International Journal of Computer Applications Technology and Research
Volume 13–Issue 06, 180 - 183, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1306.1016

Privacy-Preserving Data Mining: Techniques and Applications

George M. Kimwomi
Institute of Computing and Informatics
Technical University of Mombasa
Mombasa, Kenya

Fullgence Mwakondo
Institute of Computing and Informatics
Technical University of Mombasa
Mombasa, Kenya

Wilson Cheruiyot
Institute of Computing and Informatics
Technical University of Mombasa
Mombasa, Kenya

Abstract: Continued advancement in technology and the resulting digitalization in society has generated large electronic databases which are unstructured. The databases contain valuable hidden information which could be extracted through data mining to support organizations in strategic management. But these data sets consist of private and sensitive data about individuals and organizations which could be exposed in the data mining process thereby breaching the right to privacy of the owners. Even so, the valuable information hidden in the data need to be extracted to support organizations make strategic decisions. A range of privacy preservation techniques have been developed which could be used to protect private and sensitive data in a data mining process so that the right to privacy is not breached. This paper sought to discuss the privacy preservation techniques in data mining and their areas of application.

Keywords: Privacy preservation, data mining, privacy preservation techniques, privacy preservation applications

1. INTRODUCTION

Advancements in computing technologies including the public internet since the early 1990s brought about widespread acquisition of computers and computerization of operations in organizations and the society in general. The automated systems have tremendously increased the amount of data generated, stored and shared which has become a threat to personal privacy. The data originates from a range of sources such as production, sales, stock movement, product details, marketing, client feedback and the general society (Custers et al., 2013; Aggarwal, 2020).

Data is a valuable asset which has made many organizations such as the Amazon, Facebook and Google to thrive from knowledge uncovered from the data through data mining (Nair & Tyagi, 2021). Analysis of the data can provide important trends about markets, the society and even individuals for strategic decisions (Shah & Gulati, 2016). But these data are in an unstructured nature and cannot be used directly by any system for analysis to extract the information. To overcome this challenge, data mining which is also known as Knowledge Discovery in Databases (KDD) is used, which is a process of extracting knowledge from large unstructured data sets into a format which can be represented and understood for decision making (Aggarwal, 2020). In data mining, the unstructured data is cleaned, processed and analyzed to extract the insightful information beneficial to organizations. Data mining has been beneficial in different application areas such as health care, cyber security, commerce, banking, and transportation (Mendes & Vilela, 2017).

The unstructured data may contain personal private information of customers and data collectors such as identity, names, medical records, political orientation and financial information which is sensitive and protected by laws on privacy (Martínez et al., 2013). These data could get revealed in the data mining process which could compromise the privacy of owners. To protect sensitive and private information from exposure at any stage, privacy-preserving data mining methods are used. Privacy-preservation data mining (PPDM) are methods used to protect sensitive and private information from exposure in data mining and obtain accurate results without compromising the data (Han & Kamber, 20). They preserve the privacy of the data through masking or erasure of sensitive data.

PPDM is used in two dimensions of security namely protection of individual privacy and collective or organizational knowledge. Individual privacy is the protection of sensitive data of an individual such as name, identity number, address, medical data and political inclination. Collective data includes sensitive knowledge uncovered from data mining such as financial records which is confidential to an organization or some group which needs protection (Sharma et al., 2013).

Privacy-preservation in data mining can be implemented during data collection and publication dates, or at the time of output. Data collection can be done anonymously using a software plugin so that the data collection team is not able to know the identity of the data contributors. This can also be done during publication through hiding private data from users, during output with data mining algorithms by preventing users from determining association patterns which may provide private information (Aggarwal, 2020).

The process of privacy-preserving data mining starts by creating a new transformed data set from the old database. The new set hides sensitive information but maintains patterns and trends from the old set as shown in figure 1. PPDM employs association rule to identify relationships among data items, and clustering to group objects into classes of similar objects so as uncover new information patterns.

International Journal of Computer Applications Technology and Research
Volume 13–Issue 06, 180 - 183, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1306.1016

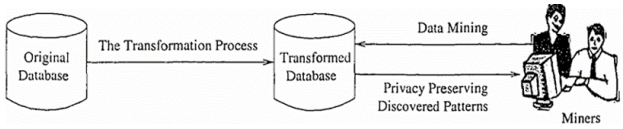

Figure 1: Privacy-preserving data mining process (Oliveira, 2005)

Privacy-preserving with association rules is used to protect sensitive information uncovered from data mining by representing it in the form of sensitive association rules which are private to the organization and therefore protected. The privacy-preserving clustering task aims to protect features of the objects which are analyzed in a cluster. While the privacy-preserving association rule data mining seeks to protect sensitive knowledge, privacy-preserving clustering aims to protect individual privacy (Oliveira, 2005). The United Nations recognized the right to personal privacy in its Universal Declaration of Human Rights in 1948 but this is limited to privacy at home, with family, and in correspondence. The scope of privacy is limited to four areas namely information handling and collection, bodily harm with invasive procedures, any form of communication, and territorial boundaries.

Due to the different areas of application in scope, privacy doesn't have a universally accepted definition (Mendes & Vilela, 2017). In relation to information handling which relates to data control in privacy-preserving data mining, privacy is defined as the right of an individual to be protected from unauthorized disclosure of electronically stored personal information (Bertino et al., 2008). Other authors have given definitions which conform to this definition in terms of control in data collection and handling so as to protect it from exposure to unauthorized persons. PPDM is an important tool which enables secure use of the large electronic databases arising from different sources without compromising individual privacy and sensitive organizational information. It facilitates secure data mining while preserving the privacy and security information.

2. PPDM TECHNIQUES

Different PPDM techniques are used to protect data privacy in a data mining process based on the location where the task is executed. The techniques are either applied in distributed manner by the individual contributors, or entrusted to a third party referred to as the Central Commodity Server. In a distributed privacy preserving situation, individual contributing parties protect their private data before publication using heavy privacy preserving techniques such as Cryptographic and Secure Multiparty Computation (SMC). Data mining can also be done at the data distribution end. In a Central Commodity Server, a trusted third party receives and protects sensitive data of the contributing parties and proceeds to undertake data mining preferably through anonymization and perturbation methods (Shah & Gulati, 2016).

The PPDM techniques can be classified using different approaches, such as by data distribution, data modification, data mining algorithms, data (or rule) hiding and privacy preservation techniques (Verykios et al., 2004). This study descried the PPDM based on their classification by data distribution approach which are explained in the next section (Shah & Gulati, 2016).

2.1 Anonymization techniques

Anonymization is a Central Commodity Server based method to which the contributing individuals entrusts the responsibility of protecting their privacy to a third party before publication. It is applied when the data needs to be published in its original form without encryption or modification while the privacy of the owners is protected. The technique uses suppression, generalization, data removal, swapping, permutation and other anonymization methods to achieve privacy.

Among different anonymization methods used, k-anonymization is the conventional choice as established from a survey on PPDM techniques (Shah & Gulati, 2016). K-anonymization works by first mapping the potentially identifiable data attributes using aggregation for numeric data while nominal data are swapped.

The other k-anonymization developed include ldiversity, (α,k) anonymity, t-closeness, psensitive k-anonymity, km - anonymization, (α,k) anonymity, (k,e) anonymity. Anonymization provides additional privacy preservation by retaining quasi-identifiers, which consist of attributes such as a rare condition, diagnosis or age, which may be combined to identify a person are masked by anonymization while retaining the utility of the data for quality conclusions (Martínez et al., 2013).Instead of removing quasi-identifiers which may affect the quality of data mining, a k-anonymization is used whereby any combination of quasi-identifier is repeated k-times or more so that the quasi-identifier attributes in every record becomes indistinguishable from at least k-1 other records. The performance of k-anonymization improves when the k-value is higher.

2.2 Perturbation Technics

Perturbation is a set of methods which use camouflage to protect data thereby distorting data before data mining in a distributed and a central commodity computing environment in a simple and efficient manner (Custers et al., 2013). Perturbation is achieved by altering an attribute value by adding noise or replacing it with anew value by the contributor thereby creating a local perturbed copy. The new attribute value can be noise or a selected new value. Users can create a multilevel privacy for a multilevel Trust PPDM which can be used to reconstruct the original data set by the miner based on the settings of the contributor.

Noise addition is allowed in a perturbation by probability distribution using a Gaussian (normal distribution) or other known distribution pattern, using noise addition, randomization or condensation (Vidya Banu & Nagaveni, 2013). A perturbed copy can be locally created by the individual contributor by adding noise. Once the local perturbed copy is generated the miner can reconstruct the perturbed version to obtain the original data distribution.

Another perturbation technique is randomization which is a simple and valuable data protection technique through distortion with some random data. Random perturbation works on discrete data which could be a character, number, classification or Boolean types which improves privacy and data mining accuracy but with increased costs (Aldeen et al., 2015).

The weakness with perturbative methods is that it modifies the original data which may not be acceptable in some classes of application like medical data mining as the quality of output is compromised.

2.3 Cryptographic Techniques

Cryptographic techniques are methods which can be used to protect data in a distributed scenario and for information sharing by using a precise privacy model and methods to prove and quantify. Techniques based on cryptography include Secure Multiparty computations (SMC), homomorphic encryption, Oblivious Transfer and Digital Envelope in a distributed environment (Shah & Gulati, 2016). SMC is a cryptographic privacy protection technique in which multiple parties compute their data securely whereby a party can only know about its own input or output results. It is used when multiple parties with individual data records could use a union of their data sets which are for a PPDM securely using some algorithm without exposing each other input. This method preserves individual privacy and leakage, and is efficient in network resource utilization (Oliveira, 2005). SMC preserves privacy from semi-honest adversaries having the right protocols but tries to input incorrect information, and malicious attacks as well.

Homomorphic is also a PPDM data mining encryption method which enables multiple parties having private data sets to collaborate in association data mining while maintaining the privacy of each other's data. Homomorphic protocol encryption was designed to support exchange of data for multiple parties in a distributed environment without a trusted central party while keeping it private to each party (Zhan et al., 2005). Homomorphic method aims to solve a data mining problem by combining data from multiple users while maintaining individual privacy.

Oblivious transfer protocol is another encryption method for privacy preservation which uses multiparty computation approach for horizontally partitioned datasets. A horizontally partitioned datasets refers to when every party has different records which have similar attributes, with the aim of conducting a global data mining to uncover insights from the data (Mendes & Vilela, 2017). Oblivious transfer encryption could be applied to datasets which are horizontally partitioned such as a chain clinic with sites having different customers while the attributes for every customer such as diagnosis are common in all the sites.

The weakness with cryptography technics is that they only prevent leakage and not the results of the output which makes perturbation to be a preferred technique (Hewage et al., 2023).

2.4 Fuzzy algorithms

Fuzzy algorithm technic are Central Commodity Server based algorithms in which a third party is entrusted to preserve data privacy by enabling anonymization without experiencing any significant information loss. The technique achieves these by creating distinct clusters of similar records which are indistinguishable. Fuzzy technique employs different algorithms such as k-means clustering, apriori algorithm, c-regression and fuzzy classifiers (Shah & Gulati, 2016).

By the use of the k-means technique, anonymization is applied to a cluster k record so that it is indistinguishable from the rest of the k-1 clusters so as to protect privacy. The fuzzy apriori is a modified algorithm which can identify sensitive rules so as to preserve its privacy in a distributed case. The fuzzy c-regression technique is used to generate synthetic data to enable computation to be done by third parties with limited risk of so as to limit the risk of disclosure to sensitive (Shah & Gulati, 2016); (Hewage et al., 2023).

2.5 Neural Network Techniques

The neural network are techniques which make use of computational models based on biological neural networks to preserve data privacy without compromising on information loss. The techniques include probabilistic neural network, Bayesian networks and kohen SOMs. The probabilistic neural network committee for peer-to-peer data mining is a distributed computation technique which selects the best of weight-based peer member. The Kohen Self Organizing Feature Members (Kohen SOMs) can maintain data with minimum probability of loss and disclosure. Bayesian networks are useful in cases of partitioned data with accuracy, complete privacy and negligible overhead (Shah & Gulati, 2016).

3. PPDM Applications

Privacy preserving data mining has found use in a range of fields which require new knowledge for various strategic applications. One of the key application areas is cloud computing which supports various online services in the collection, storage and analysis of data. This is a distributed network infrastructure which has developed with the advancement in technology where PPDM is applied. Organizations which use cloud services can entrust the service provider with their data or employ privacy preservation techniques such as encryption to protect the data. Homomorphic encryption supports encryption of data queries and results for protection from intermediate parties. Homomorphic encryption is also applied to preserve association rules in a vertical data distribution environment (Zhou et al., 2015) (Mendes & Vilela, 2017).

Homomorphic encryption method could be used for collaborative data mining on a vertically partitioned data (heterogeneous collaboration) in a scenario where privacy preservation is required in a distributed environment without a central server. The multiple parties could be rival organizations with related private data sets such as supermarkets which could be combined in a collaborative data mining for some common benefit such as buying trends of customers. Other areas of application could be with security firms, medical research and business among others (Zhan et al., 2005)

Oblivious transfer encryption technique uses multiparty computation approach on horizontally partitioned datasets. It has been applied horizontally partitioned data in chain business model such as chain clinics and chain schools which could have different records but similar record attributes. A global data mining on such a dataset could discover global trends of the data which could benefit parties in the chain (Mendes & Vilela, 2017).

PPDM is also applied to sensitive electronic health records mainly with cryptographic techniques. An example is the protection of the highly identifiable genomic data which is very sensitive and private. Homomorphic data aggregation methods were used in genome sequencing in the study of genetic information for individuals (Naveed et al., 2015).

The techniques have also been employed in Wireless Sensor Networks (WSN) of the sparsely distributed autonomous sensors used to monitor light, temperature and the general physical environment. Data collected by the sensors such as

room humidity could be sensitive as it could be used by malicious people to track occupancy behavior and use of electricity among other utilities thereby breaching individual privacy (Taban & Gligor, 2009; Mendes & Vilela, 2017b). In this case, homomorphic encryption was used to ensure sensed data of individuals preserved. k-anonymity has also been used to add synthetic data to hide sensed data values (Groat et al., 2011).

Global positioning system (GPS) and other location-based services which help to obtain accurate information on locations provide information on individuals in regards to their identity movement and residence. Access to this private information about individuals from GPS systems is a threat to privacy. k-anonymity has been recommended privacy preservation in global positioning system (GPS) and any other location-based services (Mendes & Vilela, 2017).

4. CONCLUSION

Many organizations are leveraging on the available large electronic databases and data mining technologies to extract valuable information from the data sets for strategic management. This database contains private and sensitive data about individuals and organizations which much be protected so as not to breach the rights to privacy. We have examined a range of privacy preservation techniques which have been developed to protect private so as not to breach the individual rights to privacy.

The paper has also examined the areas of application for the privacy preservation techniques used in data mining. The privacy preservation techniques for data mining and the applications could be explored further to enrich understanding on the alternative methods of classifying these techniques and their application.

3. REFERENCES

[1] Aggarwal, C. C. (2020). Data mining: The textbook. Springer.

[2] Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. SpringerPlus, 4(1), 694. https://doi.org/10.1186/s40064-015-1481-x

[3] Bertino, E., Lin, D., & Jiang, W. (2008). A Survey of Quantification of Privacy Preserving Data Mining Algorithms. In C. C. Aggarwal & P. S. Yu (Eds.), Privacy-Preserving Data Mining (Vol. 34, pp. 183–205). Springer US. https://doi.org/10.1007/978-0-387-70992-5_8

[4] Custers, B., Calders, T., Schermer, B., & Zarsky, T. (Eds.). (2013). Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases (Vol. 3). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30487-3

[5] Groat, M. M., Hey, W., & Forrest, S. (2011). KIPDA: K-indistinguishable privacy-preserving data aggregation in wireless sensor networks. 2011 Proceedings IEEE INFOCOM, 2024–2032. https://ieeexplore.ieee.org/abstract/document/5935010/

[6] Han, J., & Kamber, M. (20). Data mining: Concepts and techniques (2. ed., [Nachdr.]). Elsevier, Morgan Kaufmann.

[7] Hewage, U., Sinha, R., & Naeem, M. A. (2023). Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: A systematic. https://openrepository.aut.ac.nz/server/api/core/bitstreams/a26ff1b6-e82d-4fd0-a1b5-c316ab4aef2b/content

[8] Martínez, S., Sánchez, D., & Valls, A. (2013). A semantic framework to protect the privacy of electronic health records with non-numerical attributes. Journal of Biomedical Informatics, 46(2), 294–303. https://doi.org/10.1016/j.jbi.2012.11.005

[9] Mendes, R., & Vilela, J. P. (2017). Privacy-Preserving Data Mining: Methods, Metrics, and Applications. IEEE Access, 5, 10562–10582. https://doi.org/10.1109/ACCESS.2017.2706947

[10] Nair, M. M., & Tyagi, A. K. (2021). Privacy: History, statistics, policy, laws, preservation and threat analysis. Journal of Information Assurance & Security, 16(1). https://ak-tyagi.com/static/pdf/13.pdf

[11] Naveed, M., Ayday, E., Clayton, E. W., Fellay, J., Gunter, C. A., Hubaux, J.-P., Malin, B. A., & Wang, X. (2015). Privacy in the Genomic Era. ACM Computing Surveys, 48(1), 1–44. https://doi.org/10.1145/2767007

[12] Oliveira, S. R. D. M. (2005). Data transformation for privacy-preserving data mining. https://era.library.ualberta.ca/items/3d3164ec-6c61-421e-b221-f1d6843d6e09/download/d5d04863-06be-4c44-9e3c-d2bb91da1a01

[13] Shah, A., & Gulati, R. (2016). Privacy preserving data mining: Techniques, classification and implications-a survey. Int. J. Comput. Appl, 137(12), 40–46.

[14] Sharma, M., Chaudhary, A., Mathuria, M., & Chaudhary, S. (2013). A review study on the privacy preserving data mining techniques and approaches. International Journal of Computer Science and Telecommunications, 4(9), 42–46.

[15] Taban, G., & Gligor, V. D. (2009). Privacy-preserving integrity-assured data aggregation in sensor networks. 2009 International Conference on Computational Science and Engineering, 3, 168–175. https://ieeexplore.ieee.org/abstract/document/5283452/

[16] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. ACM SIGMOD Record, 33(1), 50–57. https://doi.org/10.1145/974121.974131

[17] Vidya Banu, R., & Nagaveni, N. (2013). Evaluation of a perturbation-based technique for privacy preservation in a multi-party clustering scenario. Information Sciences, 232, 437–448. https://doi.org/10.1016/j.ins.2012.02.045

[18] Zhan, J., Matwin, S., & Chang, L. (2005). Privacy-Preserving Collaborative Association Rule Mining. In S. Jajodia & D. Wijesekera (Eds.), Data and Applications Security XIX (Vol. 3654, pp. 153–165). Springer Berlin Heidelberg. https://doi.org/10.1007/11535706_12

[19] Zhou, J., Cao, Z., Dong, X., & Lin, X. (2015). PPDM: A privacy-preserving protocol for cloud-assisted e-healthcare systems. IEEE Journal of Selected Topics in Signal