Privacy Preserving Data Mining Technique - Issues and Challenges

(IJCSNT) International Journal of Communication Systems and Network Technologies
ISSN Number: 2053-6283
Volume 8, Issue No. 2, 55–64

https://journals.mirlabs.in/index.php/ijcsnt
DOI-10.18486/ijcsnt/8.2.104

Mayur Rathi[1], Anand Rajavat[2]

Abstract

The technology improves the ability of data storage and their processing. Therefore a significant growth on different business domains is observed. The business intelligence requires data and consumer habits but the requirement of data is never fulfilled by a single data source. In this context required to combine the different source of information for making smart decisions or for finding the conclusions. But the limitation is that nobody wants to disclose their client's private and sensitive information to others. Therefore we need a privacy preserving environment to handle the data, it's privacy and conclusion. In this presented paper privacy preserving data mining is the key area of investigation using the available literature. Additionally by concluding the literature need to establish the future prospects for the work.

Keywords

Privacy Preserving Data Mining; Multi Party Data Source; Objective Summarization; Model Improvements; Techniques and Tools

Received: 19 April 2019; **Revised:** 26 June 2019; **Accepted:** 13 August 2019;
Published: 31 August 2019;

[1]Department of Computer Science & Engineering, Shri Vaishnav Vidyapeeth Vishwavidyala, Indore, India.
[2]Department of Computer Science & Engineering, Shri Vaishnav Vidyapeeth Vishwavidyala, Indore, India.

Corresponding author:
Email: mayurrathi.31dec@gmail.com

Introduction

Data mining is an application of analyzing data, for extraction of meaningful patterns. That helps to understand and distinguish data variations in different applications. The use of data mining is also performed for making decision and predicting the future aspect of growth or downfall. Classically the utilization of data mining approaches is found in various centralized data bases. Data is stored in one place and the data mining algorithms and techniques are applied to perform the application centric task[1].

There is another scenario where data are distributed in different data sources. Additionally the data mining algorithms processes data for recovering the desired facts form data without spoiling privacy and confidentiality of data. Therefore this type of data analysis technique is known as the Privacy Preserving Data Mining (PPDM)[2].

Principally, the privacy preserving data modeling is utilized where different parties are providing data and associated for concluding a common solution. Even though, no one wants to expose their data to other party. In this context the available data may be in form of[3]:

- Vertical partitioned data
- Horizontal partitioned data

In addition of that it may be possible all parties can faith on the common party to analyze or not therefore the environment can also affect the processing of data.

- Semi trusted environment
- Fully trusted environment

In addition of that there are other issues and challenges which can affect the process of privacy preserving data mining and their techniques. Therefore in this paper the detailed study of privacy preserving data modeling is conducted. Additionally on the basis of collected literature the future objectives are established.

Background

This section incorporates the relevant information and key terms which are used during the proposed privacy preserving data mining domain.

Data Mining

Data mining is an art of handling data in order to discover the application oriented patterns and decision discovery. Therefore, it is a rich domain of applications and possibilities. A number of applications now in these days utilizing the services of data mining for improving their production, business directions, market analysis and more[4]. There are two main classes of data mining algorithms supervised and unsupervised. In supervised learning some predefined samples are required to learn on the patterns. After successful learning the supervised algorithms are able to recognize the similar patterns. On the other hand the unsupervised algorithms not required to train with the predefined samples, these algorithms are directly applicable to the data for discovering patterns.

Privacy Preserving

Now in these days every business domain gathering ends client's data. Among the gathered data some of part of data is much confidential, sensitive and private. Additionally, discloser of such kind of data to any other party can harm the end client privacy socially and financially. Therefore, during the process of data pattern discovery, it is required to preserve the private and confidential information for experimentation and other analytical task in any business or research domain[5]. In other terms keep the sensitive data confidential from the unauthorized persons and environment is known as the privacy preserving.

Data Discloser

As discussed in privacy preserving scenario some of the data is sensitive, private and confidential such kind of data distribution can harm the end client. Therefore before discloser of data for public or experimental use it is required to transform the data's actual figure, encrypt the data or encode data to non understandable format is necessary. But data discloser can be done either intentionally or by mistake. Therefore. handling of such kind of data is much essential[6]. Therefore, in some of the cases the data sensitivity scanning and their normalization is essential part of security quality of service. After that process the data is dis-closeable for public media.

Dimensionality Reduction

The privacy preserving Data mining techniques requires data from different parties and data sources. Additionally, the data is clubbed from different sources in terms of attributes and the number of instances. Therefore, it has a significant dimension of data. In addition of that as the number of parties increases the dimensions of data can be increased. The processing of higher dimensional data requires the significant amount of processing cost i.e. time and space complexity. Therefore it is required to identify the essential features from the bulk set of data to minimize the data processing cost and improving the performance of classification[7].

Impact of Noise on Data

We try to explain an example for describing this context. A person attending phone call in market or in high traffic area, it is much complicated to understand all the communication for that person. On the other hand the same person attending phone in their house minimizes the effort to understand the phone call communication. In this context noise can affect the person's perception. In the same way when a machine learning algorithm tries to learn on noisy data it is directly impact on classifier's or learning algorithm's performance. Therefore, it is required to identify the high informative features among entire information available on datasets[8].

Classification

Basically, the classification is a supervised learning technique in data mining and machine learning. In such kind of pattern learning techniques the two sets of data required first set of data is termed as training samples or patterns. These training patterns are predefined observations with their specific outcomes. The learning algorithms or classifiers are learnt on these samples and prepare the model. To recognize the similar pattern on which the learning algorithm trained on[9]. The classification algorithms can be

transparent in nature or it may be opaque. The transparent algorithms develop rules and by invoking these rules the decisions are calculated. On the other hand in opaque algorithms the weights or other factors are calculated for recognition for unknown patterns.

Decision Making

Decision making process can be supervised and unsupervised in nature. A significant amount of literature is available where the unsupervised learning techniques are also used for making decisions using the available data. But in our context the decision making is a conclusion based on the input set of data instances. Therefore, employment of both kinds of algorithms can be possible for making decision and prediction for a given unknown pattern of data instance [10]. In a number of applications, the visualization techniques are used for decision making i.e. heat map, decision trees. In addition of sometimes data mining algorithms are applied. When a journal decision required the visualization, techniques are effective way to preserve the time. On the other hand, the data mining algorithms can be applied when the precise outcomes are expected.

Data Collaboration

In business intelligence domain sometimes it is not feasible to take decision by using incomplete set of information. Therefore, it is required to delegate the additional information from other data source or party to complete the set of knowledge. For finding the essential patterns, to take big decisions in business domain required to complete analysis of data. The complete set of knowledge belongs to similar industries help to make smart decisions. Therefore, the different business owners agreed to aggregate their data for making common decisions. The delegation of data or information also needs to be secured for preserving the privacy of data [11].

Literature Survey

The privacy preserving data mining is one of the promising areas in data science. In order to improve the business and other business intelligence task there are a number of scenarios where we need to handle data sources from different data sources. Additionally, it is also required to combine and explore data to find common fruitful outcomes. The Privacy Preserving Data Modeling is the only method which solves the issue of data privacy and data utility in different situations.

For instance, consider a tour and travel industry where the hotels, restaurants, cabs, and other transportation mediums are combined each other. Additionally, the information about their clients is needed to be preserve by the service providers. In this context if they want to combine their data and want the combined conclusion for any research purpose then security and privacy on data mining is necessary [12].

Consider separate medical institutions that wish to conduct a joint research while preserving the privacy of their patients. One way to view this is to imagine a trusted third party– everyone gives their input to the trusted party, who performs the computation and sends the results to the participants. However, this is exactly what we don't want to do, for example, hospitals are not allowed to hand their raw data out, security agencies cannot afford the risk, and governments risk citizen outcry if they do. Thus, the question is how to compute the results without having a trusted party, and in a way that reveals

nothing but the final results of the data mining computation. Secure Multiparty Computation enables this without the trusted third party[13].

Chen et al.[14] conducted a review on data intensive techniques and their applications. According to this survey when the data is available in higher dimension then for classifiers performance becomes low in this context the technique employed for Dimensionality reduction can be unsupervised[15] and supervised[16] in nature. Sunitha et al.[17] perform experiments on noisy data and outlier data according to their study both the issues impact on classifiers performance. In this context Xiong et al.[18] suggests the technique of noise removal in data. In further Zhang et al.[19] shows the issues of privacy in multiparty data collaboration. Therefore it is required to deal with multiparty data with sensitive information for finding a secure computation and analysis of the data. Therefore the proposed work is intended to address the key issues and challenges involve in privacy preserving data mining efficiency and the performance of mining.

Research Gap Analysis For Existing Methods

This section provides the extracted essential research gap for concluding the proposed work.

Table 1. Research Gap Analysis of Existing Methods

Author, Publication, Year	Research Gap
Kung et al.[20], ACM Transactions on Embedded Computing Systems, July 2017	The proposed technique deals only with dimensionality issues and security concerns in a specific scenario.
Zhang et al.[21], IEEE Transactions on Computers, May 2016	The method does not provide a solution for secure data publishing to third parties while maintaining enhanced utility and assurance.
Nissim et al.[22], Information Sciences, Elsevier, 2010	The approach suffers from over-flexibility, excessive anonymity, high computational cost, and relies on data projection, which is unsuitable for all real-world datasets.
Poovammal et al.[23], Journal of Computer Science (JCS), 2009	The proposed method works only with numerical attributes and does not support categorical data.
Li et al.[24], IEEE Transactions on Knowledge and Data Engineering, March 2012	The work focuses only on high-dimensional data without considering privacy-preserving data mining (PPDM) scenarios.
Xiong et al.[18], IEEE Transactions on Knowledge and Data Engineering, March 2006	The methods perform well only in specific scenarios. The authors suggest combining multiple techniques for better generalization.
Bouzas et al.[16], IEEE Transactions on Neural Networks and Learning Systems, May 2015	The proposed algorithm is not suitable for very high-dimensional datasets.
Fletcher et al.[25], International Journal of Computer Theory and Engineering (IJCTE), February 2015	The authors do not recommend a specific technique for improving information quality.
Hua et al.[26], IEEE Transactions on Information Forensics and Security, October 2016	The work does not adequately address data utility while ensuring query processing capability.
Li et al.[27], IEEE Transactions on Information Forensics and Security, 2016	The proposed scheme increases system complexity in terms of computational resources.
Li et al.[28], IEEE Transactions on Knowledge and Data Engineering, September 2012	The authors do not consider the trustworthiness aspect of data publishing.
Arribas et al.[29], Information Processing & Management, 2012	The study does not address issues related to data aggregation and data mining security.

Issues and Challenges

Data mining techniques and applications are growing continuously. Additionally, its acceptability is also improving day by day in various real-world applications for decision making and another context. But sometimes the available data in a single source is not much sufficient to deal with the issues, therefore, it is required to gather information from other sources, to develop a complete dataset for the decision making the process. Thus the third party data contributor is worried about the data and consumers' privacy, sensitivity and confidentiality due to the risk of information discloser. In this context using the existing literature some key issues are concluded as:

1. **Network Data security issues:** in the PPDM environment, multiple parties are agreed to combine their data for finding the common data model and the decision ability. Thus the data is communicated between parties and the centralized server. The communication between servers and parties is enabled using public networks. Additionally, public networks are not secure thus there is a probability for network security issues during the communication of sensitive and confidential data using network attacks.
2. **Data discloser and data leakage issue:** although the data in PPDM is modified, transformed and/or sometimes encrypted to preserve the privacy and data owners' security. But due to the mistake of algorithm or computational losses, a transformed or modified data can provide the values of a data owner accidentally.
3. **Data noise and outcomes:** in PPDM to preserve the privacy and confidentiality of data, noise is introduced. But, the noise in data can affect the performance of the learning algorithm. Additionally, it can also impact on outcomes of the learning algorithm. Therefore, it is required to introduce the noise in data in a controlled manner.
4. **Dimensionality:** the PPDM collaborate the data of a number of parties. The collaboration of data can be possible in a horizontal or vertical manner. In both, the conditions parties involve their data by which either the number of instances is increases or the number of attributes increases. Therefore both the manner of data aggregation increases the dimensionality of data.
5. **Data utility:** in data mining applications data is processes in their absolute format and the outcome of data mining is used to get benefits for the application. But in PPDM the data is used after modification in the actual values. Therefore the outcome of the data mining can also be noisy and the issue of data utility rises.
6. **Data publishing:** in PPDM the data discloser and publishing are also performed. Therefore, to use data in other application dataset is modified with some noise for security and privacy purposes. The modification of data can impact the application's performance, the application conclusions, and decision-making ability.
7. **Mining environment:** that is an essential question in PPDM. How the data is secure, safe and confidential in third party storage. The data parties are always worried about the trust level of data processing and storage. What happened when the data processing server is compromised with the attacker? This serves the purpose of intention.

In order to achieve an effective PPDM model, we have to consider different aspects of PPDM such as working with the Efficient Data Dimensionality Reduction Techniques to preserves the computational resources. Therefore, investigation of different Dimensionality Reduction Techniques required. On the other hand, the learning algorithm's performance depends on the quality of data, therefore, need to

Measure the Impact of Noise over Classifier Performance. Moreover, there are some approaches available that are preserving the privacy of data using noise. Thus, a suitable and controlled noise composition required to maintain the utility of data.

Proposal for Privacy Preserving Data Model

In previous section the review of existing work is reported and based on the literature collection a probable model is presented in this section.

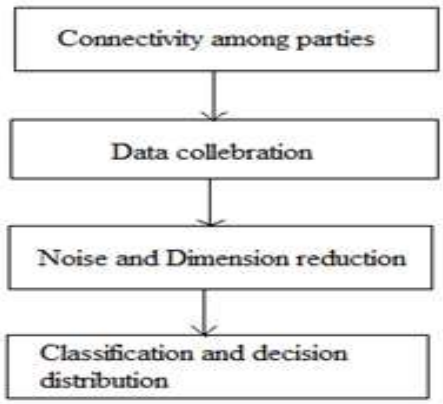

Figure 1. Praposed Data Model

Connectivity among Parties

The proposed data model is aimed to satisfy the requirements of privacy preserving in data mining. Therefore, it may be possible that there are N number of parties exist, which want to combine their data for finding effective decisions. Therefore, this phase of the system is responsible for the following two expectations:

- Connectivity of available parties with the semi trusted server
- Enable security and privacy at the client end for regulating the part of data which are need to be disclosed and essential for decision making. Therefore this phase is also enable the cryptography, data encoding, data transformation or mapping to securing the sensitive part of data. This method will solve the purpose as per pridictions.

Data Collaboration

The combination of data is required in this scenario, but the synchronization of data is also a significant need of the system. The un-synchronized data can misguide the learning algorithm and produces undesired results. Therefore, this phase of system is dedicated to verifying the data instances and their relevance class labels.

Noise and Dimensionality Reduction

After synchronization of data it is required to verify the quality of data and availability of information of the present attributes to reduce the impact of noise and cost of processing for large dimensional data. Therefore in this phase the noise measurement technique or feature selection technique is required to optimize the quality of data.

Classification and decision distribution

That is the final stage of the proposed system. This phase is responsible for mining that data, which is obtained from the previous phase. Using this data, the required application-oriented decisions are developed. Additionally, the data decisions are distributed to all the participating parties who are providing the data for experimentation and decision mining. During the distribution of final decisions and experimental data set discloser it is ensured that the sensitive and private attributes from the data set keep preserved. The attributes of any data are the issues related to originality and will help in verification.

Conclusion and Future Work

Privacy preservation in data mining has become more important in recent years, because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. If information management or possession is distributed, the disclosure of personal data becomes a problem. The privacy preserving data mining leads a number of issues and research challenges, among them accurately data processing is a key research issues. During such kind of data modeling and their publishing the following situations can affect the modeling of data:

- Data nature and dimensions
- Data integration processes and affect of modeling
- Data publishing and quality of outcomes

Privacy Preserving Data Model is a multiparty data mining environment where from multiple data sources data is combined to achieve a common objective. So the aim of proposed work is to preserve the expected level of privacy with minimum information loss, with low error rate and improvement in accuracy in multi-party data sets. In order to achieve the mentioned aim, the future objectives of the proposed work are established as:

1. The Efficient Data Dimensionality Reduction Techniques preserves the computational resources therefore need to investigate different Dimensionality Reduction Techniques for proposed system.
2. The learning algorithms performance depends on the quality of data which is used for training therefore need to Measure the Impact of Noise over Classifier Performance.
3. There are some approaches available that are preserving privacy of data using noise thus what is suitable and effective noise composition required to maintain the utility of data is need to be investigate.
4. The aim of PPDM techniques is to combine and mine the data. Additionally discover the patterns which are useful for all the participating parties. Thus need to Design and Develop the Efficient Common Data Publishing Technique.

References

[1] Mukhopadhyay A, Maulik U, Bandyopadhyay S et al. A survey of multi-objective evolutionary algorithms for data mining: Part i. *IEEE Transactions on Evolutionary Computation* 2014; 18(1): 20–35.

[2] Aldeen YAAS, Salleh M and Razzaque MA. A comprehensive review on privacy preserving data mining. *Springer Plus* 2015; 4(694): 1–36.

[3] Danasana J, Kumar R and Dey D. Mining association rules for horizontally partitioned databases using ck secure sum technique. *International Journal of Distributed and Parallel Systems* 2012; 3(6): 149–157.

[4] Ponce J and Karahoca A. *Data Mining and Knowledge Discovery in Real Life Applications*. Croatia: In-Teh, 2009.

[5] Mendes R and Vilela JP. Privacy-preserving data mining: Methods, metrics, and applications. *IEEE Access* 2016; 5: 10562–10582.

[6] Xu L, Jiang C, Wang J et al. Information security in big data: Privacy and data mining. *IEEE Access* 2014; 2: 1149–1176.

[7] Vasan KK and Surendiran B. Dimensionality reduction using principal component analysis for network intrusion detection. *Perspectives in Science* 2016; 8: 510–512.

[8] Nettleton DF, Orriols-Puig A and Fornells A. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review* 2010; 33(4): 275–306.

[9] Anitha P, Krithka G and Choudhry MD. Machine learning techniques for learning features of any kind of data: A case study. *International Journal of Advanced Research in Computer Engineering & Technology* 2014; 3(12): 4324–4331.

[10] Chitradevi B and Thinaharan N. Role of decision making in data mining systems. *International Journal of Trend in Research and Development* 2015; 2(5): 122–125.

[11] Swamy SK, Manjula SH, Venugopal KR et al. Association rule sharing model for privacy preservation and collaborative data mining efficiency. In *RAECS 2014 Proceedings of 2014 Recent Advances in Engineering and Computer Sciences*. Chandigarh, India: IEEE, pp. 1–6.

[12] Basiri A, Amirian P, Winstanley A et al. Making tourist guidance systems more intelligent, adaptive and personalised using crowd sourced movement data. *Journal of Ambient Intelligence and Humanized Computing* 2017; 9:413, https://doi.org/10.1007/s12652-017-0550-0.

[13] Kou G, Peng Y, Shi Y et al. Privacy-preserving data mining of medical data using data separation-based techniques. *Data Science Journal* 2007; 6: 429–434.

[14] Chen CLP and Zhang CY. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences* 2014; 275: 314–347. Https://doi.org/10.1016/j.ins.2014.01.015.

[15] Xu C, Tao D, Xu C et al. Large-margin weakly supervised dimensionality reduction. In *Proceedings of the 31st International Conference on Machine Learning*. Beijing, China, pp. 865–873.

[16] Bouzas D, Arvanitopoulos N and Tefas A. Graph embedded nonparametric mutual information for supervised dimensionality reduction. *IEEE Transactions on Neural Networks and Learning Systems* 2015; 26(5): 957–963.

[17] Sunitha L, Raju MB and Srinivasa BS. A comparative study between noisy data and outlier data in data mining. *International Journal of Current Engineering and Technology* 2013; 3(2): 575–577.

[18] Xiong H, Pandey G, Steinbach M et al. Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering* 2006; 18(3): 304–319.

[19] Zhang W, Lin Y, Xiao S et al. Privacy preserving ranked multi-keyword search for multiple data owners in cloud computing. *IEEE Transactions on Computers* 2015; 6(1): 1–14.

[20] Kung SY, Chanyaswad T, Chang JM et al. Collaborative pca/dca learning methods for compressive privacy. *ACM Transactions on Embedded Computing Systems* 2017; 16(3): 77–117.

[21] Zhang Q, Yang LT and Chen Z. Privacy preserving deep computation model on cloud for big data feature learning. *IEEE Transactions on Computers* 2016; 65(5): 1351–1362.

[22] Nissim M, Rokach L and Maimon O. Privacy-preserving data mining: A feature set partitioning approach. *Information Sciences* 2010; 180(14): 2696–2720.

[23] Poovammal E and Ponnavaikko M. Utility independent privacy preserving data mining on vertically partitioned data. *Journal of Computer Science* 2009; 5(9): 666–673.

[24] Li T, Li N, Zhang J et al. Slicing: A new approach for privacy preserving data publishing. *IEEE Transactions on Knowledge and Data Engineering* 2012; 24(3): 561–574.

[25] Fletcher S and Islam MZ. Measuring information quality for privacy preserving data mining. *International Journal of Computer Theory and Engineering* 2015; 7(1): 21–28.

[26] Hua J, Tang A, Fang Y et al. Privacy-preserving utility verification of the data published by non-interactive differentially private mechanisms. *IEEE Transactions on Information Forensics and Security* 2016; 11(10): 2298–2311.

[27] Li L, Lu R, Choo KKR et al. Privacy-preserving outsourced association rule mining on vertically partitioned databases. *IEEE Transactions on Information Forensics and Security* 2016; : 1847–1861.

[28] Li Y, Chen M, Li Q et al. Enabling multi-level trust in privacy preserving data mining. *IEEE Transactions on Knowledge and Data Engineering* 2012; 24(90): 1598–1612.

[29] Arribas N, Torra GV, Erola A et al. User k-anonymity for privacy preserving data mining of query logs. *Information Processing & Management* 2012; 48(3): 476–487.