, 2000.

[8] Ella Bingham and Heikki Mannila, "Random Projection in Dimensionality Reduction: Application to Image and Text Data", ACM KDD San Francisco CA USA, 2011.

[9] Tom Krenzke, Katie Hubbell, Mamadou Diallo, Amita Gopinath and Sixia Chen, "Data Coarsening and Data Swapping Algorithms", 2014.

[10] Mynavathi R., Sowmiya N. And V anitha D., "Survey of Various Techniques to Provide Multilevel Trust in Privacy Preserving Data Mining", International Journal of Engineering Science and Technology, volume 3(3): 2127-2133, 2014.

[11] Malik, M. B., Ghazi, M. A., Ali R, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", Third International Conference on Computer and Communication Technology, IEEE:26-31, 2012.

[12] Bhanumathi, S. And Sakthivel, "A New Model for Privacy Preserving Multiparty Collaborative Data Mining".

[13] International Conferences on Circuits Power and Computing Technologies (ICCPCT-2013), IEEE:845-850, 2013.

[14]Ryan Stephens and Ron Plew, "The Database Normalization Process", Sans Teach Yourself SQL in 24 Hours, 3rd Edition, 2002.

[15] Malik, M. B., Ghazi, M. A., Ali R, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", Third International Conference on Computer and Communication Technology, IEEE:26-31, 2012.

[16] Nayan G. And Devi, "A Survey on Privacy Preserving Data Mining: Approaches and Techniques", International Journal of Engineering Science and Technology, volume 3(3):2127-2133, 2011.

[17] Mynavathi R., Sowmiya N. and Vanitha D., “Survey of Various Techniques to Provide Multilevel Trust in Privacy Preserving Data Mining”, International Journal of Engineering Science and Technology, volume 3(3): 217-2133, 2014.

[18] Vinoth Kumar J, Santhi V, “A Brief Survey on Privacy Techniques in Data Mining”, IOSRH Journal of Computer Engineering (IOSR-JCE), volume 18, Issue 4, July-August 2016.

[19] Reena and R.Kuma, “Effect of Randomization for Privacy Preservation on Classification Tasks”, Processing of the International Conference on Informatics and Analaytics(ICIA-16), Pondicherry India, 2016.