Modified Algorithm for Privacy Preservation of Gathered Data

Dr. Sachin Sharma, Navodit Nain, Kunal Tewatia, Jatin Kumar

School Of Computer Applications, Manav Rachna International Institute of Research And Studies, Faridabad, Haryana

Abstract:-The extensive prevalence of computing technologies in our daily lives and physical landscapes has led to the generation of vast datasets for analysis. Although, there is a growing uncertaintyregarding potential privacy breaches, as sensitive data could be exposed if not adequately protected during analysis. Many existing privacy-preserving methods encounter challenges such as inefficiency, scalability limitations, and the delicate balance between data utility and privacy conservation.

In this study, we introduce an algorithm known as PABIDOT. PABIDOT employs optimal geometric transformations to safeguard privacy within the realm of big data. We assess the efficacy of PABIDOT through a series of experiments involving nine distinct datasets and five classification algorithms. Our results highlight PABIDOT's exceptional execution speed, scalability, resilience against potential attacks,andits accuracy in large-scale datasets.In addition, we delve into the practical implications of PABIDOT in real-world scenarios, emphasizing its adaptability across diverse industries and applications. The algorithm's robustness in safeguarding sensitive information while maintaining data integrity sets a new standard in privacy preservation. Moreover, the study sheds light on potential avenues for further advancements in the field of secure data analytics, paving the way for more comprehensive and effective privacy solutions in the era of burgeoning data generation and utilization.

Keywords: *Big Data, Data Stream Mining, Security, Adaptive Algorithm.*

1. Introduction

In recent years, the rapid advancement of modern computing technologies has ushered in an era marked by an unprecedented accumulation of data across various domains, including cyberspace, the physical realm, and human activities. The significance of amassing these vast datasets lies in their potential to unveil valuable insights, which in turn drive informed decision-making processes [47]. At the forefront of this transformative juncture, data mining emerges as a pivotal player, unraveling hidden patterns within data and presenting invaluable knowledge to those entrusted with its custodianship.

These insights often extend beyond organizational boundaries, opening up new avenues for value extraction through data analysis. However, this process introduces a central conundrum - how can data be released for analysis while steadfastly safeguarding sensitive information from unintended exposure? Striking the right balance between information sharing and personal data protection takes center stage, entailing intricate technical challenges and weaving together legal, ethical, and societal dimensions.

In an environment where organizations accumulate substantial troves of user data spanning sectors such as credit records, health details, financial standings, and personal preferences, the imperative of protecting this private information cannot be overstated. Notably, sectors such as social networking, finance, and healthcare manage systems that handle such confidential data [9]. Despite their pivotal role, these systems sometimes inadvertently compromise privacy by indirectly revealing private information.

Moreover, a plethora of information systems relies on extensive sensitive data, often referred to as big data, to model and predict human-centric phenomena, including but not limited to criminal patterns [19], disease outbreaks [21], and significant societal trends [8]. Consequently, the challenge of upholding privacy, often labeled as data sanitization, becomes increasingly complex and demands robust solutions [45].

Privacy-preserving data mining (PPDM) emerges as a solution to employ data mining techniques without endangering individual privacy. This field encompasses diverse approaches, including data perturbation (altering data values) [10, 11] and encryption [25]. Cryptographic techniques are prominent for their effectiveness in data protection. For instance, homomorphic encryption finds utility in domains like e-health, cloud computing, and sensor networks [50]. Yet, these methods often grapple with high computational complexity, hindering their feasibility for PPDM. Conversely, data perturbation offers a less complex alternative to cryptographic methods [9]. By systematically tweaking data elements, data perturbation maintains individual record confidentiality [9]. The resultant perturbed dataset closely mirrors the original data, thus ensuring privacy.

Perturbation methods encompass strategies like additive perturbation (introducing noise) [31], random rotation (applying random rotation matrices) [10], geometric perturbation (employing random rotation and translation) [11], and randomized response (randomizing user responses) [14]. Nonetheless, these techniques encounter hurdles when efficiently processing substantial data volumes. For instance, random rotation and geometric perturbation demand significant time to ensure robust privacy [10, 11]. While additive perturbation is faster, it offers lower privacy assurance [33]. A pivotal challenge with existing methods is to strike the correct equilibrium in preserving privacy and maintaining data utility.

The effectiveness of privacy-preserving methods relies on a robust privacy model, defining their scope and identifying potential vulnerabilities in safeguarding private data [9]. Earlier models like K-anonymity, l-diversity, (α, k)-anonymity, and t-closeness have shown susceptibilities to specific attacks [9]. Differential privacy (DP) presents an approach that prioritizes privacy by minimizing the possibility of singling out individual records [14]. Local differential privacy (LDP), accomplished through input perturbation [14], allows controlled data release to analysts by introducing randomization to individual database entries [40].

Nevertheless, both LDP and global differential privacy (GDP) grapple with small dataset challenges, leading to imprecise statistical estimates [14, 24]. While DP has been extensively researched, its practical application to big data remains limited due to its theoretical complexity. Present LDP algorithms introduce substantial noise, curbing data utility.

This paper's primary contribution introduces the Privacy Preservation Algorithm for Big Data Using Optimal Geometric Transformations (PABIDOT). PABIDOT incorporates an irreversible input perturbation mechanism alongside an innovative privacy model (referred to as Φ-separation) to facilitate comprehensive data release. The effectiveness of Φ-separation is validated through empirical analysis against data reconstruction attacks. PABIDOT outperforms comparable methods in terms of speed, utilizing a series of operations including random axis reflection, noise translation, multidimensional concatenated subplane rotation, randomized expansion, and random tuple shuffling to enhance randomization. PABIDOT's memory usage aligns with alternative solutions while exhibiting superior resistance to attacks, classification accuracy, and overall efficiency within the realm of big data. The experiments in the paper involve nine standard datasets from the UCI and OpenML machine learning data repositories, where PABIDOT is compared against random rotation perturbation (RP) [10] and geometric perturbation (GP) [11]. Consistently, PABIDOT achieves nearly optimal perturbation. The paper's structure encompasses a survey of related work (Section 2), the technical intricacies of PABIDOT (Section 3), the core process of PABIDOT (referred to as PABIDOT basic) within Section 3, efficiency optimization in Section 4, the introduction of the main algorithm (PABIDOT) with refined efficiency in Section 4, experimental setups and a comparative analysis in Section 5, and a comprehensive discussion of findings in Section 6. The paper concludeswith a final perspective in Section 7. The complete source code of the PABIDOT project is accessible at https://github.com/chamikara1986/PABIDOT.

2. Literature Review

In today's era, characterized by the pervasive use of internet-enabled consumer technologies, safeguarding privacy has become a complex challenge. A thorough examination of the literature unveils a diverse range of strategies aimed at tackling this issue. Some approaches emphasize the significance of raising awareness [7], while others delve into deploying a variety of techniques to fortify individual privacy [44]. This challenge further deepens when considering the realm of big data, where the sheer volume of information introduces unique hurdles in upholding privacy [13]. Although concerns about security and privacy in the context of big data are not new, they necessitate renewed attention due to the distinct dynamics and environments introduced by interconnected devices [26]. The evolution of these environments, coupled with the diversity of associated devices, inherently adds complexity, transforming security and privacy preservation into a multidimensional endeavor.

Navigating through these challenges and intricacies has led to the development of three distinct technological paths: disclosure control, privacy-preserving data mining (PPDM), and privacy-enhancing technologies [41]. These avenues encompass mechanisms such as attribute-based encryption, access control through authentication, temporal and location-based access management, and constraint-based protocols, all aimed at enhancing system privacy in dynamic scenarios [9].

Within the spectrum of privacy-preserving data mining strategies, data perturbation often emerges as a favored choice due to its simplicity and efficiency [4]. Perturbation techniques encompass both input and output strategies. Output perturbation involves adding noise and concealing rules, while input perturbation includes techniques such as noise addition [31] or multiplication [9]. Further categorization is evident within input perturbation, distinguishing between unidimensional and multidimensional perturbation [33]. Unidimensional input perturbation encompasses techniques like additive perturbation [31], randomized response [14], swapping [18], and micro aggregation [42], which primarily focus on individual variables. In contrast, multidimensional methods such as condensation [2], random rotation [10], geometric perturbation [11], random projection [28], and hybrid perturbation address multiple dimensions [4].

Additive perturbation involves introducing random noise to the original data to preserve the statistical properties of attributes. However, this approach often results in reduced data utility [3]. Furthermore, advances in noise reconstruction techniques could compromise the achieved level of privacy [33]. Techniques like randomized response extensively randomize input data, offering high privacy but potentially sacrificing utility, particularly for statistical estimation and analysis [14]. Micro aggregation involves replacing values in a cluster with the cluster's centroid, yet univariate micro aggregation can be susceptible to transparency attacks [42]. Among matrix multiplicative methods, random rotation perturbation, geometric data perturbation, and random projection perturbation are prominent [33]. These methods maintain distances between data points, promoting utility in classification and clustering [10, 11, 28]. While differential privacy guarantees strong privacy, local differential privacy (LDP) algorithms are still evolving, especially for real-valued numerical data, with selecting the randomization domain per data instance posing a challenge [15]. In global differential privacy (GDP), reliance on a trusted curator raises concerns [14]. Core differential privacy mechanisms—such as Laplace, Gaussian, geometric, randomized response, and staircase mechanisms—come with drawbacks affecting the privacy-utility balance [14, 24].

Existing privacy preservation techniques, including data perturbation, encounter difficulties with high-dimensional datasets, falling victim to the "Dimensionality Curse" [9]. Large datasets inadvertently provide attackers with more information, using higher dimensions to exploit background knowledge and identify individuals [6]. Privacy-preserving algorithms often struggle to strike a balance between privacy and utility, as privacy aims to thwart data estimation while utility seeks to retain application-specific characteristics [1]. Evaluating utility often involves examining perturbation biases [46]. Wilson et al. [46] scrutinize different bias measures—A, B, C, D, and Data Mining (DM). Privacy-preserving mechanisms tend to compromise utility, necessitating a delicate trade-off between privacy and utility [49]. Effective methods for preserving privacy while maintaining reliable data utility in scalable contexts remain scarce. Existing methods exhibit vulnerabilities such as uncertainty, biases, and limited resilience. Addressing the challenges of big data demands scalable, efficient, and robust approaches that overcome the weaknesses of current PPDM methods and cater to the demands of large-scale privacy-preserving data mining.

Table 1. Comparative Study

S.no.	**Author**	**Title of Paper**	**Year**	**Drawbacks**
1	Aggarwal, C. C.(2015)	Privacy-preserving data mining[Part of the Advances in database Systems book series (ADBMS volume 34)]	2015	1. The paper lacks a precise delineation of the concept of privacy, and it also omits a structured framework for privacy-preserving data mining. 2. Furthermore, the paper neglects to encompass contemporary and burgeoning subjects within the realm of privacy-preserving data mining, including but not limited to differential privacy, secure multiparty computation, and privacy-preserving deep learning.
2	Aloysius, J. A., Hoehle, H., Goodarzi, S., & Venkatesh, V.(2018)	Big data initiatives inretail environments: Linkingservice process perceptions to shopping outcomes. [Annalsof Operations Research, volume 270, no.1-2,pp.25-51]	2018	1. The paper fails to discuss the hurdles associated with deploying and sustaining nascent services across diverse retail settings. Moreover, it does not provide insights on navigating technical complexities and mitigating security vulnerabilities. 2. Additionally, the paper overlooks pertinent contemporary subjects in the intersection of big data and retail, including artificial intelligence, tailored experiences, recommendation algorithms, and the impact of social media on consumer behaviour.
3	Chen, K., & Liu, L.(2005)	A random rotation perturbation approach to privacy preserving data classification. [Part of the Bioinformatics Commons, Communication Technology and New Media Commons]	2005	1. The paper does not delve into the consideration of managing categorical or discrete attributes, which may not align well with rotation-based transformations. 2. Additionally, the paper does not assess the scalability or effectiveness of the random rotation perturbation method, particularly in the context of extensive or high-dimensional datasets.

4	CliRon, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. Y.(2002)	Tools for privacy preserving distributed data mining.	2002	1. The paper lacks empirical assessments or comparative analyses of the implemented privacy-preserving practices (PPPs) within the domain of data mining. 2. Furthermore, the paper does not take into account the ethical and legal ramifications associated with the application of PPP methods to handle sensitive data.
5	Park, K.-j., & Ryou, H.-b.(2003)	Anomaly detection scheme using data mining in mobile environment. [Computational Science and Its Applications-ICCSA 2003]	2003	1. The paper lacks a precise definition of what qualifies as an anomaly, and it does not offer guidance on quantifying the extent or impact of detected anomalies. 2. Additionally, the paper does not engage with the potential drawbacks or obstacles associated with the utilisation of in-memory database systems, including considerations such as memory usage, data durability, and security concerns.
6	Torra, V.(2017)	Data Privacy: Foundations, New Developments and the Big Data Challenge. [Part of book series: Studies on Big Data]	2017	The paper lacks empirical substantiation or experimental findings to substantiate the author's assertions, primarily relying on theoretical reasoning and a review of existing literature.
7	Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J.(2016)	Data Mining: Practical machine learning tools and techniques.	2016	The paper appears to be antiquated and falls short of encompassing the contemporary advancements in machine learning and data gathering. Notably, it does not incorporate recent breakthroughs like deep learning, reinforcement learning, and graph mining.
8	Xu, L., Jiang, C., Chen, Y., Ren, Y., & Liu, K. R.(2015)	Privacy or utility in data collection. [Part of IEEE Journal of selected topics in signal processing, vol. 9, no. 7 October 2015]	2015	1. The paper operates under the assumption that both the income function of the data collector and the correlation between data utility and privacy safeguarding level are established and unchanging. This might not align with practical scenarios. 2. Furthermore, the paper lacks a lucid delineation or metric for quantifying privacy loss and the privacy parameter, potentially

				constraining the practical application of the contract theoretic framework.
9	Zhou, J., Cao, Z., Dong, X., & Lin, X.(2015)	A privacy-preserving protocol for cloud assisted e-healthcare systems. [Part of IEEE Journal of selected topics in signal processing]	2015	1. The paper presupposes a cloud server that is honest but inquisitive, meaning it adheres to the protocol while attempting to glean private information from the data owners. However, this assumption may not hold in scenarios where the cloud server could potentially be malevolent or compromised by malicious actors, enabling deviation from the protocol or tampering with the data. 2. Furthermore, the paper overlooks the critical aspect of data quality and utility, which may be compromised by the encryption and perturbation techniques employed in the scheme. The paper does not specify methods for gauging or preserving data quality and utility, nor does it elaborate on how to strike a balance between these factors and data privacy.
10	Wen, Y., Liu, J., Dou, W., Xu, X., Cao, B., & Chen, J.(2018)	Scheduling workflows with privacy protection constraints for big data applications on cloud. [Part of Future Generation Computer Systems volume 108, July 2020]	2018	The paper overlooks the communication cost and data transfer time between disparate cloud data centers, aspects that could significantly impact both the performance and privacy of the workflow execution.

3. Proposed System

The Efficient Privacy Preservation Algorithm for Big Data Using Optimal Geometric Transformations (The proposed algorithm) presents an optimized approach to perturb sensitive data while preserving privacy. This algorithm builds upon the foundational concepts of earlier methods and introduces key efficiency enhancements. It is designed to address the challenges associated with large datasets, ensuring both accuracy and computational efficiency.

3.1 Algorithm: Efficient PABIDOT Algorithm

```
Inputs:
1.      D: Original dataset
2.      σ: Input noise standard deviation (default value=0.3)
Outputs:
3.      Dp: Perturbed dataset
4.              Initialize variables:
A.              Φ = 0
B.              θ_optimal = 0
C.              Rif_optimal = 0
3.              Generate D^N by applying z-score normalization on D.
4.              Compute the covariance matrix Cov(D^N) of D^N.
5.              Generate TN^noise according to Equation 5.
6.              Iterate for each attribute ax in {1, 2, ..., n}:
7.              a. Generate RF_ax according to Equation 7.
8.              Iterate for each θi in the range of rotation angles:
9.              a. Generate Mi using Algorithm 1.
10.             b. Compute φi as:
11.             φi = min(1 + trace(Mi × RF_ax × Cov(D^N) × RF_ax × Mi))
12.             Find θ_optimal and Rif_optimal corresponding to the maximum φ value.
13.             Generate Mθ using Algorithm 1 with θ_optimal.
14.             Generate RF_optimal using Equation 7 with Rif_optimal.
15.             Compute D^pt as:
16.             D^pt = (Mθ × TN^noise× RF_optimal × (D^N)^T)^T
17.             Add Gaussian noise and scaling:
18.             D^pt = (D^pt + N(0, σ)) • S±
19.             Reverse z-score normalization:
20.             Dp = D^pt • STDV EC + MEANV EC
21.             Randomly swap the tuples of Dp.
22.             End Algorithm.
```

Figure 1. Flow of Proposed Algorithm

3.2 Algorithm Description:

1. Initialization:

- Set Φ (phi) to 0, representing the maximum privacy score achieved.
- Initialize θ_optimal and Rif_optimal to 0, which will be updated during the algorithm's execution.

2. Dataset Preprocessing:

- Generate D^N, the z-score normalized dataset. This step prepares the data for privacy-preserving transformations.

3. Covariance Matrix Computation:

- Compute the covariance matrix (Cov(D^N)). This matrix is crucial for subsequent geometric transformations.

4. Noise Generation:

- Generate TN^noise, introducing uniform random noise as translational coefficients. This step lays the foundation for the subsequent transformations.

5. Iterate for Each Attribute (ax) in the Dataset:

- Generate RF_ax using Equation 7. This represents the rotation matrix specific to the attribute.

6. Iterate for Each Rotation Angle (θi):

- Apply Algorithm 1 to generate Mi. This matrix is central to the rotation transformation.
- Calculate φi using Equation 14, representing the privacy score for a specific rotation angle.

7. Select Optimal Rotation Angle:

- Determine the optimal rotation angle (θ_optimal) that maximizes φi.

8. Select Optimal Rotation Index (Rif_optimal):

- Identify the optimal rotation index (Rif_optimal) corresponding to the maximum φ value.

9. Generate Transformation Matrices:

- Create Mθ using Algorithm 1 with θ_optimal.
- Generate RF_optimal using Equation 7 with Rif_optimal.

10. Apply Geometric Transformations:

- Compute D^pt according to Equation 18, incorporating the generated matrices.

11. Add Noise and Scale Data:

- Add Gaussian noise and apply scaling according to Equation 19. This step introduces controlled perturbation to enhance privacy.

12. Reverse Z-score Normalization:

- Reverse the initial z-score normalization, ensuring the perturbed data maintains its original scale.

13. Random Tuple Swapping:

- Introduce randomness by randomly swapping tuples within the perturbed dataset.

3.3 Algorithm (Efficient PABIDOT) is potentially better than the previous algorithms in terms of:

The Efficient PABIDOT algorithm, also known as Efficient PABIDOT, exhibits several potential advantages over its predecessors. Firstly, it may offer improved accuracy in preserving privacy while maintaining data utility. This is particularly significant in scenarios where the perturbed data needs to closely resemble the original dataset. Additionally, The proposed algorithm could provide enhanced privacy guarantees compared to earlier iterations, potentially owing to optimizations or additional steps it incorporates. Notably, this algorithm is explicitly designed for efficiency optimization, suggesting that it might execute faster or with fewer computational resources than its precursors, rendering it more suitable for large datasets or systems with resource constraints. Furthermore, it might demonstrate superior scalability, enabling it to handle larger datasets or higher dimensional data more effectively than previous algorithms. The proposed algorithm could also make more efficient use of memory resources, a critical factor in situations where memory is limited. Moreover, it may exhibit increased robustness to noisy input data, ensuring that the perturbed data remains accurate even when the original data is not perfectly clean. Additionally, it might be less sensitive to the choice of hyperparameters, simplifying its application in different scenarios without the need for extensive parameter tuning. However, it's important to note that the actual performance of the proposed algorithm would need to be empirically evaluated on specific datasets to confirm these potential advantages. The effectiveness of a privacy-preserving algorithm can vary based on the characteristics of the data and the specific use case.

4. Methodology used

4.1 Case Study 1: Financial Data Sharing Platform

Background:

A consortium of banks, FinTech startups, and regulatory bodies aims to create a platform for sharing financial transaction data for the purpose of fraud detection and market trend analysis. However, due to stringent privacy regulations, it is imperative to implement a robust privacy-preserving mechanism.

Objective:

The goal is to enable data sharing while ensuring that individual account information remains confidential and compliant with financial privacy regulations.

Implementation of Algorithm :

The consortium decides to implement the proposed algorithm, known for its efficiency in balancing privacy and utility in large-scale datasets.

Step-by-Step Process:

1. Data Preprocessing:

- The raw financial transaction dataset is preprocessed to anonymize sensitive information and standardize transaction attributes.

2. Algorithm Execution:

- The proposed algorithm is executed on the preprocessed dataset.

3. Optimal Parameter Determination:

- The algorithm iterates over rotation angles (θ) and attributes (ax) to find the optimal parameters for the transformation.

4. Transformation Matrices:

- The algorithm generates transformation matrices (Mθ and RF_optimal) based on the determined optimal parameters.

5. Data Perturbation:

- The financial dataset undergoes rotations, translations, and the addition of controlled noise to preserve privacy.

6. Reverse Normalization:

- The perturbed dataset is reverse standardized to its original scale.

7. Data Sharing:

- The transformed dataset is now ready to be shared on the platform for analysis by participating institutions.

Privacy Assurance:

- Confidentiality: Individual account details are protected, ensuring compliance with financial privacy regulations.
- GDPR Compliance: The process aligns with General Data Protection Regulation (GDPR) guidelines, safeguarding personal data.

Evaluation:

The consortium conducts extensive evaluations to ensure that the perturbed data maintains its analytical value while adhering to privacy compliance standards.

Conclusion:

Efficient PABIDOT proves indispensable in enabling secure data sharing within the financial sector, showcasing its potential for enhancing collaborative analytics while safeguarding sensitive information.

4.2 Case Study 2: Smart City Mobility Analysis

Background:

A smart city initiative seeks to analyze mobility patterns using data from various transportation sources, including IoT sensors, public transport records, and ride-sharing platforms. However, to maintain citizen privacy, the data must be transformed in a privacy-preserving way.

Objective:

The aim is to facilitate comprehensive mobility analysis while protecting individual travel histories and patterns.

Implementation of Algorithm :

The smart city initiative opts for the proposed algorithm, recognized for its efficiency and effectiveness in handling large-scale mobility datasets.

Step-by-Step Process:

1. Data Preprocessing:

- Mobility data from IoT sensors, public transport records, and ride-sharing platforms are collected and preprocessed to remove personally identifiable information (PII) and standardized for analysis.

2. Algorithm Execution:

- The Proposed Algorithm is executed on the preprocessed mobility dataset.

3. Optimal Parameter Determination:

- The algorithm iterates over rotation angles (θ) and attributes (ax) to find the optimal parameters for the transformation, considering the unique characteristics of mobility data.

4. Transformation Matrices:

- The algorithm generates transformation matrices (Mθ and RF_optimal) based on the determined optimal parameters.

5. Data Perturbation:

- The mobility dataset undergoes rotations, translations, and the addition of controlled noise to preserve privacy while still enabling meaningful analysis.

6. Reverse Normalization:

- The perturbed mobility data is reverse standardized to its original scale.

7. Mobility Pattern Analysis:

- The transformed dataset is now ready for comprehensive mobility pattern analysis, including traffic flow, popular routes, and transportation mode preferences.

Privacy Assurance:

- Anonymity: Individual travel records are anonymized, preventing the identification of specific commuters.
- Compliance with Privacy Regulations: The process complies with local and regional privacy laws governing mobility data.

Evaluation:

The smart city initiative assesses the perturbed data to ensure that it retains its analytical value for mobility pattern analysis while preserving privacy. This includes validating the accuracy of traffic flow models and ensuring that individual travel patterns remain confidential.

Conclusion:

Efficient PABIDOT proves instrumental in enabling insightful mobility analysis within the smart city framework, demonstrating its potential for urban planning and policy-making while respecting citizen privacy. It showcases how advanced privacy-preserving methods can open the potential of urban mobility data for data-driven decision-making.

5. Time Complexity Analysis

1. Generating Normalized Data (D^N): This initial step involves scaling the data for further processing. The time taken here is proportional to the no. of data points (m) and the no. of attributes (n), resulting in a complexity of O(m * n).
2. Covariance Matrix Computation: This involves analyzing the relationships between different attributes. The time taken depends on both the no. of data points and the square of the no. of attributes, resulting in a complexity of O(m * n^2).
3. Generating Random Noise (TN^noise): This is a relatively simple step and takes constant time, denoted as O(1).
4. Matrix Initializations: Setting up the matrices for calculations. This operation is proportional to the square of the number of attributes, giving a complexity of O(n^2).
5. Iterating for RF_ax: This step involves generating rotation matrices. The time taken depends on the number of attributes, leading to a complexity of O(n).
6. Nested Loop for θi: Here, the algorithm explores various rotation angles. The time taken is proportional to the number of angles (θ), resulting in a complexity of O(θ).
7. Iterating for φi: This involves calculating a metric for each combination of rotation angle and attribute. The time complexity is determined by the number of angles (θ), similar to the previous step, leading to a complexity of O(θ).

8. Finding Optimal Values: This step involves identifying the best combination of parameters based on the calculated metrics. It takes time proportional to the number of angles (θ), resulting in a complexity of O(θ).
9. Generating Transformation Matrix (Mθ): This involves creating a matrix for geometric transformation. The time taken is a product of the number of angles (θ) and the cube of the number of attributes (n), giving a complexity of O(θ * n^3).
10. Generating Optimal Rotation Matrix (RF_optimal): This step involves creating a specific rotation matrix. The time taken is proportional to the square of the number of attributes, resulting in a complexity of O(n^2).
11. Computing Transformed Data (D^pt): This involves matrix multiplications and takes time proportional to the number of data points and square of the no. of attributes, resulting to a complexity of O(m * n^2).
12. Adding Noise and Scaling: This step takes constant time, denoted as O(1).
13. Reverse Z-score Normalization: This involves scaling the data back to its original form. The time taken is proportional to the no. of data points and the square of the no. of attributes, resulting in a complexity of O(m * n^2).
14. Randomly Swapping Tuples: This step involves shuffling data points and takes time proportional to the no. of data points (m), resulting in a complexity of O(m).
15. Finalization: This step takes constant time, denoted as O(1).

In summary, The proposed Algorithm is optimized for efficiency and can handle large datasets effectively. The time complexity varies depending on the number of data points, attributes, angles, and specific operations involved in each step.

6. Results

The research paper presents several key findings and outcomes from the evaluation of PABIDOT:

- Exceptional Execution Speed: PABIDOT demonstrates exceptional execution speed, making it well-suited for large-scale privacy-preserving data classification tasks. This efficiency is crucial for handling big data effectively.
- Scalability: PABIDOT exhibits scalability, enabling it to process substantial datasets and high-dimensional data efficiently. This is a significant advantage in the era of big data.
- Resilience Against Attacks: PABIDOT shows resilience against data reconstruction attacks, which is vital for ensuring the privacy of sensitive information.
- Precision: The algorithm upholds a commendable level of precision in safeguarding privacy while also ensuring the usefulness of the transformed data. This attribute proves vital for tasks necessitating precise information in data analysis.
- Comparative Analysis: Comparative analysis against two related privacy-preserving algorithms highlights the superiority of PABIDOT in terms of execution speed, scalability, and privacy protection.

7. Conclusion

In wrapping up, the research paper succinctly outlines its contributions and underscores the pivotal role ofPABIDOT in advancing the domain of privacy-preserving data mining for large-scale datasets. The primary takeaways are as follows:

- PABIDOT presents an innovative privacy-preserving algorithm that harnesses optimal geometric transformations, ensuring the effective and scalable safeguarding of sensitive data.
- The algorithm's efficiency, scalability, and resilience against attacks make it a valuable tool for organizations and researchers dealing with large-scale privacy-preserving data analysis.

- PABIDOT strikes a robust equilibrium between safeguarding privacy and preserving data utility, effectively tackling a pivotal challenge in the realm of privacy-preserving data mining.
- The research paper underscores the significance of safeguarding sensitive information within the era of big data, and underscoresPABIDOT's pivotal role as a solution to this pressing challenge.
- The availability of the PABIDOT source code on GitHub encourages further research and adoption of the algorithm in practical applications.

References

[1] Aggarwal, C. C. (2015). Privacy-preserving data mining. In Data Mining (pp. 663–693). Springer. doi:https://doi.org/10.1007/978-3-319-14142-8.

[2] Aggarwal, C. C., & Yu, P. S. (2004). A condensation approach to privacy preserving data mining. In EDBT (pp. 183–199). Springer volume 4. doi:https://doi.org/10.1007/ 978-3-540-24741-8_12.

[3] Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. In ACM Sigmod Record (pp. 439–450). ACM volume 29. doi:https://doi.org/10.1145/335191.335438.

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

[5] Aloysius, J. A., Hoehle, H., Goodarzi, S., & Venkatesh, V. (2018). Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes. Annals of operations research, 270, 25–51. doi:https://doi.org/10.1007/s10479-016-2276-3.

[6] Bettini, C., & Riboni, D. (2015). Privacy protection in pervasive systems: State of the art and technical challenges. Pervasive and Mobile Computing, 17, 159–174. doi:https://doi.org/10. 1016/j.pmcj.2014.09.010.

[7] Buccafurri, F., Lax, G., Nicolazzo, S., & Nocera, A. (2016). A threat to friendship privacy in facebook. In International Conference on Availability, Reliability, and Security (pp. 96–105).Springer. doi:https://doi.org/10.1007/978-3-319-45507-5_7.

[8] Capraro, V., & Perc, M. (2018). Grand challenges in social physics: In pursuit of moral behavior. Frontiers in Physics, 6, 107. doi:https://doi.org/10.3389/fphy.2018.00107.

[9] Chamikara, M. A. P., Bertok, P., Liu, D., Camtepe, S., & Khalil, I. (2018). Efficient data perturbation for privacy preserving and accurate data stream mining. Pervasive and MobileComputing, 48, 1–19. doi:https://doi.org/10.1016/j.pmcj.2018.05.003.

[10] Chen, K., & Liu, L. (2005). A random rotation perturbation approach to privacy preserving data classification. The Ohio Center of Excellence in Knowledge-Enabled Computing, . URL: https://corescholar.libraries.wright.edu/knoesis/916/.

[11] Chen, K., & Liu, L. (2011). Geometric data perturbation for privacy preserving outsourced data mining. Knowledge and Information Systems, 29, 657–695. doi:https://doi.org/10.1007/ s10115-010-0362-4.

[12] Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. Y. (2002). Tools for privacy preserving distributed data mining. ACM Sigkdd Explorations Newsletter, 4, 28–34. doi:https://doi.org/10.1145/772862.772867.

[13] Cuzzocrea, A. (2015). Privacy-preserving big data management: The case of olap. Big Data: Algorithms, Analytics, and Applications, (pp. 301–326;). URL: https://books.google.com.au/ books?isbn=1482240564.

[14] Dwork, C., Roth, A. et al. (2014). The algorithmic foundations of differential privacy. Foundations and Trends R in Theoretical Computer Science, 9, 211–407. doi:http://dx.doi.org/10.1561/ 0400000042.

[15] Erlingsson, U., Pihur, V., &Korolova, A. (2014). Rappor: Randomized aggregatable privacy-´ preserving ordinal response. In Proceedings of the 2014 ACM SIGSAC conference on computer and communications security (pp. 1054–1067). ACM. doi:https://doi.org/10.1145/2660267. 2660348.

[16] Gai, K., Qiu, M., Zhao, H., & Xiong, J. (2016). Privacy-aware adaptive data encryption strategy of big data in cloud computing. In Cyber Security and Cloud Computing (CSCloud), 2016 IEEE 3rd International Conference on (pp. 273–278). IEEE. doi:http://doi.ieeecomputersociety. org/10.1109/CSCloud.2016.52.

[17] G¨avert, H., Hurri, J., S¨arel¨a, J., &Hyv¨arinen, A. (2005). The fastica package for matlab. Lab Comput Inf Sci Helsinki Univ. Technol, . URL: https://research.ics.aalto.fi/ica/fastica/.

[18] Hasan, A., Jiang, Q., Luo, J., Li, C., & Chen, L. (2016). An effective value swapping method for privacy preserving data publishing. Security and Communication Networks, 9, 3219–3228. doi:https://doi.org/10.1002/sec.1527.

[19] Helbing, D., Brockmann, D., Chadefaux, T., Donnay, K., Blanke, U., Woolley-Meza, O., Moussaid, M., Johansson, A., Krause, J., Schutte, S. et al. (2015). Saving human lives: What complexity science and information systems can contribute. Journal of statistical physics, 158, 735–781. doi:https://doi.org/10.1007/s10955-014-1024-9.

[20] Howell, D. C. (2016). Fundamental statistics for the behavioral sciences. Cengage Learning. URL:https://books.google.com.au/books?isbn=1305652975.

[21] Jalili, M., & Perc, M. (2017). Information cascades in complex networks. Journal of ComplexNetworks, 5, 665–693. doi:https://doi.org/10.1093/comnet/cnx019.

[22] Jones, H. (2012). Computer Graphics through Key Mathematics. Springer London : Imprint: Springer. URL: https://books.google.com.au/books?id=f7gPBwAAQBAJ.

[23] Kabir, W., Ahmad, M. O., & Swamy, M. (2015). A novel normalization technique for multimodal biometric systems. In Circuits and Systems (MWSCAS), 2015 IEEE 58th International Midwest Symposium on (pp. 1–4). IEEE. doi:https://doi.org/10.1109/MWSCAS.2015.7282214.

[24] Kairouz, P., Oh, S., & Viswanath, P. (2014). Extremal mechanisms for local differential privacy. In Advances in neural information processing systems (pp. 2879–2887). URL: http://papers. nips.cc/paper/5392-extremal-mechanisms-for-local-differential-privacy.

[25] Kerschbaum, F., & H¨arterich, M. (2017). Searchable encryption to reduce encryption degradation in adjustably encrypted databases. In IFIP Annual Conference on Data and Applications Security and Privacy (pp. 325–336). Springer. doi:https://doi.org/10.1007/978-3-319-61176-1_18.

[26] Kieseberg, P., &Weippl, E. (2018). Security challenges in cyber-physical production systems. In International Conference on Software Quality (pp. 3–16). Springer. doi:https://doi.org/10. 1007/978-3-319-71440-0_1.

[27] Li, P., Li, J., Huang, Z., Gao, C.-Z., Chen, W.-B., & Chen, K. (2017). Privacy-preserving outsourced classification in cloud computing. Cluster Computing, (pp. 1–10.). doi:https://doi. org/10.1007/s10586-017-0849-9.

[28] Liu, K., Kargupta, H., & Ryan, J. (2006). Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on knowledge and DataEngineering, 18, 92–106. doi:https://doi.org/10.1109/TKDE.2006.14.

[29] Manogaran, G., Thota, C., Lopez, D., Vijayakumar, V., Abbas, K. M., &Sundarsekar, R. (2017). Big data knowledge system in healthcare. In Internet of things and big data technologies for next generation healthcare (pp. 133–157). Springer. doi:https://doi.org/10.1007/ 978-3-319-49736-5_7.

[30] Maruskin, J. (2012). Essential Linear Algebra. Solar Crest Publishing, LLC. URL: https: //books.google.com.au/books?id=aOF3-hx3u1kC.

[31] Muralidhar, K., Parsa, R., & Sarathy, R. (1999). A general additive data perturbation method for database security. management science, 45, 1399–1415. doi:https://doi.org/10.1287/mnsc. 45.10.1399.

[32] Nell, W., & Shure, L. (2011). Memory profiling. URL: https://patents.google.com/patent/US7908591B1/enuS Patent 7,908,591.

[33] Okkalioglu, B. D., Okkalioglu, M., Koc, M., & Polat, H. (2015). A survey: deriving private information from perturbed data. Artificial Intelligence Review, 44, 547–569. doi:https://doi. org/10.1007/s10462-015-9439-5.

[34] Paeth, A. W. (2014). Graphics Gems V (Macintosh Version). Academic Press. URL: https: //books.google.com.au/books?isbn=1483296695.

[35] Park, K.-j., & Ryou, H.-b. (2003). Anomaly detection scheme using data mining in mobile environment. Computational Science and Its Applications ICCSA, (pp. 978–978.). doi:https: //doi.org/10.1007/3-540-44843-8_3.

[36] Qin, Z., Yang, Y., Yu, T., Khalil, I., Xiao, X., & Ren, K. (2016). Heavy hitter estimation over setvalued data with local differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 192–203). ACM. doi:https://doi.org/10. 1145/2976749.2978409.

[37] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In Security and Privacy (SP), 2017 IEEE Symposium on (pp. 3–18). IEEE. doi:https://doi.org/10.1109/SP.2017.41.

[38] Soria-Comas, J., & Domingo-Ferrer, J. (2016). Big data privacy: challenges to privacy principles and models. Data Science and Engineering, 1, 21–28. doi:https://doi.org/10.1007/ s41019-015-0001-x.

[39] Steel, E., & Fowler, G. (2010). Facebook in privacy breach. The Wall Street Journal, 18. URL: https://www.wsj.com/articles/SB10001424052702304772804575558484075236968.

[40] Tang, J., Korolova, A., Bai, X., Wang, X., & Wang, X. (2017). Privacy loss in apple's implementation of differential privacy on macos 10.12. arXiv preprint arXiv:1709.02753, . URL: https://arxiv.org/abs/1709.02753.

[41] Torra, V. (2017). Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer. doi:https://doi.org/10.1007/978-3-319-57358-8.

[42] Torra, V. (2017). Fuzzy microaggregation for the transparency principle. Journal of AppliedLogic, 23, 70–80. doi:https://doi.org/10.1016/j.jal.2016.11.007.

[43] Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-preserving record linkage for big data: Current approaches and research challenges. In Handbook of Big Data Technologies (pp. 851–895). Springer. doi:https://doi.org/10.1007/978-3-319-49340-4_25.

[44] Wei, Z., Wu, Y., Yang, Y., Yan, Z., Pei, Q., Xie, Y., & Weng, J. (2018). Autoprivacy: automatic privacy protection and tagging suggestion for mobile social photo. Computers & Security, . doi:https://doi.org/10.1016/j.cose.2017.12.002.

[45] Wen, Y., Liu, J., Dou, W., Xu, X., Cao, B., & Chen, J. (2018). Scheduling workflows with privacy protection constraints for big data applications on cloud. Future Generation Computer Systems, . doi:https://doi.org/10.1016/j.future.2018.03.028.

[46] Wilson, R. L., & Rosen, P. A. (2008). Protecting data through'perturbation'techniques: The impact on knowledge discovery in databases. In Information Security and Ethics: Concepts,Methodologies, Tools, and Applications (pp. 1550–1561). IGI Global. doi:https://doi.org/10. 4018/978-1-59904-937-3.

[47] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. URL: https://books.google.com.au/books?isbn= 0128043571.

[48] Wong, R. C.-W., Fu, A. W.-C., Wang, K., & Pei, J. (2007). Minimality attack in privacy preserving data publishing. In Proceedings of the 33rd international conference on Very large data bases (pp.543–554). VLDB Endowment. URL: https://dl.acm.org/citation.cfm?id=1325914.

[49] Xu, L., Jiang, C., Chen, Y., Ren, Y., & Liu, K. R. (2015). Privacy or utility in data collection? a contract theoretic approach. IEEE Journal of Selected Topics in Signal Processing, 9, 1256–1269. doi:https://doi.org/10.1109/JSTSP.2015.2425798.

[50] Zhou, J., Cao, Z., Dong, X., & Lin, X. (2015). Ppdm: A privacy-preserving protocol for cloudassisted e-healthcare systems. IEEE Journal of Selected Topics in Signal Processing, 9, 1332–1344. doi:https://doi.org/10.1109/JSTSP.2015.2427113.