Multiparty differential privacy over vertically partitioned data

Publication History
Received: 23 January 2015
Accepted: 3 March 2015
Published: 19 April 2015

Citation
Uماماهسواری E, Mary Selvan. Multiparty differential privacy over vertically partitioned data. Discovery, 2015, 30(134), 347-354
MULTIPARTY DIFFERENTIAL PRIVACY OVER VERTICALLY PARTITIONED DATA

E. UMAMAHESWARI¹, MARY SELVAN²

¹II M.E. Department of Computer Science & Engineering, St. Peter’s College of Engineering & Technology, Chennai, 600054, India
²Assistant Professor, Department of Computer Science & Engineering, St. Peter’s College of Engineering & Technology, Chennai, 600054, India

Abstract- Data integration methods enable different data providers to flexibly integrate their expertise and deliver highly customizable services to their customers. Data mining of private data is one of the key to success for an organization, it is challenging task to implement data mining in distributed database. Collaboration of different organization brings mutual benefits to the party involved. So different organizations want to collaborate and execute efficient data mining algorithm. Privacy preserving data publishing addresses the problem of disclosing sensitive data where different attributes for the same set of individuals are held by multi parties. Differential privacy guarantees strongest privacy among multi parties. To ensure differential privacy, distributed Gaussian mechanism is proposed for vertically partitioned data for multi parties in the trusted third party model. To achieve differential privacy noise are added with true function using Gaussian mechanism. Random value protocol and Yao’s protocol are used to provide the security, according to the definition of secure multiparty computation.

Keywords- Differential privacy, secure data integration, classification analysis

I. INTRODUCTION

1.1 OVERVIEW

Data mining technology is a means of knowledge discovery to efficiently large quantities of data to find interesting patterns and trends. Mining contains various algorithms such as association rule mining, clustering, classification and ranking. Data mining classification algorithms are centralized algorithm and works on centralized database. In this information age, organization uses distributed database. In business environment, data sharing is an essential requirement for making better decisions and providing high quality services. Collaboration is the important key to success. It brings mutual benefit and heavy results that helps in decision making. In distributed database it describes two concerns. First one is privacy preservation of private data. Second one is to extract efficient result after running the multiparty privacy algorithm over vertical portioned data in a secure way. Data distribution among various sites can be achieved through horizontally or vertically portioned data. Vertically partitioned data in which each party collects different information about the same set of entities. For example bank and insurance company might gather different information about the same set of people. In vertical partitioned data, a portion of each instance is presented at each site. But no site has complete information about any entity. Privacy preservation of private data can be achieved by adding laplacian noise with true count function. But before add the laplacian noise each party should be anonymizing their private data. Ensure security by using cryptographic techniques on integrated data set.

1.2 DISTRIBUTED DATABASE

A distributed database is a collection of data which logically belongs to the same system but is distributed over the sites in a computer network. It is divided into three types such as 1. Horizontal partitioning of data: Data can be horizontally partitioned among different parties over the same set of attributes. It involves putting different rows into different tables. Perhaps customers with ZIP codes less than 50000 are stored in Customers East, while customers with ZIP codes greater than or equal to 50000 are stored in Customers West. The two partition tables are then Customers East and Customers West, while a view with a union might be created over both of them to provide a complete view of all customers. 2. Vertical partitioning of data vertically partitioned data in which each party collects different information about the same set of entities. Vertical partitioning involves creating tables with fewer columns and using additional tables to store the remaining columns. Normalization also involves this splitting of columns across tables, but vertical partitioning goes beyond that and partitions columns even when already normalized. 3. Mixed partitioning of data: Database is first horizontally partitioned then vertically partitioned or vice versa. Mixed partition is sometimes called arbitrary partition. This paper concern vertically partitioned data only; it selects data column by column instead of selecting row by row.
1.3 PRIVACY PRESERVING TECHNIQUE

Privacy preserving is a very active research area in data mining. Discovering knowledge through a combination of different databases raises security issue [1]. Although data mining results usually do not violate privacy of individuals, it cannot be assured that an unauthorized person will not access the data is partitioned over different sites and data is not encrypted, it is impossible to derive new knowledge about the other sites. Data mining techniques try to identify regularities in data, which are unknown and hard to discover by individuals. Regularities or patterns are to be revealed over the entire data, rather than on individuals. To find such disclosure of patterns, the mining process has to access and use individual information. The scalar product protocol allows more than two parties for computation. The main goal of this protocol is to secure the private data of other parties such that a party can know its own result and data.

1.4 DIFFERENTIAL PRIVACY

Differential privacy aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records[2]. Consider a trusted party that holds a dataset of sensitive information (e.g. medical records, voter registration information, email usage) with the goal of providing global, statistical information about the data publicly available, while preserving the privacy of the users whose information the data set contains. Such a system is called a statistical database. The notion of indistinguishability, later termed differential privacy, formalizes the notion of “privacy” in statistical databases. Many differentially private algorithms rely on adding controlled noise to functions with low sensitivity. Differential privacy is designed to protect the privacy between neighboring databases which differ only in one row. This means that no adversary with arbitrary auxiliary information can know if one particular participant submitted his information. However this is also extendable if they want to protect databases differing in rows, which amounts to adversary with arbitrary auxiliary information can know if particular participants submitted their information.

II. LITERATURE REVIEW

2.1 ANONYMIZATION

The centralized anonymization method can be viewed as “integrate then generalize” approach, where the general government health agency first integrates the data from different hospitals then performs generalization. The distributed anonymization method can be viewed as “generalize and integrate” approach, to achieve distributed anonymization each hospital anonymize the patient data independently and then integrate.

Anonymization is to generalize the records into equivalence groups so that each group contains at least k records with respect to some quasi identifiers attributes (QID). When the number of quasi identifier is large that is the dimensionality of data is high, most of the data is suppressed in order to achieve K-anonymity, resulting in poor data quality for data analysis. In order to overcome this bottleneck LKC privacy model is used for high dimensional health care data [3]. LKC –Privacy is ensure that every combination of values in QID with maximum length L in the data table T is shared by at least k records.

2.2 DIFFERENTIAL PRIVACY OF DATA

Anonymization algorithm for the non interactive setting based on the generalization technique. The proposed differential generalization algorithm first generalizes the raw data and then adds noise to guarantee differential privacy. In generalization each record represents the information of an individual with attributes [2]. To anonymize a data set D, generalization replaces a value of an attribute with a more general value. The exact general value is determined according to the attribute partition, categorical and numerical partition is the two types of attribute partition in the generalization techniques. The general idea is to anonymize the raw data by a sequence of specializations, starting from the topmost general state. The specialization process can be viewed as pushing the “cut” of each taxonomy tree downwards and the cut contains exactly one value on each root to leaf path. After that Noise is generated according to the Laplace distribution with probability density function.

2.3 SECURE DATA INTEGRATION

In business environment data sharing is an essential requirement for making better decisions and providing high quality services. While data sharing can help their clients obtain the required information or explore new knowledge, it can also be misused by adversaries to reveal sensitive information that was not available before the data integration. In order to overcome this problem, they proposed two algorithms for secure data integration [7]. First is algorithm for semi honest parties, goal of this algorithm to find the winner candidate by using top down specialization approach to generalize a single table. Second is algorithm for malicious parties, in this algorithm using trusted third party to verify the integrity of data. Each party generalizes its own attributes independently and then provides the anonymous data set to trusted third party to integrate the data securely.
2.4 DECISION TREE CLASSIFIER

A decision tree is a popular classification method in data mining. The most important feature of decision tree classifier is their ability to break down a complex decision making process into a collection of simpler decisions. Privacy preserving c4.5 decision tree classification algorithm based on secure set union protocol without third party [14]. In secure union methods, each party needs to give their rules without revealing the owner. The union of items can be evaluated using SMC protocol method. All parties locally generate their public key pair for a commutative encryption scheme and then distribute their public key to every other participant. In secure set intersection protocol, each party encrypts its items with its key and passes it along to the other parties. On receiving the set of items, a party encrypts each item and permutes the order before sending it to the next party. This process repeated until every item has been encrypted by every party and then counts the no of values. That is present in all of the encrypted item sets. This process can be done by any party. None of the parties is able to know the items in the intersection set because of the encryption.

2.5 VERTICAL PARTITIONED DATA

Data can be vertically partitioned among different parties over the different set of attributes in same entities [5]. In CART algorithm root node has been created and then computes gini index for all attribute present in data set, after that attribute of minimum gini index is selected as the attribute maximizes the impurity reduction. Each record has been divided in between two parties and both parties share the class labels of all records then determine whether the both parties remain the same single class or not. If they belong to the same single class, then returns the leaf node with that specific class value. Both parties share the class labels of all records and the names of all attributes, so they both know whether list is empty or not. If list is empty, just scan dataset and stasitise the most frequent class, making the leaf with the most frequent class label. After that queue is initialized to contain the root node while queue is not empty then pop out the first node from queue and then evaluate the gini index for each attribute to find the best split attribute, if the split attribute is continual then find its partition value for best split attribute. calculate the classified mistakes of each node and carry on the tree pruning.

2.6 DISCOVER FREQUENT PATTERN

Discovering frequent pattern from data is a popular exploratory technique in data mining. Frequent item set mining is a fundamental problem in data mining. In order to overcome this problem two efficient algorithms are proposed for discovering the k most frequent patterns in a data set of sensitive records. Exponential mechanism based algorithm takes the transaction data set and then preprocessing the data set using FIM algorithm. In sampling the truncated frequencies are used to sample K item sets and then perturb the true frequencies of the itemset sampled by adding Laplace noise. Basic idea of Laplace mechanism based algorithm is to add independent Laplace noise to the frequencies of all itemsets and select the K itemsets with the highest perturbed frequencies [12].

2.7 NAÏVE BAYES CLASSIFIER

Classification is a popular data mining technique used to predict group membership for data tuples. In classification rule mining, a set of database tuples act as a training sample and it is analyze to produce a model of the data or classifier that can be used for classifying a new tuple. The goal of privacy preserving classification is to build precise classifiers without disclosing personal information in the data being mined. In naïve bayes classifier has three layers such as input layer, intermediate layer, output layer. In input layer all participating parties that are involved in the classification process individually calculate probability or model parameters for all class value of each attribute value for every attribute. In intermediate layer no party has all the attributes. They must collaborate to find total probability for all classes. For this they proposed secure multiplication protocol, no party is able to know the probabilities or model parameter, not even data of the other parties. Only first party will know the probability for all classes. Based on the total probability of all class value, first party will find the class with the highest total probability and finally classify the new tuple. Send this class value to all other parties. Protocol is secure to classify each new tuple [9].

III.PROPOSED METHOD

Distributed Gaussian mechanism is proposed based on the trusted third party model for vertically partitioned data among multi parties to achieve a differential privacy. The Gaussian mechanism anonymize the raw data using generalization technique. To achieve differential privacy noise is added with anonymous data using Gaussian distribution. The system uses homomorphic encryption scheme to provide secure way integration according to the definition of secure multi party computation (SMC).

The system architecture is shown in figure 3.1. Data can be vertically partitioned among different parties over the different set of attributes in same entities. These distributed data can be integrated for making better decisions and providing high quality services. Each party generates local anonymous database using anonymous technique. To anonymize a data set, generalization replaces a value of an attribute with a more
general value. The exact general value is determined according to the attribute partition, which is divided into two types: 1. categorical attributes partition, 2. Numerical attributes partition. Categorical attribute partitions are defined by a set of nodes from the taxonomy tree such that it covers the whole tree, and each leaf node belongs to exactly one partition.

Figure 3.1: System Architecture

Gaussian distribution mechanism first computes the true output and then perturbs the output by adding noise. The noise is generated according to the gaussian distribution with probability density function; its variance is $\Omega^2$ and its mean is $\mu$. The gaussian mechanism guarantees the perturbed output to satisfy the differential privacy in gaussian distribution variance. The gaussian variance is considered as privacy parameter that determines the magnitude of noise.

System modules:

- Generalization of Data
- Privacy Model
- Security Model
- Performance Analysis

1. Generalization of Data

The generalization replaces the specific value with more general value to make the information less precise while preserving truthfulness of the information. The generalization of data is done based on the attributes values. The purpose of generalization is used to generate anonymous data from raw data. The exact general value is determined according to the attribute partition, which is divided into two types: 1. categorical attributes partition, 2. Numerical attributes partition. Categorical attribute partitions are defined by a set of nodes from the taxonomy tree such that it covers the whole tree, and each leaf node belongs to exactly one partition.

2. Privacy Model

Differential privacy provides one of the strong privacy guarantees. Standard mechanism to achieve differential privacy is to add random noise to the true output of the function. Random noise is generated according to the Gaussian distribution with probability density function; its variance is $\Omega^2$ and its mean is $\mu$. The Gaussian mechanism guarantees the perturbed output to satisfy differential privacy. In Gaussian distribution variance is considered as privacy parameter that determines the magnitude of noise.

3. Security Model

Homomorphic encryption scheme is used to provide secure way integration according to the definition of secure multi party computation (SMC). Each party encrypts the information using their own private key. All parties generate their public key pair and distribute their key to trusted third party. Random value protocol is used to generate random values for secure multiparty computation.

4. Performance Analysis

Distributed exponential mechanism is used for providing differential privacy for two party interactions in semi honest
adversary model. Distributed Gaussian mechanism provides differential privacy for multi party’s interaction in trusted third party model. Performance analysis is calculated between distributed exponential mechanism and distributed Gaussian mechanism.

IV.IMPLEMENTATION

Distributed gaussian mechanism is used to achieve differential privacy. First step of this algorithm is anonymizing the raw data using top-down specialization (TDS) approach. Initially in TDS all values are generalized to the topmost value in its taxonomy tree and cut contains the topmost value for each attribute. At the each iteration TDS performs the best specialization that has the highest general value among the partitioned data. The exact general value is determined according to the attribute partition, which is divided into two types: 1. categorical attributes partition, 2. Numerical attributes partition. Categorical attribute partitions are defined by a set of nodes from the taxonomy tree such that it covers the whole tree, and each leaf node belongs to exactly one partition. For example artist is the general value of dancer according to the taxonomy tree of job. Due to specializing on a single attribute depends only on that attribute and class attribute, each party knows about class and attributes they have.

Second step of this algorithm is to ensure \( \epsilon \)-differential privacy, which guarantees that an adversary learns nothing more about an individual, regardless of whether her record is present or absent in the data. A standard mechanism to achieve differential privacy is to add a random noise to the true output of a function. The noise is calibrated according to the sensitivity of the function. The sensitivity of a function is the maximum difference of its outputs from two data sets that differ only in one record. Distributed gaussian mechanism takes a data set \( D \), a function \( f \), and the parameter \( \Omega \) that determines the magnitude of noise as input. Gaussian distribution mechanism first computes the true output \( f(D) \) and then perturbs the output by adding noise. The noise is generated according to the gaussian distribution with probability density function; its variance is \( \Omega^2 \) and its mean is 0. The gaussian distributed mechanism guarantees that perturbed output \( f(D) = f(D) + \text{gau}(\Delta f/\Omega) \) satisfies \( \epsilon \)-differential privacy, where \( \text{gau}(\Delta f/\Omega) \) is a random variable sampled from the gaussian distribution.
CONCLUSION

In this paper the distributed gaussian mechanism for multi-parties over vertically portioned data has been done. The distributed gaussian mechanism achieves differential privacy and supports effective classification analysis. The proposed solution connects the classical generalization technique with output perturbation to effectively anonymize raw data. Generalization replaces the specific value with more general value to make the information less precise while preserving truthfulness of the information.

REFERENCES


