International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Smart Privacy Preserving and Security Mining Techniques

MR. Labhade Vishal S. and Prof. Shaikh I.R

Department of Computer Engineering, SND College of Engg. Yeola

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

In recent years privacy preserving data mining has become popular and continuously evolving field of study, preserving privacy and ensuring the security of data has emerged as important issues. Although there are several frameworks and tools to handle such issues, they mostly implement data anonymity techniques. Data mining also opens new threats and challenge to privacy and information security Thus, this paper presents a novel Privacy-Preserving and Security Mining, which focuses on privacy-preserving data mining and data security. It is an open-source data mining library, there are several algorithms in PPDM like: (1) data anonymity, (2) privacy- preserving data mining, and (3) privacy- preserving utility mining. Our Data mining tool has a user- friendly interface that allows running algorithms and displaying the results

Keywords: *Privacy-preserving data mining, privacy- preserving utility mining, security, data anonymity*

Introduction

There are many data mining tools and frameworks have been designed such as Weka, Knime, Mahout, and SPMF. Different tool has different functionalities such as Association Rule Mining (ARM), Sequential Pattern Mining (SPM), and Utility-Driven Mining (UDM). In recent decade with the availability of more and more powerful data mining tools, it has become easy to find implicit and interesting and useful information in datasets. However, confidential or private and sensitive information may be occasionally revealed by data mining algorithms or tools, which may cause important security and privacy issues. Thus, privacy-preserving data mining and data security problems have become important research topics, with a focus on subjects such as Data Anonymity, PrivacyPreserving Data Mining, and Privacy- Preserving Utility Mining. There are some major problems affecting research progress on those topics is that many researchers do not share their source code or there project implementations online, mainly for the important topics like PPDM and PPUM. Thus, those who want to carry research on these topics or apply these techniques need to implement the algorithms again. This more time, requires programming skills and advanced Understanding of data mining, and is prone to errors. Several famous software such as ARX, GUPT, Harvard University Privacy Tools Project, and Data Anonymization were presented. However, they mostly focused on data anonymization. But there was no focus on the topics of PPDM or PPUM. But these two topics are important since only useful and meaningful information should be discovered for decision-making and the side-effects of the sanitization progress should also be considered. To overcome this limitation and encourage research and application of privacy preserving and data security, the Privacy- Preserving and Security mining techniques are proposed as a novel opensource library. The goal of Security mining techniques is to provide a common library to share the source codes of algorithms. We expect that this will increase the importance of this topic and its real-life applications. The implemented algorithms can be considered as references, which can be used for performance comparison and usage in various applications and domains. It is fast and lightweight, implemented in Java, and without dependencies to other projects. It offers a userfriendly interface, from which when a user selects an algorithm result is displayed in a simple text file format to facilitate understanding and interoperability. This paper introduces the purpose of the three categories of techniques offered, also its main features.

Literature Survey

Data anonymization is a promising process within the discipline of privacy preserving data mining used to protect the information in opposition to identity disclosure. Information loss and long-established attacks possible on the anonymize information are critical challenges of anonymization. Not too long ago, knowledge anonymization utilizing information mining strategies has showed gigantic improvement in information utility. Nonetheless the prevailing

approaches lack in robust handling of attacks. As a result J. Jesu Vedha Nayahi et al. proposed an anonymization algorithm established on clustering and resilient to similarity attack and probabilistic inference attack is proposed. It prevents the privacy revelation of the sensitive information. The privacy defense mechanism avoids the identity and attributes disclosure. The privateness is executed by means of the high accuracy and consistency of the person expertise, i.e., the precision of the personal data. For addressing the drawback of identical privateness safety for all relocating objects in trajectory knowledge, Elahe Ghasemi Komishani et al. proposed PPTD, a novel process for keeping privateness in data publishing established on the concept of personal privacy. They targets to strike stability between the conflicting objectives of information utility and knowledge privacy in line with the privateness standards of relocating objects. They combine sensitive attribute generalization and trajectory nearby personalized privacy model for trajectory data publishing.

They performed experiments on two artificial trajectory datasets and concluded that PPTD is powerful for maintaining personalized privateness in trajectory information publishing. The usual data publishing ways will do away with the sensitive attributes and generate the considerable records to attain the goal of privacy safety. In the big data environment, the requirement of using information (e.g., data mining) come to be more and more quite a lot of, which is beyond the scope of the normal procedure. Tong Li et al. Presents a cryptographic data for publishing system that preserves the information integrity (i.e., the long-established knowledge structure is preserved) and achieves anonymity without deletion of any attribute or utilization of redundancy. The safety analysis suggests that their process is secure underneath proposed security model. Surbhi Sharma et al. Show how the exclusive departments of same group combine their data without harming the privateness of the client for making robust selections in efficient and correct manner. For that reason the approaches vertically information combination and decision mining is established. To mine the choices from the information a C4.5 resolution tree is used. Additionally the efficiency of the method is computed in phrases of accuracy, error rate, memory consumption and time consumption. In the end to justify the effects of the proposed data mining system the normal J4.5 tree utilizing WEKA instrument is used with same knowledge for comparative performance learn. The experimental results show the mighty performance and protection within the given privacy preserving procedure.

PPSM Architecture

Privacy Preserving Security Mining contains three main modules: (1) a user-friendly graphical-interface, (2) a data processing module, and (3) a visualization module. The overall architecture, consisting of these modules is illustrated in Fig 1.

User Interface Module: This module provides a user-friendly graphical interface. This interface lets the end-user select and run algorithms suitable for them for generating desired output. Generally, inputs are given as transactional datasets or utility-based datasets. Based on the choice of an algorithm, user can adjust parameters. The output location for result is also set by user. To facilitate interoperability with other software, results are stored in a form of plain text file format generally in notepad.

Processing Module: This module consists of 13 algorithms for data anonymity, PPDM and PPUM. The algorithms implemented and integrated into Privacy Preserving Security

Mining, including the conventional GSP, PTA, Greedy, SIF- IDF, HHUIF, MSICF, MSU MAU, MAU MIN and the Evolutionary sGA2DT, pGA2DT, cpGA2DT, PSO2DT, pGAPPUM algorithms. Implemented algorithms are shown in table I.

The key features and novelties of PPSM are as follows. First, it offers a total of 13 algorithms, organized in three categories, which can provide a great variety of techniques, suitable for the needs of different users.

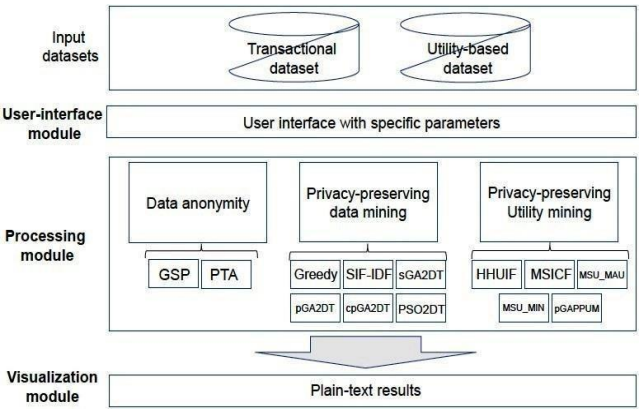

Fig. 1: The three-module architecture.

Visualization Module: The default output for results is a stored in plain text format, which can be easily visualized and interpreted by users. It can automatically launch the default text editor (e.g. Notepad for the Windows platform) to let the user inspect results. This module provides flexibility since a result file can be opened using various text editors, and is compatible with all operating systems where Java can be run.

Purposes and Features

The Privacy Preserving Security Mining project categorizes into three categories: data anonymity Privacy preserving data mining, and privacy-preserving utility mining. The purpose of each category and the included algorithms are explained below.

Data Anonymity: The purpose of Data Anonymity is to

Anonymize data so that each data record is in distinguished- able from at least (*k*-1) other records according to the *k*- anonymity principle. This methodology is used to handle attributed datasets. Two algorithms named GSC and PTA algorithms are offered for Data Anonymity.

Privacy-Preserving Data Mining (PPDM): The purpose of PPDM is to sanitize a database for hiding sensitive information while ensuring that useful and meaningful knowledge can still be extracted for decision-making. It usually considers three factors such as hiding failure, missing cost, and the artificial cost for evaluation. The dataset dissimilarity between the original dataset and the sanitized one is sometimes considered as another factor to show the performance. Six algorithms have been implemented in Model, including Greedy, SIF- IDF, GA-based approaches and particle- swarm optimization based algorithms. **Privacy-Preserving Utility Mining (PPUM):** PPUM is an extension of PPDM it handles databases having utility information. This has emerged as an important need since high- utility itemset mining (HUIM) has become an important data mining topic in recent years. PPUM is used to reveal High- Utility Patterns (HUPs) from databases while ensuring that sensitive high utility itemsets remain hidden. PPUM also considers perfor- mance factors that are similar to PPDM but it focuses on utility values. It offers five algorithms for PPUM, namely HHUIF, MSICF, MSU MAU, MSU MIN, and pGAPPUM .

V PPDM algorithm

Input: V, N //The inputs are the vectors of V, composed of confidential numerical attributes only, and the uniform noise vector N, while the output is the transformed vector subspace V'.//
Output: V'
Step 1. **For** each confidential attribute A_j in V, where $1 \le j \le d$ **do**
1. Select the noise term e_j in N for the confidential attribute A_j
2. The *j-th* operation $op_j \leftarrow \{Add\}$
Step 2. **For** each $v_i \in V$ **do**
For each a_j in vi $= v_i = (a_1, \ldots, a_d)$, where a_j is the observation of the *j-th* attribute **do**
1. $a_j' \leftarrow trasform(a_j, op_j, e_j)$
Transform(aj; opj ; ej)
End

There are six PPDM algorithms, five PPUM algorithms, and two data anonymity algorithms

Table I: The implemented algorithms

Algorithm	Description
GSC [20]	The algorithm use deleting QID items for Achieve sensitive *k*-anonymity on transaction data.
PTA [14]	The state-of-the-art algorithm for not only anonymize transactional data with a small information loss but also to reduce the computational complexity of the anonymization process.
Greedy [22]	The first greedy algorithm for hiding the number of sensitive rules.
SIF-IDF [11]	The first algorithm uses the TF-IDF concept for Hiding the sensitive item sets.
sGA2DT [12]	The first algorithm to apply the genetic algorithms (GAs) for hiding sensitive information.
pGA2DT [12]	The improved algorithm of the GA-based algorithm for hiding sensitive information.
cpGA2DT [13]	The first algorithm to hide the sensitive information using the compact GAs.
PSO2DT [15]	The first PSO-based algorithm for hiding sensitive information.
HHUIF [24]	The first algorithm to handle the sensitive high utility patterns.
MSICF [24]	The improved algorithm of HHUIF for hiding sensitive high-utility item sets.
MSU MAU [17]	The improved algorithms for hiding the sensitive high utility patterns that are using maximum criteria.
MSU MIN [17]	The improved algorithms for hiding the sensitive high utility patterns using minimum criteria.
GAPPUM [16]	The first evolutionary algorithm for hiding sensitive high utility patterns.

System Analysis

Mathematical Model A mathematical model is a description of a system using mathematical concepts and language. The process of developing a mathematical model is termed as mathematical modeling. As the project is having finite input and finite output, it comes under P-Problem. Let the system be described by S, S= {I, P, R, O} Where, S: is a System. I: is Input R: is set of Rules O: Final Output. I= {I1; I2} Where, I1 = Enter Input In the form of dataset I2 = Select Algorithm to apply P is set of procedure or function or processes or methods. P = {P1, P2, P3}; Where, P1 = Process Input. P2 = Apply selected Algorithm P3 = Generate Result R is set of Rules R1, R2; R1 = Enter Valid Information. R2 = Match the appropriate algorithm. Output= {O1, O2} Where, O1, O2 = Generate Result in the form of Text P is set of procedure or function or processes or methods. P = {P1, P2, P3}; Where, P1 = Check login

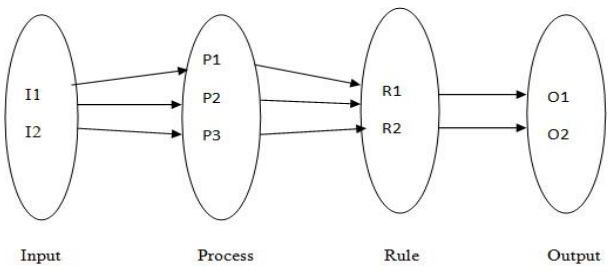

Fig 2:Venn Diagram

System Requirements

Hardware requirements: Hard disk: 128 GB RAM: 512 MB Processor: Pentium and above Input device: Keyboard and Mouse Output device: Monitor Software requirements: Operating System: Windows 7/Linux Front End: HTML, JS, Bootstrap Back End: MySQL, Oracal 10 g UML Design: StarUml VIII. PROPOSED SYSTEM. 1) Classification in data mining methodology aims at constructing a model (classifier) from a training data set that can be used to classify records of unknown class labels. 2) There are different algorithms in proposed system each one has its different functionality

3) Different types of parameter consider for fatality. Linear regression is useful for finding relationship between two variables. 4)Integrate additional visualization capabilities and processing tools to help preparing data for anonymization.

Conclusion and Results

In this paper, we have presented a novel Privacy Preserving and Security Data Mining, which provides 13 algorithms, organized in three categories: privacypreserving data mining, privacy- preserving utility mining, our initial results indicates that deterministic algorithms for privacy preserving are a promising framework for controlling disclosure of sensitive data and knowledge and data anonymity. We are providing a user-friendly interface for selecting algorithms and results are provided in a flexible and easyto-understand text file format. The current development of is focused on adding more baseline algorithms for the three major categories. Moreover, more algorithms will be added to the software related to security problems of the Internetof-Things (IoT) and cloud security, for the next releases. In the current platform, a greater effort has been put on developing the data processing module to provide rich functionality and user friendly Interface.

Acknowledgment

I would like to take this opportunity to express my profound gratitude and deep regard to my Project Guide Prof. I. R. Shaikh, for his exemplary guidance, valuable feedback and constant encouragement throughout the duration of the project Also Prof. V. N. Dhakane (PG coordinator) who provided facilities to explore the subject with more enthusiasm. I express my immense pleasure and thankfulness to all the teachers and staff of the Department of Computer Engineering, S.N.D. College of Engineering and Research Centre, Yeola, Nasik for their co-operation and support. ABOUT AUTHORS Mr. Labhade Vishal Sarjerao, he received his BE from. SRES College of Engineering, Koprgaon currently perusing. ME from Computer Engineering Department, S.N.D College of Engineering and Research Centre, Yeola, Nasik.His main area of interest is Data Mining and Information Retrieval Guide Prof. Shaikh I.R.Currently working as Assistant Professor Computer Engineering Department, S.N.D College of Engineering and Research Centre,Yeola, Nasik.

References

[1] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases,” International Conference on Very Large Data Bases, pp. 487–499, 1994

[2] R. Agrawal and R. Srikant, “Mining sequential patterns,” IEEE International Conference on Data Engineering, pp. 3–14, 1995

[3] R. Agrawal and R. Srikant, “Privacy-preserving data mining,” ACM SIGMOD Record, vol. 29(2), pp. 439–450, 2000

[4] D. Agrawal and C. Aggarwal. On the Deisgn and Quantification of Privacy Preserving Data Mining Algorithms. Proceedings Of PODS, pages 247–255, 2001.

[5] R. Agrawal and R. Srikant. Privacy Preserving Data Mining. Proceedings of SIGMOD Conference, pages 45–52, 2000.

[6] P. Fournier-Viger, J. C. W. Lin, A. Gomariz, T. Gueniche,A.Soltani