An adaptive mechanism for anonymizing set-valued data

[1] R.ARPITHA, [2]K.SIRISHA

[1]PGScholar, Dept., of CSE, Chadalawada Ramanamma Engineering College, Tirupati, A.P., INDIA.

[2]Asst. Prof., Dept., of CSE, Chadalawada Ramanamma Engineering College, Tirupati, A.P., INDIA.

Abstract

Classification is a fundamental problem in data analysis. Training a classifier requires accessing a large collection of data. Releasing person-specific data, such as customer data or patient records, may pose a threat to individual's privacy. Even after removing explicit identifying information such as Name and SSN, it is still possible to link released records back to their identities by matching some combination of non-identifying attributes such as {Sex,Zip, Birth date}. A useful approach to combat such linking attacks, called k-anonymization , is anonymizing the linking attributes so that at least k released records match each value combination of the linking attributes. Previous work attempted to find an optimal k-anonymization that minimizes some data distortion metric. We argue that minimizing the distortion to the training data is not relevant to the classification goal that requires extracting the structure of predication on the "future" data. In this paper, we propose a anonymization solution for classification. Our goal is to find a anonymization, not necessarily optimal in the sense of minimizing data distortion, that preserves the classification structure. We conducted intensive experiments to evaluate the impact of anonymization on the classification on future data. Experiments on real life data show that the quality of classification can be preserved even for highly restrictive anonymity requirements.

KEY WORDS : Privacy protection, anonymity, security, integrity, data mining, classification, data sharing

1. INTRODUCTION

Data Mining is also called knowledge discovery in databases (KDD). Data mining is about finding new information in a lot of data. The information obtained from data mining is hopefully both new and useful. The data is saved with a goal. For example, a store wants to save what has been bought. They want to do this to know how much they should buy themselves, to have enough to sell later. Saving this information, makes a lot of data. The data is usually saved in a database. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. Privacy preserving data analysis, and data publishing have received considerable attention in recent years as promising approaches for sharing data while preserving individual privacy. In a non-interactive model, a data provider (e.g., hospital) publishes a "sanitized" version of the data, simultaneously providing utility for data users (e.g., researchers), and privacy protection for the individuals represented in the data (e.g., patients). When data are gathered from multiple data providers or data owners, two main settings are used for anonymization . One approach is for each provider to anonymize the data independently which results in potential loss of integrated data utility. A more desirable approach is *collaborative data publishing* , which anonymizes data from all providers as if they would come from one source , using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols. The address the issue of privacy preserving data mining specifically,then consider a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. Our work is motivated by the need to both protect privileged information and enable its use for research or other purposes. The above problem is a specific example of secure multi-party computation.

2. RELATED WORK

A simple privacy-preserving reformulation [1] of a linear program whose equality constraint matrix is partitioned into groups of rows. Each group of matrix rows and its corresponding right hand side vector are owned by a distinct private entity that is unwilling to share or make public its row group or right hand side vector. By multiplying each privately held constraint group by an appropriately generated and privately held random matrix, the original linear program is transformed into an equivalent one that does not reveal any of the privately held data or make it public.
The solution vector of the transformed secure linear program is publicly generated and is available to all entities. privacy-preserving classification and data mining, wherein the data to be classified or mined is owned by different entities that are unwilling to reveal the data they hold or make it public, has spread to the field of optimization and in particular linear programming. In a number of shortcomings in the privacy-preserving linear programming literature are pointed out. In a method for handling privately held vertical partitions of a linear programming constraint matrix and cost vector is proposed that is based on private random transformations of the corresponding problem variables. The BIRCH algorithm [2] is a well known algorithm for clustering for effectively computing clusters in a large data set. As the data is typically distributed over several sites, clustering over distributed data is an important problem. The data can be distributed in horizontal, vertical or arbitrarily partitioned databases. But, because of privacy issues no party may share its data to other parties. The problem is how the parties can cluster the distributed data without breaching privacy of others data. The solutions in arbitrarily partitioned database setting generally work for both horizontal and vertically partitioned databases. It give a procedure for securely running BIRCH algorithm over arbitrarily partitioned database. Introduce secure protocols for distance metrics and give a procedure for using the semetrics in securely computing clusters over arbitrarily partitioned database. The Privacy preserving [3] Data mining has been a popular research area for more than a decade due to its vast spectrum of applications. The aim of privacy preserving data mining researchers is to develop data mining techniques that could be applied on databases without violating the privacy of individuals. This work propose methods for constructing the dissimilarity matrix of objects from different sites in a privacy preserving manner which can be used for privacy preserving clustering as well as database joins, record linkage and other operations that require pair-wise comparison of individual private data objects horizontally distributed to multiple sites.ID3 Algorithm [4] describes, Privacy and security concerns can prevent sharing of data, derailing data mining projects. Introduce a generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data distributed over two or more parties. Along with the algorithm, it give a complete proof of security that gives a tight bound on the information revealed. While this has been done for horizontally partitioned data. It present an algorithm for vertically partitioned data: a portion of each instance is present at each site, but no site contains complete information for any instance. This problem has been addressed, but the solution is limited to the case where both parties have the class attribute Ease of Use

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

3. PROPOSED WORK

3.1 Anonymization for set valued data

The baseline algorithm utilizes a data provider-aware algorithm with adaptive verification strategies to ensure high utility and *m*-privacy for anonymized data. The SMC implements the m privacy anonymization in a distributed environment while preserving security. For a privacy constraint C that is generalization monotonic m-privacy with respect to. C is generalization monotonic. Most existing generalization-based anonymization algorithms are modified to guarantee m-privacy with respect to C. The Adoption is straightforward every time a set of records is tested for privacy fulfillment check m-privacy with respect to. C. The Binary Space Partitioning (BSP) recursively chooses an attribute to split data points in multidimensional domain space until data cannot be split any further without breaching *m*-privacy with respect to. C.

The Features of BSP takes into account the data provider as an additional dimension for splitting uses privacy fitness score as a general scoring metric for selecting the split point. It adapts its *m*-privacy checking strategy for efficient verification.

3.2 K- Anonymity is Set Valued Data

The K- Anonymity is set valued data privacy model consider Let I = {I1, I2... I|I|} be the set of items from which elements of the sets are drawn and Let D = {t1, t2... t|D|} be a transactional database over I where each transaction within D is a non-empty subset of I. The Equivalence class in transactional database D consists of a multi set of transactions. An equivalence class for D is the set of all transactions with identical sets of items S. The k-anonymity in set-valued data transactional database D is k-anonymous if every transaction in D occurs at least k times, or equivalently the size of each equivalence class in D is at least k. The Transactional database is k-anonymous if each transaction is identical to at least k - 1 others. The states that given any m or fewer items chosen from any transaction there are at least k-1 other transactions containing same set of m items.

4. CONCULSION

This performed a tradeoff analysis of system methodologies utilizing M- Partition Privacy to answer user queries. Finally, applied our analysis results to the design of a M – Partition Privacy algorithm to identify and apply the best design parameter settings in J2EE. Then implemented the proposed scheme, and conducted comprehensive performance analysis and evaluation, which showed its efficiency and advantages over existing schemes.

REFERENCES

[1] Olvi L. Mangasarian, “Privacy-Preserving Horizontally Partitioned Linear Programs”.2003.

[2] P. Krishna Prasad and C. Pandu Rangan “Privacy Preserving BIRCH Algorithm for Clustering over Arbitrarily Partitioned Databases”.

[3] Ali Inan, Yücel Saygin, Erkay Sava, Ayça Azgin Hinto lu, Albert Levi “Privacy Preserving Clustering on Horizontally Partitioned Data”.

[4] Jaideep Vaidya and Chris Clifton “Privacy-Preserving Decision Trees over Vertically Partitioned Data”

[5] Zhiqiang Yang and Rebecca N. Wright, “Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data” IEEE
TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 9, SEPTEMBER 2006

[6] Khuong Vu and Rong Zheng Jie Gao “Efficient Algorithms for K-Anonymous Location Privacy in Participatory Sensing” 2012
Proceedings IEEE INFOCOM

[7] Sebastian Schrittwieser, Peter Kieseberg, Isao Echizen, Sven Wohlgemuth, Noboru Sonehara, and Edgar Weippl “An Algorithm for k-anonymity-based Fingerprinting”

[8] S. Goryczka, L. Xiong, and B. C. M. Fung, “m-Privacy for collaborative data publishing,” in Proc. of the 7th Intl. Conf. on Collaborative

Author Profiles

R.ARPITHA
M.Tech Scholar,
CSE Dept,
Chadalawada Ramanamma
Engineering College, Tirupati.
A.P, India.

K. SIRISHA(Asst. Prof).,
Dept. of Computer Science *and Engineering*
Chadalawada Ramanamma Engineering College
Tirupathi, India