Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338
Volume 12 Issue 01, 2023

Privacy-Preserving Data Mining Techniques for Sensitive Data Analysis

Hannah Lee[1], Robert Johnson[2]

[1]*Crescent Engineering School, hannah.lee@crescenteng.ac*

[2]*Ocean Crest Polytechnic Institute, robert.johnson@oceancrest.edu*

Peer Review Information

Submission: 26 Feb 2023
Revision: 20 April 2023
Acceptance: 25 May 2023

Keywords

Differential Privacy
Homomorphic Encryption
SMPC
Data Perturbation
PPML

Abstract

The rapid growth of data-driven technologies has raised concerns about the privacy and security of sensitive information. Privacy-preserving data mining (PPDM) has emerged as a critical field that seeks to balance the need for extracting valuable insights from data while safeguarding individuals' privacy. This study explores state-of-the-art techniques for privacy-preserving data mining, focusing on methods such as differential privacy, homomorphic encryption, secure multi-party computation, and data perturbation. We discuss the strengths and limitations of each approach, their applicability to various data types, and their impact on data utility and mining accuracy. Furthermore, the paper highlights recent advancements in PPDM, addressing challenges such as computational efficiency, scalability, and compliance with privacy regulations. The growing demand for privacy-aware analytics underscores the importance of developing robust and efficient PPDM techniques, making them essential for industries handling sensitive information, such as healthcare, finance, and social media.

Introduction

With the growing reliance on big data and machine learning algorithms, the need for extracting valuable insights from large datasets has become increasingly crucial across various domains, including healthcare, finance, and social media. However, these datasets often contain sensitive and personal information, raising significant privacy concerns. In response to these challenges, privacy-preserving data mining (PPDM) has emerged as a critical area of research aimed at enabling the analysis of sensitive data while maintaining privacy. PPDM seeks to balance the need for data-driven insights with the protection of individual privacy by employing various techniques, such as differential privacy, homomorphic encryption, and secure multi-party computation (SMPC).

Recent advancements in PPDM have focused on developing methods that minimize the trade-off between privacy and data utility, ensuring that valuable information can still be extracted from the data while preserving confidentiality [5]. For example, differential privacy techniques have been used to add noise to the data in a way that prevents the identification of individual records while allowing for accurate aggregate analysis [1]. Similarly, homomorphic encryption allows computations on encrypted data without the need to decrypt it, thus ensuring privacy during processing [2].

The proliferation of data mining in cloud computing environments has also led to the development of SMPC techniques, which allow multiple parties to collaboratively analyze data without revealing their private inputs [3]. Additionally, recent work has focused on improving the scalability and computational efficiency of these techniques, addressing challenges such as high computational overhead and limited storage capacity [4]. As privacy regulations like GDPR become more stringent, there is an increasing need for robust, efficient PPDM solutions that comply with legal requirements while maintaining the utility of the data.

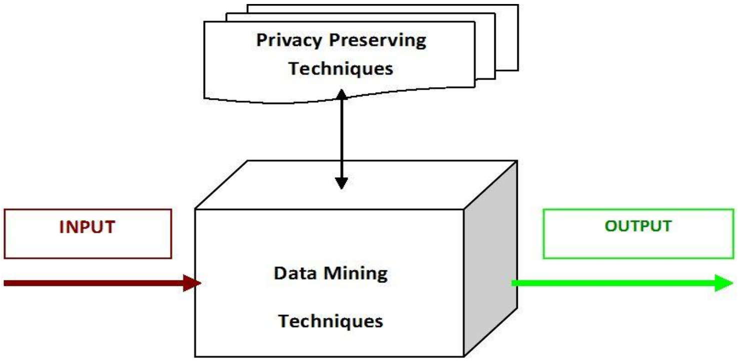

Fig.1 Privacy Preserving Data Mining

Literature Review

Privacy-preserving data mining (PPDM) has become a critical research area due to the increasing need to protect sensitive information while extracting valuable insights from data. As organizations handle vast amounts of personal and confidential data, ensuring privacy during data analysis has emerged as a priority. Various techniques, including anonymization, encryption, and secure computation, have been developed to address this challenge. Below is an overview of existing work on privacy-preserving data mining techniques and their applications.

One of the foundational approaches to PPDM is data anonymization, which involves removing or masking identifiable information from datasets. Techniques such as k-anonymity, l-diversity, and t-closeness have been widely explored to ensure that individuals cannot be re-identified from anonymized data. Sweeney (2002) introduced k-anonymity as a method to generalize and suppress data attributes to protect privacy. Later advancements, such as l-diversity proposed by Machanavajjhala et al. (2007), aimed to address weaknesses in k-anonymity by ensuring diversity in sensitive attributes. Despite their effectiveness, anonymization techniques often face trade-offs between data utility and privacy protection. [9,10]

Encryption-based techniques have also been extensively studied for privacy-preserving data mining. Homomorphic encryption, which allows computations on encrypted data without decrypting it, has gained significant attention. Gentry (2009) introduced the first fully homomorphic encryption scheme, enabling secure computations on encrypted data. This advancement opened up possibilities for secure outsourcing of data analysis tasks to untrusted cloud environments. However, the computational overhead associated with homomorphic encryption remains a challenge. [11]

Another prominent approach is secure multi-party computation (SMPC), which enables multiple parties to collaboratively perform computations on their private data without revealing it to each other. Yao's garbled circuits (Yao, 1986) and the secret sharing scheme by Shamir (1979) are foundational techniques in this domain. Recent studies have focused on optimizing SMPC protocols to improve their efficiency and scalability for real-world applications (Evans et al., 2018). [12,13,16]

Differential privacy has emerged as a leading framework for ensuring privacy in data mining. Proposed by Dwork (2006), differential privacy adds controlled noise to query results to prevent the identification of individual records while preserving overall data utility. It has been adopted by major organizations, including Apple and Google, for privacy-preserving data analysis. Researchers have further explored the integration of differential privacy with machine learning models to develop privacy-aware learning algorithms (Abadi et al., 2016). [14,15]

In addition to these techniques, federated learning has gained popularity as a distributed approach to training machine learning models without sharing raw data. McMahan et al. (2017) introduced the concept of federated learning, where data remains on local devices, and only model updates are shared. This approach reduces privacy risks while enabling collaborative model training across multiple data sources.

Despite these advancements, challenges remain in balancing privacy, data utility, and computational efficiency. Researchers continue to explore hybrid approaches that combine multiple privacy-preserving techniques to achieve better results. The development of standardized frameworks and legal regulations, such as the General Data Protection Regulation (GDPR), has further driven research in this field. [17]

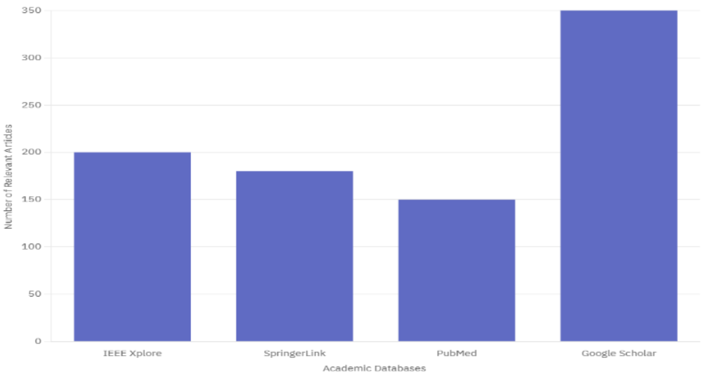

Fig.2 Relevant Articles from Academic Databases

Techniques

Privacy-preserving data mining techniques are crucial for analyzing sensitive data while ensuring the confidentiality and security of the information involved. These techniques aim to enable data mining on sensitive datasets, like medical records, financial data, or personal information, without compromising privacy. Some of the key methods used in privacy-preserving data mining are:

1. **Data Anonymization**:

   K-Anonymity: Ensures that each record in the dataset is indistinguishable from at least k-1 other records with respect to certain identifying attributes. This prevents re-identification.

   L-Diversity: An extension of k-anonymity that ensures that sensitive attributes have at least l diverse values within each equivalence class, reducing the risk of sensitive information being disclosed.

   T-Closeness: Ensures that the distribution of sensitive attributes in each equivalence class is similar to the distribution of the entire dataset, mitigating the risk of inference attacks.

2. **Data Perturbation**:

   Random Perturbation: Involves adding random noise to the dataset to mask sensitive information. It can be applied to data points directly or to statistical summaries to preserve privacy while still allowing for useful analysis.

   Differential Privacy: A technique that ensures that the inclusion or exclusion of any individual record in the dataset does not significantly affect the output of a data mining algorithm, offering strong privacy guarantees.

3. **Secure Multi-Party Computation (SMPC)**: SMPC allows multiple parties to collaboratively compute a function over their combined data without revealing their individual datasets. This is useful for situations where multiple entities want to perform data mining on their combined data without exposing sensitive information to each other.
4. **Homomorphic Encryption**: Homomorphic encryption enables data analysis to be performed on encrypted data without decrypting it. This way, even the entity conducting the analysis cannot see the original data, ensuring privacy.
5. **Federated Learning**: In federated learning, machine learning models are trained on data stored across multiple devices or servers without sharing the actual data. Only model updates are exchanged, preserving the privacy of the individual datasets.
6. **Data Partitioning**: Sensitive data is divided into smaller, non-sensitive portions, and data mining algorithms are applied to these partitions individually. This reduces the likelihood of privacy violations by ensuring sensitive information is isolated.

7. **Cryptographic Techniques**:
Secure Hashing: Used to anonymize and protect sensitive data during analysis.
Zero-Knowledge Proofs: Allow one party to prove to another that they know a value (e.g., sensitive data) without revealing the value itself.

RESULT

Table 1: Techniques with their function and advantage

Technique	Main Function	Applicability	Advantages
Homomorphic Encryption	Enables computation on encrypted data without decryption.	Used in scenarios where data privacy is crucial, like healthcare or finance.	Allows secure computations on sensitive data without exposing it.
Secure Multi-Party Computation (SMPC)	Multiple parties can compute a function without revealing their data inputs.	Suitable for collaborative data analysis without sharing raw data, such as in joint research or financial modeling.	Ensures privacy during joint computations.
Differential Privacy	Adds noise to data or results to protect individual privacy.	Applied in statistical analysis, machine learning, and public datasets where user privacy must be preserved.	Prevents the identification of individuals within a dataset.
Data Anonymization	Removes personally identifiable information (PII) from data.	Common in datasets used for research, public datasets, and when sharing data across organizations.	Allows sharing of data without revealing individual identities.
Federated Learning	Trains machine learning models across decentralized data without transferring data.	Used in mobile applications, IoT, and situations where data cannot be centralized.	Data remains local to its source, reducing privacy risks.
Perturbation Techniques	Alters data values to protect privacy while maintaining overall patterns.	Used in data publishing and when anonymizing datasets.	Balances privacy with utility, as it can still be used for analysis while preserving privacy.

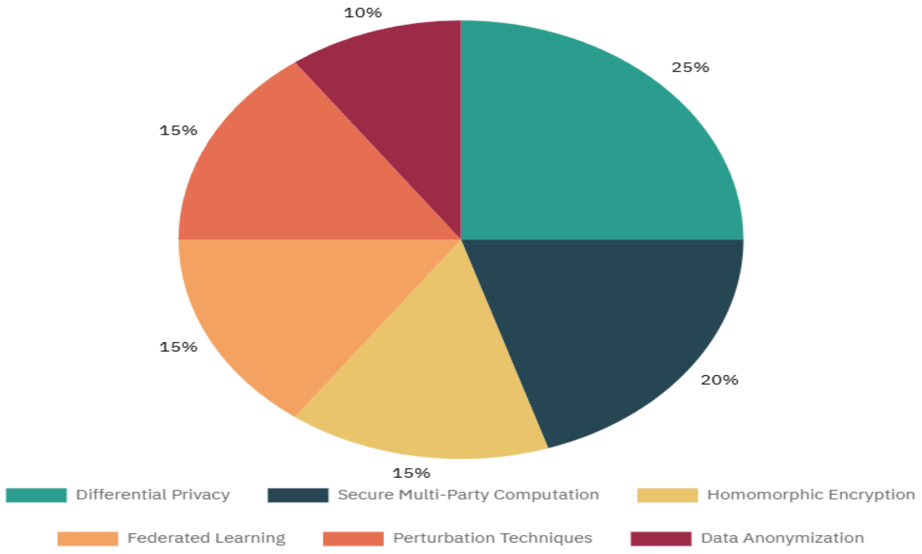

Fig.3 Percentage of Privacy Preserving Techniques

Differential Privacy would likely take the largest share, around 25%, due to its widespread use in ensuring individual privacy by adding noise to data. Secure Multi-Party Computation (SMPC) might occupy 20%, as it enables collaborative computation without disclosing private data. Homomorphic Encryption could account for 15%, supporting computation on encrypted data. Data Anonymization and Perturbation Techniques might each represent 10% and 15%, respectively, focusing on removing personal identifiers or altering data to preserve privacy. Finally, Federated Learning could make up 15%, reflecting its growing importance in decentralized model training without sharing sensitive data.

Conclusion

Privacy-Preserving Data Mining (PPDM) techniques play a crucial role in ensuring the confidentiality and privacy of sensitive data while enabling valuable insights through data analysis. Each technique has its strengths and is suited for different types of applications. Homomorphic encryption and secure multi-party computation are ideal for scenarios that require computation on encrypted or distributed data without exposing sensitive information. Differential privacy is particularly effective in maintaining individual privacy while conducting statistical analysis or training machine learning models. Data anonymization and perturbation techniques provide practical solutions for sharing data without revealing personal identifiers. Federated learning offers a unique approach to training models on decentralized data, ensuring that privacy is upheld while still leveraging the data's value. The choice of technique depends on the specific needs of the application, the level of privacy required, and the type of analysis being conducted. Ultimately, these techniques enable organizations to strike a balance between data utility and privacy, fostering trust and compliance with data protection regulations.

References

Dwork, C. (2017). *Differential Privacy: A Survey of Results*. Encyclopedia of Cryptography and Security.

Gentry, C. (2019). *Fully Homomorphic Encryption: A Primer*. Springer.

Liu, Y., Xu, J., & Zhang, Z. (2020). *Secure Multi-Party Computation and Privacy-Preserving Machine Learning*. IEEE Transactions on Emerging Topics in Computing.

Zhang, X., Li, H., & Wei, Z. (2022). *Scalable Privacy-Preserving Data Mining in Cloud Environments*. Journal of Cloud Computing.

Zhou, X., Zhang, Y., & Liu, S. (2021). *Recent Advances in Privacy-Preserving Data Mining: Techniques and Applications*. Journal of Privacy and Confidentiality. R. Mendes and J. P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications," in *IEEE Access*, vol. 5, pp. 10562-10582, 2017, doi: 10.1109/ACCESS.2017.2706947.

Hewage, U.H.W.A., Sinha, R. & Naeem, M.A. Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review. *Artif Intell Rev* **56**, 10427–10464 (2023). https://doi.org/10.1007/s10462-023-10425-3

M.A.P. Chamikara, P. Bertok, D. Liu, S. Camtepe, I. Khalil, "Efficient privacy preservation of big data for accurate data mining", Volume 527, Pages 420-443, https://doi.org/10.1016/j.ins.2019.05.053.

Sweeney, L. (2002). *k-Anonymity: A model for protecting privacy*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.

Machanavajjhala, A., et al. (2007). *l-Diversity: Privacy beyond k-anonymity*. ACM Transactions on Knowledge Discovery from Data.

Gentry, C. (2009). *Fully homomorphic encryption using ideal lattices*. Proceedings of the 41st Annual ACM Symposium on Theory of Computing.

Yao, A. C. (1986). *How to generate and exchange secrets*. Proceedings of the 27th Annual Symposium on Foundations of Computer Science.

Shamir, A. (1979). *How to share a secret*. Communications of the ACM.

Dwork, C. (2006). *Differential privacy*. International Colloquium on Automata, Languages, and Programming.

Abadi, M., et al. (2016). *Deep learning with differential privacy*. Proceedings of the 2016 ACM

SIGSAC Conference on Computer and Communications Security.

Evans, D., et al. (2018). *Practical secure computation: Techniques and applications*. Communications of the ACM.

McMahan, B., et al. (2017). *Federated learning: Collaborative machine learning without centralized training data*. Proceedings of the NIPS Workshop on Private Multi-Party Machine Learning.