A Review of Privacy Preserving Clustering in Data Mining Using Piecewise Vector Quantization

Shikha Jawre[1] Pradeep Pandey[2]
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]SAM College of Engineering and Technology, Bhopal, India

Abstract — Privacy preservation has become a critical challenge in data mining applications due to the rapid growth of internet-based and cloud-based data sharing environments. While data mining techniques enable the extraction of valuable knowledge from large datasets, they also raise serious concerns regarding the confidentiality of sensitive information. Various privacy preserving data mining (PPDM) techniques such as cryptography, anonymization, perturbation, and secure multiparty computation have been proposed to address these issues. Among data mining methods, clustering techniques play a significant role in privacy preservation by grouping similar data while minimizing information disclosure. This review paper presents an analytical study of privacy preserving data mining techniques with a particular focus on privacy preserving clustering using Piecewise Vector Quantization (PVQ). The paper discusses existing PPDM approaches, highlights the role of clustering, explains the PVQ-based privacy preservation mechanism, identifies research gaps, and outlines future research directions.

Keywords: Privacy Preserving Data Mining, Clustering, Piecewise Vector Quantization, Data Security, Collaborative Data Mining

I. Introduction

In the current digital era, the integrity and security of data transmitted over the internet and cloud networks have become a major concern. Organizations increasingly rely on data mining techniques to extract hidden and useful knowledge from large datasets for applications such as healthcare, banking, marketing, and national security. However, mining sensitive data without violating individual privacy remains a significant challenge. Traditional data protection techniques such as cryptography, steganography, and watermarking provide security during data transmission but are often inefficient for data analysis tasks. As a result, privacy preserving data mining techniques have emerged to ensure that sensitive information is protected during the mining process. Clustering, classification, and association rule mining are widely used data mining techniques for privacy preservation. Among these, clustering is particularly effective because it groups similar records without directly exposing individual data values.

This review focuses on privacy preserving clustering techniques with special emphasis on Piecewise Vector Quantization (PVQ), a method that balances privacy protection and data utility

II. Privacy Preserving Data Mining (PPDM)

Privacy preserving data mining refers to techniques that enable knowledge extraction from data while preventing the disclosure of sensitive information. According to the reviewed PDF, PPDM techniques can be broadly classified into three major categories

A. Secure Multiparty Computation (SMC)

SMC-based techniques allow multiple parties to collaboratively perform data mining tasks without revealing their private data to each other. These techniques are widely used in distributed and collaborative data mining environments but suffer from high computational complexity.

B. Secret Sharing-Based Techniques

In secret sharing approaches, sensitive data is divided into multiple shares and distributed among different parties. No single party can reconstruct the original data independently, thus ensuring privacy.

C. Perturbation-Based Techniques

Perturbation techniques preserve privacy by modifying original data values using noise addition or transformation. While these methods are computationally efficient, they may lead to information loss if not carefully designed.

III. Related Work

Privacy Preserving Data Mining (PPDM) has attracted significant research attention due to the increasing need to protect sensitive information during data analysis. Several studies have explored different privacy preservation techniques using data mining algorithms, particularly association rule mining, clustering, and classification.

Kumaraswamy et al. proposed a key distribution–less privacy preserving data mining system, where local association rules generated by individual parties are securely combined using a commutative RSA algorithm. The combined rule set is then used for data classification and mining. Their experimental results showed improved accuracy when compared using C4.5 and C5.0 decision tree–based systems, demonstrating the effectiveness of secure rule sharing in collaborative data mining environments

1) A_review_on_Privacy_Preservatio…

Kantarcioglu and Jiang focused on incentive-compatible privacy preserving data analysis, emphasizing truthful data sharing among participating parties. Although their approach guarantees that no information other than the final mining result is disclosed, the authors highlighted the challenge of verifying the correctness of input data provided by different participants in distributed environments

2) A_review_on_Privacy_Preservatio…

Tassa introduced a protocol for secure association rule mining in horizontally distributed databases based on the Fast Distributed Mining (FDM) algorithm. The protocol employs secure multi-party computation techniques to compute the union of private subsets and verify element inclusion without

revealing individual datasets, thereby ensuring data confidentiality during collaborative mining

3) A_review_on_Privacy_Preservatio…

Sasikala and Banu addressed efficiency issues in privacy preserving data mining and proposed a privacy preserving clustering approach using Piecewise Vector Quantization (PVQ). Their work demonstrated that traditional distortion-based techniques can significantly increase mining runtime. To overcome this limitation, they introduced a modified K-means LBG algorithm combined with PVQ, which minimizes information loss while maintaining privacy and clustering accuracy

4) A_review_on_Privacy_Preservatio…

Agrawal and Haritsa presented FRAPP, a matrix-theoretic framework for random perturbation in privacy preserving data mining. Their approach showed that accuracy can be significantly enhanced by carefully designing perturbation matrices while still satisfying strict privacy constraints. This framework provided a generalized understanding of existing perturbation-based techniques

5) A_review_on_Privacy_Preservatio…

Somayyeh and Mohammed Reza proposed a classification framework for privacy preserving distributed data mining techniques, categorizing them into Secure Multi-Party Computation–based, Secret Sharing–based, and Perturbation–based methods. Their framework enabled systematic comparison and evaluation of privacy preserving techniques under distributed data scenarios

6) A_review_on_Privacy_Preservatio…

Overall, the reviewed studies indicate that while cryptographic and secure multi-party approaches provide strong privacy guarantees, they often suffer from high computational complexity. Perturbation and transformation-based techniques, such as Piecewise Vector Quantization, offer a practical balance between privacy protection and data utility, particularly in clustering-based data mining applications.

IV. Research Gap and Challenges

Based on the reviewed PDF, the following research gaps are identified

- High information loss in random perturbation techniques
- High computational overhead in cryptographic and SMC-based methods
- Limited research on clustering-specific privacy preservation
- Need for efficient transformation-based techniques with minimal data loss

PVQ addresses some of these challenges, but further improvements are required for scalability and adaptive quantization.

V. Role of Clustering in Privacy Preservation

Clustering is an unsupervised data mining technique that groups data objects based on similarity. In privacy preserving scenarios, clustering helps in hiding individual data records by representing them as part of a group. This reduces the risk of sensitive data disclosure while retaining useful patterns for analysis.

However, traditional clustering algorithms may still leak private information if raw data values are directly used. Therefore, privacy preserving clustering techniques are required to transform data before clustering is performed.

VI. Privacy Preserving Clustering Using Piecewise Vector Quantization

Piecewise Vector Quantization (PVQ) is a transformation-based privacy preserving technique used in clustering applications. As discussed in the reviewed work, PVQ operates by dividing data vectors into segments and applying quantization independently to each segment

A. Working Principle of PVQ

- Original data vectors are partitioned into multiple pieces
- Each piece is quantized separately using predefined codebooks
- The quantized vectors replace original values during clustering
- Sensitive information is concealed while preserving similarity relationships

B. Advantages of PVQ

- Reduces direct exposure of original data values
- Preserves clustering accuracy
- Minimizes information loss compared to random perturbation
- Efficient for large datasets

C. Limitations of PVQ

- Selection of optimal quantization levels is challenging
- Privacy-utility tradeoff must be carefully balanced
- Not suitable for all data types without customization

VII. METHOD OF PRIVACY PRESERVATION

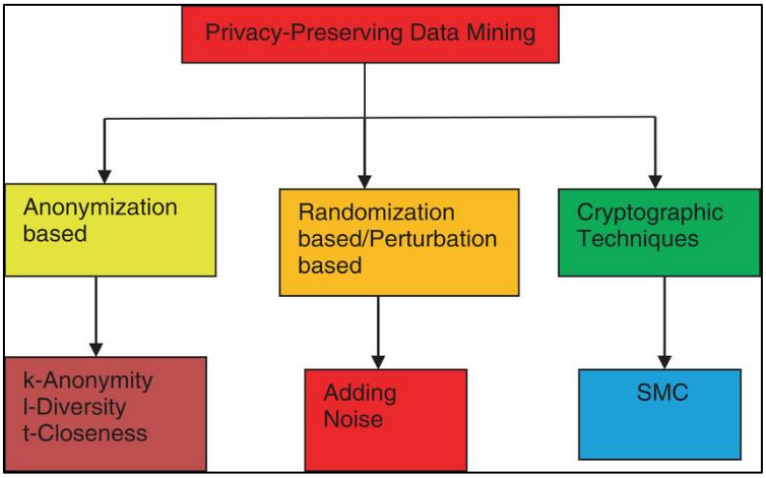

Fig. 1.A: PPDM framework.

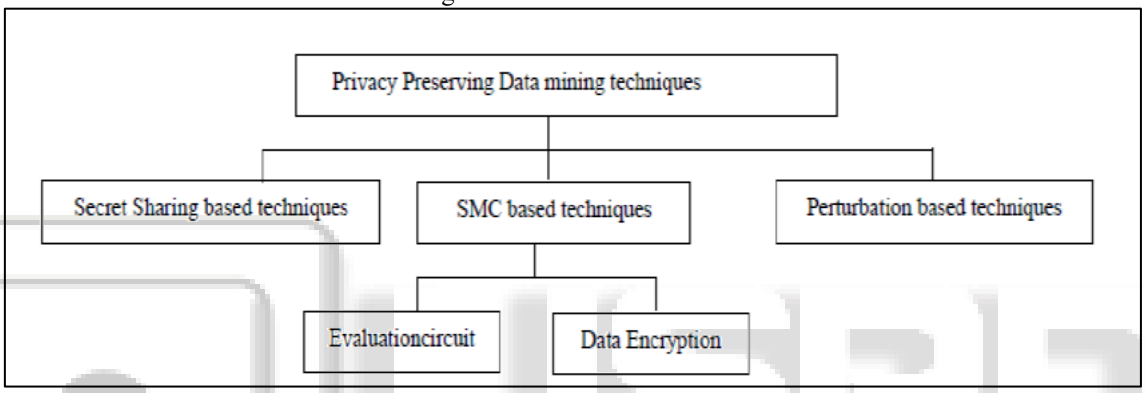

Fig. 1.B: PPDM framework.

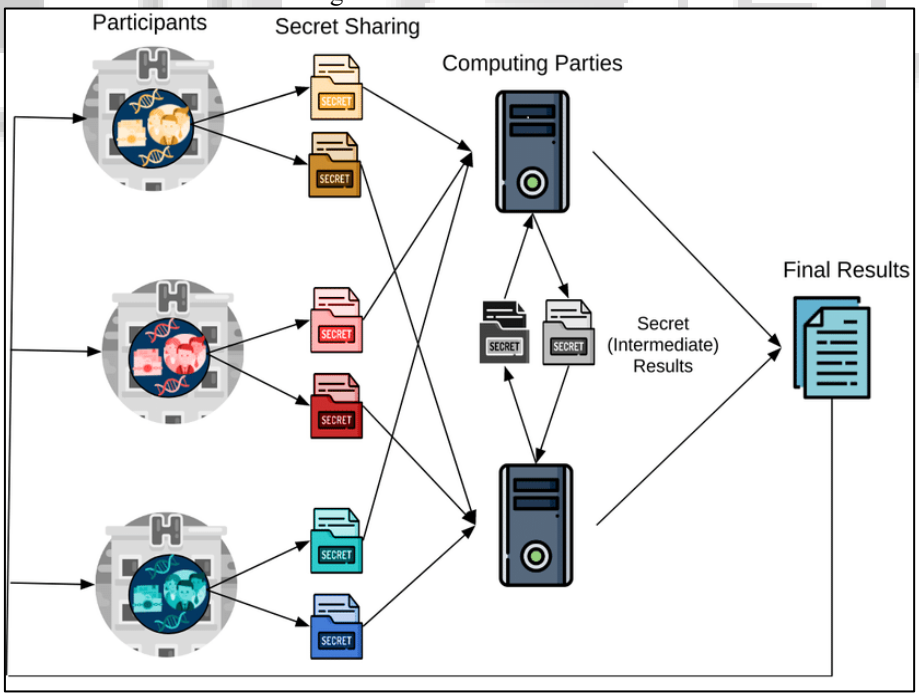

Fig. 1.C: PPDM Techniques classification framework.

The figure illustrates the classification framework of Privacy Preserving Data Mining techniques. At the top level, PPDM techniques are divided into Secret Sharing–based, Secure Multiparty Computation–based, and Perturbation–based techniques. The framework highlights how different privacy preservation mechanisms are applied depending on the data distribution and security requirements. This classification provides a clear understanding of how privacy preservation is achieved during the data mining process

1) Secure Multiparty Computation (SMC) Based Techniques

Secure Multiparty Computation allows multiple parties to collaboratively perform data mining tasks without revealing their private data to one another. Each participant contributes encrypted or transformed data, and the final mining result is obtained without exposing individual inputs.

Key characteristics:

- Suitable for distributed and collaborative data mining
- Strong privacy guarantees
- High computational and communication overhead

SMC-based techniques are widely used in horizontally and vertically partitioned databases where data owners do not fully trust each other.

2) Secret Sharing Based Techniques

In secret sharing approaches, sensitive data is divided into multiple shares and distributed among different parties. No single party can reconstruct the original data independently. Only when a sufficient number of shares are combined can the original data be recovered.

Key characteristics:

- Strong resistance to data leakage
- Ensures data confidentiality even in distributed environments
- Increased complexity in share management

Secret sharing is particularly effective in collaborative mining scenarios where multiple agents or organizations are involved.

3) Perturbation Based Techniques

Perturbation techniques preserve privacy by modifying original data values before the mining process. Noise or transformation is applied to mask sensitive information while retaining the overall statistical properties of the data.

Common perturbation methods include:

- Random noise addition
- Data distortion
- Transformation-based methods such as Piecewise Vector Quantization (PVQ)

Among these, PVQ is considered efficient because it reduces information loss while preserving similarity relationships, making it suitable for clustering applications.

Key characteristics:

- Low computational cost
- Easy to implement
- Possible information loss if noise is excessive

VIII. Comparative Analysis of PPDM Techniques

This comparison indicates that PVQ-based clustering provides an effective balance between privacy protection and data usability.

Technique	Privacy Level	Data Utility	Computational Cost
Cryptography	Very High	Low	Very High
Anonymization	Medium	Medium	Medium
Random Perturbation	Medium	Low	Low
SMC	Very High	High	Very High
PVQ-Based Clustering	High	High	Moderate

Table.1 Analysis of PPDM Techniques

IX. Problem Formulation

Privacy preservation plays a crucial role in ensuring data hiding and data security in modern information systems. Traditionally, cryptographic techniques have been widely used to protect sensitive data during storage and transmission. However, with the rapid growth of large-scale databases and data-driven applications, cryptographic methods alone are often insufficient or computationally expensive for data analysis tasks. Consequently, data mining techniques have emerged as important tools for privacy preservation.

Data mining refers to a set of automated techniques used to extract valid, implicit, potentially useful, and understandable patterns from large databases using modern computing technologies. Over the past few decades, data mining has been successfully applied in various domains such as marketing, finance, medical diagnosis, banking, manufacturing, and telecommunication. While these applications provide significant benefits, the mining of sensitive datasets using conventional data mining tools may unintentionally disclose private or confidential information.

The extracted knowledge patterns, although valuable for decision-making and strategic planning, raise serious concerns regarding individual privacy during data collection, processing, and analysis. To address these challenges, privacy-preserving data mining techniques such as association rule mining, clustering, and classification have been proposed. Additionally, adaptive noise-based data transformation methods are used to mask original data values while retaining analytical utility.

Matrix decomposition techniques play an important role in privacy-preserving data mining, particularly in classification tasks. Different forms of matrix decomposition, including horizontal, vertical, and diagonal partitioning of data, are employed to protect sensitive information. These approaches eliminate the dependency on third-party involvement while preserving data utility. Singular Value Decomposition (SVD), in particular, helps prevent the loss of mixed and extracted data during the decomposition process, thereby maintaining the balance between privacy protection and data accuracy.

X. Conclusion and Future Work

This paper presented a comprehensive review of privacy preservation techniques based on data mining approaches. The review highlighted that data mining provides a wide variety of algorithms and methodologies for preserving privacy during knowledge discovery. It was observed that many existing approaches employ hybrid techniques that combine multiple privacy-preserving mechanisms to enhance data security. Transformation-based techniques were found to be effective for privacy preservation; however, challenges related to computational complexity, noise management, and information loss still remain. These issues limit the

applicability of existing methods in large-scale and real-time data mining environments.

As a future research direction, the use of a single-point vector decomposition method is suggested for privacy preservation in data mining applications. This approach is expected to improve efficiency by reducing data loss and minimizing computational overhead. By utilizing a single selection point for data decomposition, the proposed method can enhance the security strength of privacy preservation while maintaining high data utility. Further research can focus on validating this approach across different datasets and real-world applications.

REFERENCES

[1] Wu, X., Zhu, X., Wu, G. Q., Ding, W.: Data mining with big data, IEEE Transactions on Knowledge and Data Engineering, 26(1), pp. 97–107, (2014).
[2] Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann Publishers, Elsevier, (2011).
[3] Agrawal, R., Srikant, R.: Privacy-preserving data mining, Proceedings of the ACM SIGMOD Conference, ACM Press, pp. 439–450, (2000).
[4] Kantarcioglu, M., Kardes, O.: Privacy-preserving data mining in the malicious model, International Journal of Information and Computer Security, 2(4), pp. 353–375, (2009).
[5] Sasikala, S., Banu, N.: Privacy preserving data mining using Piecewise Vector Quantization (PVQ), International Journal of Advanced Research in Computer Science and Technology, 2, pp. 302–306, (2014).
[6] Agrawal, S., Haritsa, J. R.: FRAPP: A framework for high-accuracy privacy-preserving mining, Proceedings of the 21st International Conference on Data Engineering, Springer, pp. 1–39, (2005).
[7] Kantarcioglu, M., Jiang, W.: Incentive compatible privacy-preserving data analysis, IEEE Transactions on Knowledge and Data Engineering, 25(6), pp. 1323–1335, (2013).
[8] Tassa, T.: Secure mining of association rules in horizontally distributed databases, IEEE Transactions on Knowledge and Data Engineering, 26(4), pp. 970–983, (2014).
[9] Nanavati, N. R., Jinwala, D. C.: Privacy preserving approaches for cyclic association rules in distributed databases, Proceedings of IEEE International Conference on Privacy, Security, Risk and Trust, pp. 368–371, (2011).
[10] Kumar, A.: A review on privacy preservation and collaborative data mining, International Journal of Computers & Technology, 14(12), pp. 6368–6372, (2015).
[11] A. Anchlia, "Enhancing Query Performance Through Relational Database Indexing," International Journal of Computer Trends and Technology, vol. 72, no. 8, p. 130, Aug. 2024, doi: 10.14445/22312803/ijctt-v72i8p119.
[12] Q. M. Alzubi, M. Anbar, Z. N. M. Alqattan, M. A. Al-Betar, and R. Abdullah, "Intrusion detection system based on a modified binary grey wolf optimisation," Neural Computing and Applications, vol. 32, no. 10, p. 6125, Feb. 2019, doi: 10.1007/s00521-019-04103-1.
[13] P. Beaumont and M. Huth, "Constrained Bayesian Networks: Theory, Optimization, and Applications," arXiv (Cornell University), Jan. 2017, doi: 10.48550/arxiv.1705.05326.
[14] W. Ziarko, "Variable precision rough set model," Journal of Computer and System Sciences, vol. 46, no. 1, p. 39, Feb. 1993, doi: 10.1016/0022-0000(93)90048-2.
[15] A. Mukherjee and A. K. Das, "Einstein-operations on fuzzy soft multi sets and decision making," Boletim da Sociedade Paranaense de Matemática, vol. 40, p. 1, Feb. 2022, doi: 10.5269/bspm.32546.
[16] K. A. Dhanya, S. Vajipayajula, K. Srinivasan, A. Tibrewal, S. K. Thangavel, and T. G. Kumar, "Detection of Network Attacks using Machine Learning and Deep Learning Models," Procedia Computer Science, vol. 218, p. 57, Jan. 2023, doi: 10.1016/j.procs.2022.12.401.
[17] I. S. Thaseen, B. Poorva, and P. S. Ushasree, "Network Intrusion Detection using Machine Learning Techniques," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), p. 1, Feb. 2020, doi: 10.1109/ic-etite47903.2020.148.
[18] T. Khorram, "Network Intrusion Detection using Optimized Machine Learning Algorithms," European Journal of Science and Technology, Jun. 2021, doi: 10.31590/ejosat.849723.
[19] A. M. Bamhdi, I. Abrar, and F. Masoodi, "An ensemble based approach for effective intrusion detection using majority voting," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 19, no. 2, p. 664, Feb. 2021, doi: 10.12928/telkomnika.v19i2.18325.
[20] P. M. Corea, Y. Liu, J. Wang, S. Niu, and H. Song, "Explainable AI for Comparative Analysis of Intrusion Detection Models," arXiv (Cornell University), Jun. 2024, doi: 10.48550/arxiv.2406.09684.
[21] P. Dini, A. Elhanashi, A. Begni, S. Saponara, Q. Zheng, and K. Gasmi, "Overview on Intrusion Detection Systems Design Exploiting Machine Learning for Networking Cybersecurity," Applied Sciences, vol. 13, no. 13, p. 7507, Jun. 2023, doi: 10.3390/app13137507.
[22] W. H. Aljuaid and S. S. Alshamrani, "A Deep Learning Approach for Intrusion Detection Systems in Cloud Computing Environments," Applied Sciences, vol. 14, no. 13, p. 5381, Jun. 2024, doi: 10.3390/app14135381.

