Privacy Protection in Data Mining

Yilai Chen 12013025

1 INTRODUCTION

With the advancement of information technology, ensuring high data availability while protecting the privacy of data subjects has become a major focus for researchers worldwide. Key privacy protection techniques include data encryption, data distortion, and data anonymization, with anonymization being the primary method in privacy-preserving data mining. Researchers have proposed various anonymization models, including k-anonymity, l-diversity, (,k)-anonymity, and t-closeness. This paper analyzes and compares these four main models, highlighting their characteristics, advantages, and disadvantages. It also summarizes common anonymization techniques and current methods for measuring anonymization quality. Looking ahead, as anonymization continues to be a crucial method for privacy protection in data mining, several challenges remain to be addressed, and future research directions are explored. [3]

2 GENERAL BASIC PROCESS

The process of privacy-preserving data mining primarily includes the following steps: (1) defining the data mining objectives, (2) data cleaning, which involves removing or hiding sensitive data from the raw dataset, (3) data preprocessing, where the cleaned data is transformed and encrypted, (4) data mining, where appropriate algorithms are selected for extracting useful information, (5) evaluating the results, using metrics such as privacy, which measures the likelihood of unauthorized identification of private information, effectiveness, which assesses the accuracy of the final results, and complexity, which considers the time and space complexity of the algorithms, and (6) visual representation, which involves using visualization tools to represent the mined knowledge.

In this field, many related algorithms have been proposed. The following is a brief list of some algorithms we have found on the Internet: perturbation algorithms, including randomized perturbation algorithms (randomly adding noise, data conversion matrix, k-means and multiple random hashes) and multiplication perturbation algorithms (rotation perturbation algorithms and projection perturbation algorithms). k-anonymity algorithms, the two main algorithms of k-anonymity are generalization and suppression. Association rule hiding algorithms, common association rule hiding algorithms include heuristic algorithms, boundary-based algorithms and precise algorithms.

And generally speaking, the evaluation indicators can be defined as follows: Data quality. In data mining that needs to be privacy-protected, some processing is usually performed on the data, which may be destructive to the data, thus further causing deviations in the data sets before and after mining. However, it will not affect the results of data mining or the impact is very small. Therefore, the quality of data is usually measured based on accuracy, consistency, and completeness. Performance cost. In the final analysis, privacy-preserving data mining is still data mining, and performance is still an important consideration for mining methods. The performance is measured by evaluating the time complexity and space responsibility. Scalability. It usually refers to how to effectively protect the attributes of the data set when performing related magnification and reduction operations on the target data set, and has good performance indicators. Privacy degree. This is the most core indicator and the most difficult indicator to measure. There are a lot of results based on information entropy research, which is also a relatively accepted method. Agrawal first proposed the method of information entropy, and Berino made an effective extension.

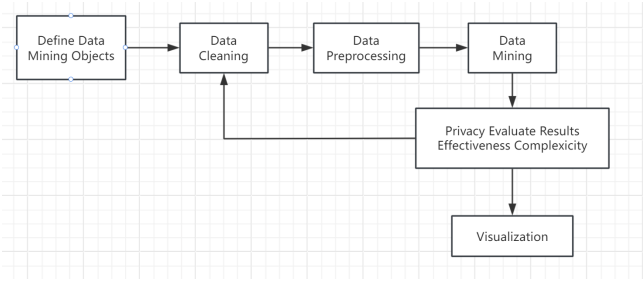

Figure 1. Privacy-Preserving Data Mining Process

3 PRIVACY PROTECTION TECHNOLOGIES

3.1 Data Distortion Techniques

In order to protect the privacy of published databases, many effective data mining techniques have been proposed to hide sensitive information. The purposes of privacy protection are as follows: (1) hide sensitive information contained in the original data; (2) the hidden data has the same characteristics as the original data; (3) obtain the same data accuracy as the original data set. Privacy-preserving data mining algorithms, such as classification, association rule discovery, and clustering, require the selection of data to be modified or purified, and the selection of purified data is an NP-hard problem. To deal with this complex problem, distortion methods such as random perturbation, blocking, and compression are used. Perturbation-based association rule mining Statistical significance is used to judge the occurrence of rules in

the data set, and support and confidence are used as metrics. All association rules are greater than or equal to the user-defined support and confidence, but from the user's perspective, some rules are sensitive and some are not. Association rule hiding technology is to purify the original data set using the following method. All sensitive rules can only appear in the original data mining, and rules with the same (or greater than) confidence and support are not allowed to appear when the data set is purified. Non-sensitive rules that can be mined in the original data set can also be mined on the purified data set with the same support and confidence.Sensitive rules that cannot be mined in the original dataset cannot be mined in the purified dataset either. For association rule mining with hidden large itemsets, the optimal purification problem is NP-hard. Reference proposes a major project to purify sensitive rules by purifying sensitive sets. The approach taken in this work is to prevent the generation of sensitive rules by hiding the frequent itemsets from which they come, or by reducing the confidence of sensitive rules below a user-specified threshold. These two approaches lead to the generation of three strategies for hiding sensitive rules. Regarding these three strategies, it is important to mention that the flexibility of data modification has side effects. In addition to non-sensitive association rules being hidden, non-frequent rules may also become frequent rules.[1]

3.1.1 Mining Association Rules Using Blocks. Another perturbation of association rules by data modification methods is data block. Blocking methods replace the attribute values of data items with question marks, and replacing actual values with unknown values instead of false values is very popular in medicine. Reference proposes a method for mining association rules using blocks, appropriately changing the definition of minimum support, replacing minimum support intervals and minimum confidence, and replacing confidence intervals. We believe that privacy is not violated as long as the support of a sensitive rule is lower than the middle of the support interval or the confidence of a sensitive rule is lower than the middle of the confidence interval. Either 1 or 0 should be mapped to a question mark, otherwise the original value of the question mark is exposed. Reference describes in detail the effectiveness of the chunking approach, which uses perturbation rules to reconstruct text.

3.1.2 Chunking-based Classification Rule Mining. Reference provides a new framework that combines classification rule analysis and parsimony degradation, in which the goal of the data curator is to chunk the values of the class labels. In this way, the information receiver cannot build an information model for the un-degraded data. Parsimony degradation is a framework for formalizing the phenomenon of pruning information from a dataset to degrade the information. In parsimony degradation, a cost metric is assigned to the potentially degraded information that is not sent to the downgraded level. The main goal to be accomplished in this work is to find out whether the functional loss of not degrading the data is worth the additional confidentiality.

3.2 Distributed Privacy Preserving Mining

Within the realm of privacy-preserving data mining, numerous encryption-based approaches have emerged to address the challenge of cooperative data analysis while safeguarding individual data privacy. This entails secure multiparty computation (SMC) techniques, particularly pertinent in distributed environments, where the focus is on transforming diverse data mining methodologies into secure multiparty computation tasks. These tasks encompass data classification, clustering, association rule mining, generalization, and aggregation.

One aspect of interest is distributed association rule mining, wherein two or more parties collaborate to analyze their data without divulging the underlying information. Vertical partitioning entails segregating different attributes of items across various sites, while horizontal partitioning involves distributing transactions among multiple databases.

In vertically partitioned data association rule mining, the emphasis lies in mining private association rules by securely determining the support count of specific itemsets. This involves computing the support count in a secure manner, allowing for the determination of frequent itemsets while preserving privacy. Each party contributes to the computation by providing sub-item sets represented as vectors, facilitating the calculation of support counts through vector dot products.

Conversely, horizontally partitioned data association rule mining involves distributing transactions across multiple sites. Here, the total support count of an itemset is derived from the sum of local support counts across all sites. A globally supported itemset is identified based on its support count relative to the total transaction database size, ensuring that privacy is maintained while deriving meaningful association rules.[1]

3.3 Reconstructed Technology

Recent advancements in privacy-preserving data mining have explored techniques involving data perturbation or reconstruction at the convergence layer. In a notable study cited as, researchers examined the construction of a decision tree classifier using perturbed values from individual records as training data. Acknowledging the inherent inaccuracy in estimating original values, the study focused on accurately estimating the original distribution. To achieve this, Bayesian methods were considered for reconstructing the original distribution.

Building upon this work, reference introduced improvements to the Bayesian reconstruction process by incorporating the Expectation-Maximization (EM) algorithm within a distributed data framework. The study demonstrated that the

EM algorithm not only maximized estimation accuracy for the original data distribution amidst disruptions but also exhibited robustness when handling large datasets. Moreover, highlighted the impact of background knowledge on privacy estimation, indicating that reconstructions of the distribution could potentially lead to decreased privacy levels when accessed by data miners.

These findings underscore the evolving landscape of privacy-preserving data mining, where innovative methodologies strive to balance the imperatives of data utility and individual privacy.[1]

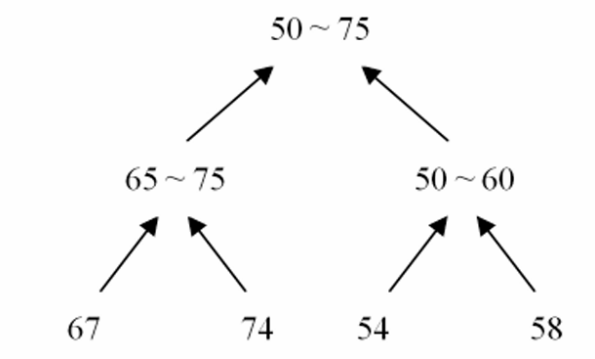

Figure 2. Generalization of Age[3]

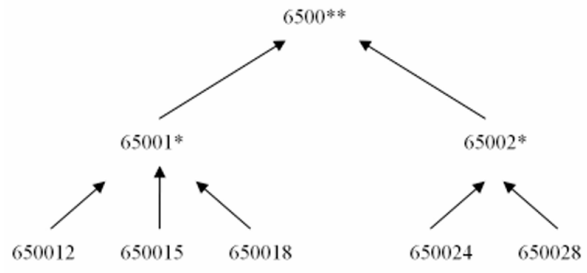

Figure 3. Global Generalization of ZIP[3]

3.4 Anonymous Privacy Protection

In the realm of privacy protection, a critical consideration is the selective release of data to mitigate privacy risks while maintaining data utility. This study focuses on data anonymity techniques, which involve striking a balance between privacy disclosure risks and data utility. This entails selectively releasing data and information to ensure that sensitive data remains within acceptable privacy risk thresholds.

Data anonymity revolves around two key principles. Firstly, there's the imperative to design robust anonymity methods that not only safeguard privacy effectively but also retain practical utility. Secondly, there's a need to develop efficient anonymity algorithms that align with specific anonymity principles. As research delves deeper into anonymity, the practical application of anonymous data emerges as a central research focus.

One seminal principle in data anonymity is the concept of k-anonymity, introduced by Samarati and Sweeney, which mandates that each released record cannot be distinguished from at least k-1 other records, forming equivalent classes. However, the flaw of k-anonymity lies in its failure to impose constraints on sensitive data, leaving room for privacy breaches through various attacks, such as background knowledge attacks.

To address these limitations, (Ɪ, k)-anonymity was introduced, aiming to improve upon k-anonymity by ensuring that each released equivalence class adheres to a certain threshold of attribute value occurrence, denoted by Ɪ.

Traditional data publishing methods, including k-anonymity, l-diversity, and t-closeness, often rely on generalization techniques, which significantly reduce data accuracy and utility. For instance, static data release mechanisms like l-diversity ensure that the disclosure risk of published datasets remains below a certain threshold, while dynamic data publishing principles, such as m-invariance, maintain disclosure risks within acceptable limits.[1]

3.5 Bottom-up Generalization

The growing capacity to gather, store, retrieve, cross-reference, analyze, and link vast amounts of electronic records has provided substantial benefits to millions of people. For example, correlating personal records of chemical exposure with death records can help identify carcinogenic substances. However, these advancements also bring about significant privacy and liability concerns due to the potential for uncovering new information. As illustrated, a sensitive medical record can be uniquely connected to an individual voter record in a public voter list through common attributes such as postal code, birth date, and gender. In reality, safeguarding individual data sources does not guarantee protection when these sources are cross-referenced, as "the whole is greater than the sum of its parts." An important area of research is developing methods to prevent the deduction of private information through record linkage while still reaping the benefits of data sharing and mining.[2]

3.5.1 Method Overview. Sensitive information refers to details that are specific to a small number of individuals. In contrast, data mining typically uses information shared by many individuals to ensure the statistical significance of patterns. Therefore, sensitive information should be excluded for reliable data mining. This idea leads us to use data mining requirements to identify which information can be shared and which sensitive information should be masked. This approach, known as data mining-based privacy protection, turns data mining from a privacy threat into a privacy solution.

An anonymity problem was addressed where a data holder

wants to release person-specific data (R) without it being linked to external sources (E) through shared attributes called virtual identifiers. One solution is to generalize specific values into less specific but semantically consistent ones to create K-anonymity. This means that if a record in R is linked to some external information, at least K-1 other records will have the same virtual identifier, making the inference ambiguous. For example, generalizing “birth date” to “birth year” means all individuals born in the same year share a common birth year in the medical records, even though most of these linkages are not real.

Our focus here is on using a classifier. We introduce a data mining technique known as iterative bottom-up generalization to achieve K-anonymity while maintaining the utility of the generalized data for classification. This generalization process is guided by a hierarchical structure for each attribute within the virtual identifier. The main challenge is to identify the optimal generalization to ascend the hierarchy in each iteration. Since evaluating all potential candidates in every iteration is impractical, we offer a scalable solution that considers at most one generalization per attribute in each iteration. The evaluation workload is proportional to the number of distinct virtual identifier values that undergo generalization. We evaluate the effectiveness and scalability of this method.

3.5.2 Key Concepts. Definition 1: Anonymity: A virtual identifier, denoted as VID, is a set of attributes shared by R and E. a(vid) represents the number of records in R with value vid on VID. The anonymity of VID, denoted as A(VID), is the minimum a(vid) for any value vid on VID. A vid is called an anonymous vid if a(vid) = A(VID). R satisfies the anonymity requirement <VID, K> if A(VID) K, where K is specified by the data holder.

We transform R to meet the anonymity requirement by generalizing a specific value on VID to a less specific but semantically consistent value. Generalization increases the probability that a specific value on VID occurs by chance, and therefore reduces the probability that a connection through this value represents a real-life fact. The generalization space is specified by a classification hierarchy for each attribute in VID, which is provided by the data holder or data recipient. The hierarchy is a tree with leaf nodes representing domain values and parent nodes representing less specific values. R is generalized by a series of generalizations, where each generalization replaces all child values c in the hierarchy with their parent value p. Before generalizing a value c, all values below c should first be generalized to c.

Definition 2: Generalization: A generalization, written as c→p, replaces all child values c with their parent value p. A generalization is valid if all values below c have been generalized to c. If vid contains a value in c, then vid is generalized by c→p.

3.5.3 Algorithm. The algorithm describes our bottom-up generalization process. In the i-th iteration, we generalize R to the "best" generalization Gbest according to the IP metric. The algorithm has no requirements on efficiency, because lines 2 and 3 need to calculate IP(G) for all candidate generalizations G. Let’s look at this calculation process in more detail.

Consider a candidate generalization G: c→p in an iteration. |Rc and freq(Rc, cls) can be maintained after each iteration. |RD|. and freq(Rp, cls) can be obtained by aggregating |Rc| and freq(Rc, cls). Therefore, I(G) can be easily calculated, that is, without access to vids. In fact, any metric on a single attribute (plus class labels) can be calculated in this way. A(V I D) is the result of applying the previous generalization. However, calculating Ac(V I D) depends on the "effect" of G, which can only be obtained after applying G and requires access to vids. This is a new scalability challenge.

Our insight is that most generalizations G do not affect A(V I D), and therefore, AG(VID) = A(VID). In fact, if a generalization G fails to generalize all anonymous vids, then G will not affect A(VID). For such G, P(G) = 0, IP(G) = , and our metric does not require AG(V I D).

4 CONCLUSION

Privacy preservation technology has surfaced as a burgeoning realm of scholarly exploration, permeating various sectors with its manifold applications in recent times. This article endeavors to furnish an all-encompassing examination of privacy safeguarding methodologies within the domain of data mining. Our journey commences with a deep dive into the prevailing panorama of privacy preservation research, unraveling its fundamental methodologies. Following this, we navigate through an array of privacy safeguarding approaches, encompassing distortion, encryption, and privacy alongside anonymity techniques, underpinned by pertinent examples drawn from scholarly literature.

Due to its interdisciplinary essence, the realm of privacy protection technology offers a vast array of uncharted territories awaiting exploration. Take, for instance, the domains of mobile data mining and data stream mining, which present captivating puzzles surrounding privacy in the realm of data mining, thus beckoning towards promising realms for future investigation. With the continued proliferation of spatial and geographical data, we anticipate the emergence of innovative applications harnessing user mobility patterns. Furthermore, the notion of incremental privacy protection data release introduces a captivating realm for exploration, entailing the intricate challenge of adapting data mining algorithms to seamlessly integrate incremental data processing.

In addition to specialized inquiries within particular domains, there exists an urgent imperative for the formulation of a comprehensive framework to evaluate and contrast a myriad of privacy protection data mining algorithms. Such a

framework holds the potential to streamline decision-making processes and propel advancements in privacy protection technology within the sphere of data mining.

References

[1] Xinjun Qi and Mingkui Zong, *An overview of privacy preserving data mining. Procedia Environmental Sciences*, volume 12, pages 1341–1347, Elsevier, 2012.

[2] Ke Wang, Philip S. Yu, and Sourav Chakraborty, *Bottom-up generalization: A data mining solution to privacy protection*. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 249–256, IEEE, 2004.

[3] TAN, Ying, *Progress in Anonymous Privacy-Preserving in Data Mining*. In *Yunnan University of Finance and Economics*, Department of InformationKunming 650221, China.