PRIVACY PRESERVING MINING ASSOCIATION RULE FROM OUTSOURCED TRANSACTION

Authors: MOHANRAJ GOVINDARAJ & SARITHA KARTHIK

Abstract:
Spurred by developments such as cloud computing, there has been considerable recent interest in the paradigm of data mining-as-a-service. A company (data owner) lacking in expertise or computational resources can outsource it’s to a third party service provider (server). However, both the items and the association rules of the outsourced database are considered private property of the corporation (data owner). To protect corporate privacy, the data owner transforms its data and ships it to the server, sends mining queries to the server, and recovers the true patterns from the extracted patterns received from the server. In this paper, we study the problem of outsourcing the association rule mining task within a corporate privacy-preserving framework. We propose an attack model based on background knowledge and devise a scheme for privacy preserving outsourced mining. Our scheme ensures that each transformed item is indistinguishable, w.r.t. the attacker’s background knowledge, from at least k-1 other transformed items. Our comprehensive experiments on a very large and real transaction database demonstrate that our techniques are effective, scalable, and protect privacy.

INTRODUCTION

Today there is large amount of data proceed in every day from different sources. That large amount of data stored in different database. This data store in storage devices in from of row data. Data mining is the process of discovering interesting pattern and knowledge from large amount of data. Following Example where data mining techniques are used are Direct mail marketing, bioinformatics, credit card fraud detection, text analysis and market basket analysis. Extracting knowledge from row data, There Database Amish Desai Computer Science and Engineering Assi.Prof ParullInstitute of Technology, Vadodara Gujarat,INDIA-39 1 760 desaiamishI986@gmail.com is some technique to deal with security. Privacy preserving in data mining is one of the technique that deal with security of the knowledge that extracted by data mining technique. There are various Data Mining Tasks: Classification Clustering Association Rule Mining Sequential Pattern Mining Regression.

PRIVACY PRESERVING DATA MINING

Data mining is the process of gathering information about the user specific data, also called knowledge discovery, on internet. The problem with data mining output is that it also discloses some information, which is considered to be private and personal. Effortless access to such personal data causes a peril to individual privacy [9]. Recent research in the area of privacy preserving data mining has considerate effort to determine a trade-off between privacy and the need for knowledge discovery, which is necessary in order to improve decision-making processes and other human activities. PPDM cope with the problem of learning accurate models over aggregate data, while protecting privacy at the level of individual records. The main purpose of privacy preserving data mining is to design competent frameworks and algorithms that can extract relevant knowledge from a large amount
of data without revealing of any sensitive information [9]. It protects sensitive information by providing sanitized database of original database on the internet or a process is used in such a way that private data and private knowledge remain private even after the mining process. It is PPDM due to which the benefits of data mining be enjoyed, without compromising the privacy of concerned individuals.

ASSOCIATION RULE MINING

Association rule mining one of the task of the data mining. Association rule mining is important field to under privacy preserving data mining. R. Agrawal was first proposed the basic concept of the Association rule mining. Association rule is basically using the concept of IF-THEN relationship among the different data. Following example of shows the concept of the Association rule. "If the customer buy a laptop, then he/she is 85% likely to also purchase anti-virus ". Analysis of the above example that laptop is somewhat related to anti-virus because every time customer buy a computer then he/she buy anti-virus. Association rule is used for market basket analysis. Let \( I = \{I_1, I_2, ..., I_n\} \) be a set of item. Let \( D \) be a database of transactions where each transaction \( T \) is a set of item such that \( T \) belongs to \( I \).

For every transaction is associated to an identifier, called TID. A transaction \( T \) is contain \( A \) if and only if \( A \) belongs to \( T \). An association rule is applied of the form \( A-B \). Where \( A \) and \( B \). And \( AB \) belongs to \( <1> \). Every association rule must be satisfy two contain support and confidence. Support of rule \( A-B \) is the transaction database that contain support count of \( AUB \). Support for rule \( (A-B) \) can be calculated using below formula in (1).

\[
\text{IAUBI} \text{support (A-B)} \in D
\]

Confidence of rule \( A-B \) is the transaction database that contain \( A \) also contain \( B \). The confidence for rule \( (A-B) \) can be calculated using below formula in (2). Confidence

ASSOCIATION RULE HIDING

Association rule hiding is one technique to PPDM (Privacy Preserving Data Mining). Association rule hiding methodology aim is to sanitize the original data. so it may be applied to following condition: (1) sanitized database is not reveal any sensitive rules. (2) Sanitized database is mining of all non-sensitive rules. (3) Sanitized database is not add any new rules, not present in database \( D \). Association rule hiding is the depend on support and confidence of the rule, There is two way to hide any rule (i) Decrease support up to certain threshold. (ii) Decrease confidence up to certain threshold. Related Works

There are many methodologies used for maintaining privacy in transaction database. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language can be used for
developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into account for developing the proposed system.

This section provides background to the research through a review of some of the literature on privacy. The literature review is focused on those areas central to the scope of this research.

What is privacy? It is an almost customary feature of any analysis of privacy to begin with a disclaimer about the inherent difficulty of defining exactly what ‘privacy’ is and disaggregating its various dimensions. It is something that is taken for granted and most people would have a sense of what privacy is but have difficulty putting it into words. The concept and meaning of privacy has long been debated by philosophers, social scientists, academic lawyers and other scholars. All definitions, to some extent, are based on assumptions about individualism and about the distinction between the realms of civil society and the state. However, many gloss over essential cultural, class-related and gender differences. Literature on privacy tends to give readers an overwhelming sense that privacy is a deeply contested concept, which often varies according to context and environment. (Bennett & Grant, 1999)

According to Bennett and Raab (2003), in Western culture, the modern claim to privacy and the contemporary justification for information privacy as a public policy goal was derived from a notion of a boundary between the individual and other individuals, and between the individual and the state. This concept of privacy rests on a construct of society as comprising relatively autonomous individuals and on notions of differences between the privacy claims and interests of different individuals. According to John Stuart Mill (as cited in Bennett & Raab, 2003), there should be certain ‘self-regarding’ activities of private concern, contrasted with ‘other-regarding’ activities to community interest and regulation. Shils (as cited in Bennett & Raab, 2003) argued that privacy is essential for the strength of American pluralistic democracy because it bolsters the boundaries between competing and countervailing centres of power. Dr Alan Westin, a leading academic (whose book Privacy and Freedom has shaped virtually all current thinking about privacy as a public issue), reinforced the importance of privacy for liberal democratic societies – in contrast to totalitarian regimes:

A balance that ensures strong citadels of individual and group privacy and limits both disclosure and surveillance is a prerequisite for liberal democratic societies. The democratic society relies on publicity as a control over government, and on privacy as a shield for group and individual life.

Westin also addresses the specific functions that privacy plays. It promotes freedom of association. It shields scholarship and science from unnecessary interference by government. It permits the use of a secret ballot and protects the voting process by forbidding government surveillance of a citizen’s past voting record. It restrains improper police conduct such as unreasonable search and seizure. It also serves to shield those institutions, such as the press, that operate to keep government accountable.
In a seminal law review article Samuel Warren and Louis Brandeis (1890) defined privacy simply as “the right to be let alone” – to go about life free from unreasonable interference by external forces.

Privacy has also been defined comprehensively:

Privacy is a concept related to solitude, secrecy, and autonomy, but it is not synonymous with these terms; for beyond the purely descriptive aspects of privacy as isolation from the company, the curiosity, and the influence of others, privacy implies a normative element: the right to exclusive control of access to private realms… the right to privacy asserts the sacredness of the person;… any invasion of privacy constitutes an offence against the rights of the personality – against individuality, dignity, and freedom. Arnold Simmel.

Privacy can be divided into the following facets

Territorial privacy – concerning the setting of limits on intrusion into the domestic and other environments such as the workplace or public space.

- Privacy of the person – this is concerned with protecting a person against undue interferences such as physical searches and drug testing, and information that violates his or her moral sense;
- Privacy of communications, covering the security and privacy of mail, telephones, email and other forms of communication;
- Privacy in the information context – this deals with the gathering, compilation and selective dissemination of personal information such as credit data and medical records.

The discourse on privacy as a policy issue has largely focused on information privacy and it is this facet of privacy that this research project will focus on. In this sense, privacy can be defined as “the claim of individuals, groups or institutions to determine for themselves when, how and to what extent information about them is communicated to others.” (Westin, 1967, p7).

However, the rise to prominence of Internet communications and e-commerce has led to privacy of communications (and transmission) attracting more attention and concern. The increased concern with privacy of communications has caused some confusion between the meanings of information privacy and information security and the terms are often used interchangeably. As Clarke noted (as cited in Bennett & Raab, 2003), the term ‘privacy’ is used by some people to refer to the security of data or security of data during transmission as protection against various risks, such as data being accessed or modified by unauthorised persons. These aspects, however, are only a small fraction of the considerations within the field of ‘information privacy’. That is, data security is a necessary but not sufficient condition for information privacy. An organisation might keep the personal information it collects highly secure, but if it should not be collecting that information in the first place, the individual’s information privacy rights are clearly violated.

Proposed Work

The particular problem attacked in our paper is outsourcing of pattern mining within a corporate privacy-preserving framework. A key distinction between this problem and the abovementioned PPDM problems is that, in our setting, not only the underlying data but also the mined results are not intended for sharing and must remain private. In particular, when the server possesses background knowledge and
conducts attacks on that basis, it should not be able to guess the correct candidate item or item set corresponding to a given cipher item or item set with a probability above a given threshold. We proposed to solve this problem by using \( k \)-privacy, i.e., each item in the outsourced dataset should be indistinguishable from at least \( k-1 \) items regarding their support. In this paper, our goal is to devise an encryption scheme which enables formal privacy guarantees to be proved, and to validate this model over large-scale, real-life transaction database.

This work is to devise encryption schemes such that formal privacy guarantees can be proven against attacks conducted by the server using background knowledge, while keeping there source requirements under control.

System Models

THE PATTERN MINING TASK

The reader is assumed to be familiar with the basics of association rule mining. We let \( I = i_1, ..., i_n \) be the set of items and \( D = t_1, ..., t_m \) a transaction database (TDB) of transactions, each of which is a set of items. We denote the support of an item set \( S \subseteq I \) as \( \text{supp}(D(S)) \) and the frequency by \( \text{freq}(D(S)) \). Recall, \( \text{freq}(D(S)) = \text{supp}(D(S))/|D| \). For each item \( i \), \( \text{supp}(D(i)) \) and \( \text{freq}(D(i)) \) denote respectively the individual support and frequency of \( i \). The function \( \text{supp}(D) \) projected over items, is also called the **item support table**. The well-known frequent pattern mining problem: given a TDB \( D \) and a support threshold \( \sigma \), find all item sets whose support in \( D \) is at least \( \sigma \). In this paper, we confine ourselves to the study of a (corporate) privacy preserving outsourcing framework for frequent pattern mining.

PRIVACY MODEL

We let \( D \) denote the original TDB that the owner has. To protect the identification of individual items, the owner applies an encryption function to \( D \) and transforms it to \( D^* \), the encrypted database. We refer to items in \( D \) as **plain items** and items in \( D^* \) as **cipher items**. The term item shall mean plain item by default. The notions of plain item sets, plain transactions, plain patterns, and their cipher counterparts are defined in the obvious way. We use \( I \) to denote the set of plain items and \( E \) to refer to the set of cipher items.

Attack Model

The server or an intruder who gains access to it may possess some background knowledge using which they can on the encrypted database \( D^* \). We generically refer to these agents as an **attacker**. We adopt a conservative model and assume that the attacker knows exactly the set of (plain) items \( I \) in the original transaction database \( D \) and their true supports.

We assume the service provider (who can be an attacker) is **semi-honest** in the sense that although he does not know the details of our encryption algorithm, he can be curious and thus can use his background knowledge to make inferences on the encrypted transactions. We also assume that the attacker always returns (encrypted) item sets together with their exact support. The data owner (i.e., the corporate) considers the true identity of:

1. every cipher item,
(2) every cipher transaction, and
(3) every cipher frequent pattern as the intellectual property which should be protected.
We consider the following attack model

- **Item-based attack:**
  The semi honest service provider can attack the owners data depend upon the single item identity.

- **Set-based attack:**
  The service provider attack the owners data depend upon the many item identities. In this method the attacker can easily attacks the data correctly but they can’t use that data because that data’s are in ciphertext form.

Data owners are using the separate E/D Module.

**ENCRYPTION/DECRYPTION SCHEME:**

**Encryption:**

In this section, we introduce the encryption scheme, which transforms a TDB D into its encrypted version D*. Our scheme is parametric w.r.t. k > 0 and consists of three main steps: (1) using 1-1 substitution ciphers for each plain item; (2) using a specific item k-grouping method; (3) using a method for adding new fake transactions for achieving k-privacy.

The constructed fake transactions are added to D(once items are replaced by cipher items) to form D*, and transmitted to the server.

**Decryption:**

When the client requests the execution of a pattern mining query to the server, specifying a minimum support threshold σ, the server returns the computed frequent patterns from D*. Clearly, for every item set S and its corresponding cipher item set E, we have that supp D(S) ≤ supp D_\(E\). For each cipher pattern E returned by the server together with supp D_(E), the E/D module recovers the corresponding plain pattern S. It needs to reconstruct the exact support of S in D and decide on this basis if S is a frequent pattern. To achieve this goal, the E/D module adjusts the support of E by removing the effect of the fake transactions. Supp D(S) = supp D_\(E\)-supp D_\(\neg D\)(E). This follows from the fact that support of an item set is additive over a disjoint union of transaction sets. Finally, the pattern S with adjusted support is kept in the output if supp D(S) ≥ σ. The calculation of supp D_\(\neg D\)(E) is performed by the E/D module using the synopsis of the fake transactions in D* \ D.

**CONCLUSION**

We proposed a protocol for secure mining of association rules in horizontally distributed databases that improves significantly upon the current leading protocol in terms of privacy and efficiency. One of the main ingredients in our proposed protocol is a novel secure multi-party protocol for computing the union (or intersection) of private subsets that each of the interacting players holds. Another ingredient is a protocol that tests the inclusion of an element held by one player in a subset held by another. Those protocols exploit the fact that the underlying problem is of interest only when the number of players is greater than two. One research problem that this study suggests, namely, to devise an efficient protocol for inequality verifications that uses the existence of a semi honest third party. Such a protocol might enable to further improve upon the communication and computational costs of the second and third stages of the protocol. Other research problems that this study suggests is the implementation of the
techniques presented here to the problem of distributed association rule mining in the vertical setting the problem of mining generalized association rules, and the problem of subgroup discovery in horizontally partitioned data.

REFERENCES