PRIVACY-PRESERVING SENSITIVE INFORMATION MASKING SCHEME

Capstone Project Proposal

Submitted by:

(102116050) VINAYAK LAL

(102166003) S NITISH KUMAR

(102116051) KESHAV MITTAL

(102116055) DURGA HARSHA VARDHAN TILLAPUDI

BE Third Year-

CPG No. 192

Under the Mentorship of

DR.ROHAN SHARMA

Assistant Professor

DR.SHIVANI SHARMA

Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

MARCH 2024

TABLE OF CONTENTS

Mentor Consent Form

I hereby agree to be the mentor of the following Capstone Project Team

Project Title: Privacy-Preserving Sensitive Information Masking Scheme		
Roll No	**Name**	**Signatures**
102116050	Vinayak Lal	
102166003	S Nitish Kumar	
102116051	Keshav Mittal	
102116055	Durga Harsha Vardhan Tillapudi	

NAME of Mentor: … Dr. Rohan Sharma …... . . . . .

SIGNATURE of Mentor: . . . . . .

NAME of Co-Mentor(if any): …Dr. Shivani Sharma…... . . . . .

SIGNATURE of Co-Mentor: . . . . . .

Project Overview

With the rapid growth of information technology and e-commerce applications, it has become increasingly easy to discover useful information and interesting relationships in huge amount of data. Data mining, also called knowledge discovery in database (KDD), provides a set of techniques commonly used to analyze relationships among purchased products for market basket analysis and also in the healthcare industry, which is producing a large amount of data that is highly sensitive. We are required to analyze and understand the associations and patterns in the data for the drug, vaccine, and critical symptom analysis. Knowledge discovered using KDD techniques can be generally classified as association rules.

As data mining techniques can be used to discover implicit information in very large databases, private or secure information, such as credit card numbers, personal identification numbers, telephone numbers, and other confidential data, may also be easily revealed by those techniques. Similarly, sharing medical data with any third party may cause serious privacy threats. Another important issue is that information shared among business collaborators may also be analyzed using data mining techniques to reveal sensitive knowledge that may be leaked to competitors.

So, Privacy-preserving data mining (PPDM) has become an important research field in recent years, as approaches for PPDM can discover important information in databases while ensuring that sensitive information is not revealed. Several algorithms have been proposed to hide sensitive information in databases. They apply addition and deletion operations to perturb an original database and hide the sensitive information. However, deleting information causes utility reduction. Optimizing utility and finding an appropriate set of transactions/itemsets to be perturbed for hiding sensitive information while preserving other important information is an NP-hard problem. Therefore, we must develop and design a model that can transform data so that no confidential information gets revealed but must retain high utility.In this project, we shall develop new model to transform the dataset and will be using the basics of various optimization techniques. Particle Swarm Optimization (PSO) is a nature-inspired optimization technique that mimics the collective behavior of swarms, such as bird flocking and fish schooling. PSO has been widely utilized in various optimization problems due to its simplicity, efficiency, and ability to converge to optimal solutions but has drawbacks too.

In summary, by combining robust algorithm design with the optimization power of various sanitization algorithms, this project aims to develop privacy-preserving masking scheme that effectively safeguard sensitive information while maintaining data utility and confidentiality without deleting data and still maintaining high utility.

Problem Statement

Amid the growing concerns over privacy and security in the era of big data, there's a pressing need for innovative methods to protect sensitive information while preserving data utility. Traditional deletion practices in privacy-preserving data mining (PPDM) often lead to unintended data loss and reduced utility, prompting the search for alternative techniques. Our project is dedicated to addressing this challenge by introducing a novel approach that seamlessly combines advanced data aggregation methods with cutting-edge evolutionary optimization techniques, particularly focusing on Particle Swarm Optimization (PSO).

Through skillful concealment of sensitive patterns and meticulous minimization of data loss and artificial pattern generation, our meticulously crafted solution aims to navigate the intricate landscape of privacy preservation and data utility. By achieving an optimal balance between the two, our framework provides a robust foundation for secure data transformation and fosters smooth collaboration with third-party entities. In every aspect of its design and execution, our project prioritizes professionalism and conciseness, underscoring our commitment to excellence in the realm of data privacy and security.

Need Analysis

In today's digital age, the proliferation of data collection and analysis has led to unprecedented advancements in various fields, ranging from healthcare and finance to e-commerce and social media. However, this abundance of data also brings forth significant concerns regarding privacy and security. As organizations and individuals increasingly rely on data-driven technologies, the need to protect sensitive information from unauthorized access and misuse has become more critical than ever before. Therefore, a thorough need analysis is essential to understand the pressing challenges and requirements that drive the development of privacy-preserving sensitive information masking schemes.

1. Protection of Individual Privacy:

The primary motivation behind the need for privacy-preserving masking schemes lies in safeguarding individual privacy. With the digitization of personal information, individuals are increasingly vulnerable to privacy breaches and identity theft. For instance, healthcare records, financial transactions, and online behavior patterns contain sensitive information that, if exposed, can lead to severe consequences for individuals. Therefore, there is a growing demand for robust mechanisms to anonymize and protect personal data while enabling legitimate data analysis and utilization.

2. Regulatory Compliance and Legal Obligations:

Governments and regulatory bodies worldwide have introduced stringent data protection regulations to ensure the privacy and security of personal information. Regulations such as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA) impose legal obligations on organizations to implement measures for safeguarding sensitive information failing to which can result in severe penalties and reputational damage. Hence, there is a compelling need for organizations to adopt privacy-preserving techniques to ensure compliance with regulatory requirements.

3. Facilitating Data Sharing and Collaboration:

In today's interconnected world, data sharing and collaboration are essential for advancing research, innovation, and decision-making processes. However, sharing sensitive information poses inherent risks, as it may compromise individual privacy and confidentiality. Privacy-preserving masking schemes solve this challenge by enabling data sharing while protecting sensitive attributes. By anonymizing sensitive information, organizations can collaborate securely without exposing individuals to privacy risks.

Literature Survey

1. A sanitization approach for hiding sensitive itemsets based on particle swarm optimization (PSO)

- **Authors:** Jerry Chun-Wei Lin, Qiankun Liu, Philippe Fournier-Viger, Tzung-Pei Hong, Miroslav Voznak, Justin Zhan

- **Summary:** This research paper proposes a novel sanitization approach that utilizes PSO for hiding sensitive itemsets in datasets. The approach aims to minimize the distortion introduced to non-sensitive itemsets while effectively concealing sensitive ones. By formulating the problem as an optimization task and leveraging the collective behavior of swarms, the proposed approach demonstrates promising results in preserving data utility while ensuring privacy.

- **Key Contributions:**

  - Formulating a sanitization approach for hiding sensitive itemsets using PSO.
  - Experimental evaluation demonstrating the effectiveness of the proposed approach in preserving data utility and privacy.

2. VIDPSO: Victim item deletion based PSO inspired sensitive pattern hiding algorithm for dense datasets

- **Authors:** Shalini Jangra, Durga Toshniwal

- **Summary:** This paper presents a sensitive pattern hiding algorithm inspired by PSO, specifically designed for dense datasets. The algorithm, called VIDPSO, focuses on identifying and deleting victim items to conceal sensitive patterns effectively. By incorporating PSO's optimization capabilities, VIDPSO achieves significant improvements in terms of privacy preservation and data utility compared to existing techniques.

- **Key Contributions:**

  - Development of a sensitive pattern-hiding algorithm tailored for dense datasets.
  - Utilization of PSO-inspired optimization techniques to enhance privacy preservation.

3. Privacy-preserving data mining using differential privacy

- **Authors**: Cynthia Dwork
- **Summary:** This seminal paper introduces the concept of differential privacy as a framework for privacy-preserving data mining. Differential privacy aims to ensure that the presence or absence of an individual's data does not significantly affect the outcome of data analysis. By adding carefully calibrated noise to query responses, differential privacy offers strong privacy guarantees while allowing for meaningful data analysis.
- **Key Contributions:**

- Introduction of the differential privacy framework for privacy-preserving data mining.
- Discussion on mechanisms for achieving differential privacy, such as Laplace noise addition and randomized response.

4. Privacy-preserving data publishing: A survey

- **Authors:** Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu
- **Summary:** This survey paper provides a comprehensive overview of privacy-preserving data publishing techniques. It covers a wide range of approaches, including k-anonymity, l-diversity, t-closeness, and differential privacy. The survey discusses the strengths and limitations of each technique and provides insights into their real-world applicability across various domains.
- **Key Contributions:**

- Overview of privacy-preserving data publishing techniques.
- Comparative analysis of different approaches based on their privacy guarantees and computational complexity.

5. Privacy-preserving collaborative filtering using randomized perturbation-based approach

- **Authors**: Yee Wei Law, Ninghui Li, and David W. Cheung
- **Summary:** This research paper proposes a randomized perturbation-based approach for privacy-preserving collaborative filtering. The approach introduces random noise to user-item ratings to protect individual privacy while maintaining the accuracy of recommendation systems. By incorporating perturbation techniques, the proposed approach offers robust privacy guarantees without compromising recommendation quality.
- **Key Contributions:**

- Development of a privacy-preserving collaborative filtering approach using randomized perturbation.
- Evaluation of the approach's effectiveness in preserving privacy and recommendation accuracy through experimental validation.

6. Privacy-preserving machine learning: Threats and solutions

- **Authors:** Al-Rubaie, Mohammad, and J. Morris Chang
- **Summary:** This paper provides an overview of privacy threats in machine learning models and proposes solutions for mitigating these threats. It discusses various attacks on machine learning models, such as membership inference and model inversion attacks, and presents privacy-preserving techniques, including differential privacy, federated learning, and secure multi-party computation.
- **Key Contributions:**

- Identification of privacy threats in machine learning models.
- Presentation of privacy-preserving solutions to mitigate these threats and protect sensitive information.

This literature survey provides insights into existing systems, products, and projects in the field of privacy-preserving sensitive information masking field. Each paper contributes valuable insights and methodologies that can inform the development of novel privacy-preserving techniques in this domain.

Novelty

Our project introduces a novel approach to privacy-preserving sensitive information masking, departing from conventional deletion practices by integrating data aggregation techniques with innovative strategies focused on enhancing privacy and utility. While drawing inspiration from evolutionary optimization methods such as Particle Swarm Optimization (PSO), Genetic Algorithm, Ant Colony Optimization, Grey Wolf Optimization, and Honey Bee Optimization, our approach innovatively focuses on adding artificial data to decrease the frequency of sensitive information rather than deleting it. This unique strategy ensures improved privacy protection while enhancing data utility. Additionally, our methodology emphasizes real-time data transformation techniques to ensure privacy without compromising utility. Moreover, our project stands out for its practical applicability, offering secure data transformation algorithms for collaboration with third parties, thus bridging the gap between privacy preservation and real-world impact. Key aspects of our project's novelty include:

- Integration of data aggregation techniques instead of data deletion with innovative strategies for privacy preservation and utility enhancement
- Emphasis on adding artificial data to decrease the frequency of sensitive information and pattern that need to be hidden.
- Real-time data transformation techniques for privacy preservation
- Practical applicability through secure data transformation algorithms for collaboration with third parties
- Delicate balance between privacy preservation and data utility

Objectives

1.Developing a Novel Privacy-Preserving Masking Scheme: The primary objective of this project is to design and implement a novel privacy-preserving masking scheme that effectively conceals sensitive information while preserving data utility and confidentiality.

2.Overcoming the Drawbacks in already existing algorithms like Particle Swarm Optimization (PSO) and Its Variants: We aim to develop efficient masking algorithm that can dynamically optimize parameters and adapt to different types of sensitive data, enhancing the effectiveness and efficiency of the masking scheme and overcoming the drawbacks of already existing privacy preserving data mining and pattern hiding techniques like PSO, Ant Colony Optimization, etc.

3.Optimization of Masking Parameters: Another objective is to optimize the parameters of the masking algorithm, ensuring optimal performance and adaptability across various datasets and privacy requirements.

4.Comprehensive Evaluation and Validation: We aim to conduct a comprehensive evaluation and validation of the proposed masking scheme, assessing its effectiveness in preserving privacy, maintaining data utility, and scaling to real-world scenarios. This includes the development of novel evaluation criteria and benchmarks to measure the robustness and applicability of the masking scheme.

5.Cross-Domain Application and Impact: Lastly, we seek to explore the cross-domain application and impact of the masking scheme, demonstrating its versatility and transferability across diverse domains such as healthcare, finance, and social media. By showcasing the applicability of the scheme in different contexts, we aim to highlight its potential to address privacy challenges across various industries.

Methodology

- Data Input and Separation: The process begins with users inputting data through our website, which is then segregated into sensitive and non-sensitive categories. Utilizing pattern mining or knowledge mining techniques such as Apriori and FP Growth, correlations between attributes are identified. A predetermined threshold, often set at 50% correlation, is applied to filter out attributes with lower correlations, ensuring the selection of pertinent parameters.

- Victim Identification and Selection: Following the separation of sensitive data, the next step involves identifying and selecting the "victim" item or transaction that will undergo modification, addition , or other changes. Victim selection is based on transactional or item-level attributes, with a focus on identifying the portion of data that contains the sensitive information requiring transformation.

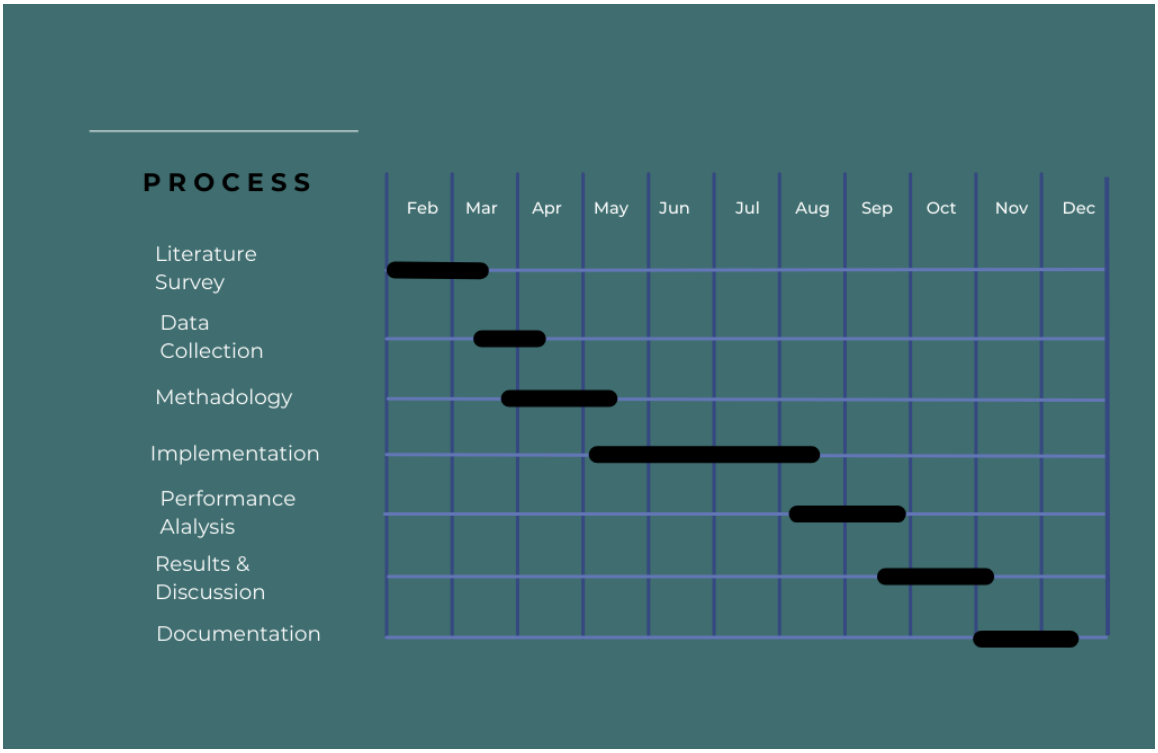
- Identification of Portion of Data to be Modified: Once the victim item or transaction is identified, algorithms such as Particle Swarm Optimization (PSO) and other optimization techniques are employed to determine the specific portion of the data where the sensitive information will be modified, changed, or deleted through the victim. These algorithms facilitate the precise identification of the data subset requiring modification, ensuring efficient and targeted data transformation while minimizing data loss.

- Privacy Preservation Transformation Techniques: Following data separation, privacy preservation techniques are applied to mask the resulting sensitive patterns while ensuring they remain below the specified threshold. Various Privacy-Preserving Data Mining (PPDM) algorithms, including techniques inspired by Particle Swarm Optimization (PSO), Genetic Algorithm, Ant Colony Optimization, are employed for this purpose. These algorithms ensure that sensitive information is effectively concealed while preserving data utility.

- Minimization of Data Loss and Constraints: Throughout the process, a primary objective is the minimization of data loss while effectively concealing sensitive patterns. Constraints are enforced to mitigate the frequency of failed attempts to hide sensitive data (FTH) and minimize non-sensitive data inadvertently hidden (NTH). Additionally, measures are taken to mitigate the impact of artificial pattern generation, ensuring that artificially introduced elements do not inadvertently become sensitive trends.

- Integrated Sanitized dataset: Once the sensitive data is masked or hidden, the non-sensitive data is directly integrated with the transformed data without requiring further modification. This streamlined approach ensures efficient data transformation while maintaining the integrity of both sensitive and non-sensitive information. This combined dataset is then made available for further analysis or utilization, securely preserving sensitive information while maximizing data utility.

Work Plan


PROCESS
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Literature Survey
Data Collection
Methadology
Implementation
Performance Alalysis
Results & Discussion
Documentation

Project Outcomes & Individual Roles

Project Outcomes:

- Development of a novel approach to privacy-preserving sensitive information masking, integrating data aggregation techniques with innovative strategies for privacy preservation and utility enhancement.
- Design and implementation of algorithms for adding artificial data to decrease the frequency of sensitive information while maintaining data utility.
- Creation of a secure data transformation application capable of real-time data transformation for collaboration with third parties.
- Demonstration of improved privacy protection and enhanced data utility compared to traditional deletion-based methods.
- Documentation of methodologies and algorithms for knowledge hiding within extensive datasets, ensuring privacy without compromising utility.

Individual Roles:

- Website Frontend Developer: Harsha and Keshav are responsible for designing and developing the user interface (UI) and user experience (UX) of the secure data transformation application.
- Backend Developer: Vinayak and Nitish are tasked with building the backend infrastructure and implementing core functionalities of the secure data transformation application.
- Documentation and Reporting Specialist: Keshav, Vinayak, and Harsha are responsible for preparing comprehensive documentation of methodologies, algorithms, and project outcomes.
- Algorithm Developer: Vinayak, Harsha, Keshav, and Nitish are involved in analyzing, designing and implementation of the algorithm.

Course Subjects

- Database Management Systems (DBMS): Understanding of database concepts and proficiency in designing and implementing database schemas for secure data storage and retrieval.

- Web Development: Proficiency in frontend technologies such as HTML, CSS, and JavaScript for designing user interfaces and backend technologies such as Node.js or Django for building server-side logic and APIs.

- Algorithm Design and Analysis: Knowledge of algorithmic techniques for designing efficient algorithms to add artificial data and decrease the frequency of sensitive information while maintaining data utility.

- Data Privacy and Security: Understanding of privacy-preserving techniques and security measures to ensure the confidentiality and integrity of sensitive information during data transformation and collaboration with third parties.

- Documentation and Technical Writing: Proficiency in documenting methodologies, algorithms, and project outcomes in a clear and comprehensive manner, including user manuals and technical specifications.

- Software Engineering: Knowledge of software development methodologies and best practices for managing the development lifecycle of the secure data transformation application.

- Optimization Techniques: Understanding of optimization techniques such as Particle Swarm Optimization (PSO), Genetic Algorithm, and other evolutionary algorithms for enhancing privacy preservation and data utility in real-time data transformation.

- Real-time Data Processing: Knowledge of techniques and frameworks for processing data in real-time to ensure privacy preservation and utility enhancement during data transformation.

REFERENCES

[1] Al-Rubaie, Mohammad, and J. Morris Chang. "Privacy-preserving machine learning: Threats and solutions." IEEE Security & Privacy 17.2 (2019): 49-58.

[2] Amiri, Ali. "Dare to share: Protecting sensitive knowledge with data sanitization." Decision Support Systems 43.1 (2007): 181-191.

[3] Dwork, Cynthia. "Differential privacy." International colloquium on automata, languages, and programming. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[4] Fung, Benjamin CM, et al. "Privacy-preserving data publishing: A survey of recent developments." ACM Computing Surveys (Csur) 42.4 (2010): 1-53.

[5] Jangra, Shalini, and Durga Toshniwal. "VIDPSO: Victim item deletion based PSO inspired sensitive pattern hiding algorithm for dense datasets." Information Processing & Management 57.5 (2020): 102255.

[6] Lin, Jerry Chun-Wei, et al. "A sanitization approach for hiding sensitive itemsets based on particle swarm optimization." Engineering Applications of Artificial Intelligence 53 (2016): 1-18.