Information Sciences 267 (2014) 267–286

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Towards semantically secure outsourcing of association rule mining on categorical data

Junzuo Lai [a,b,*], Yingjiu Li [a], Robert H. Deng [a], Jian Weng [a,b], Chaowen Guan [a], Qiang Yan [a]

[a] Department of Computer Science, Jinan University, China
[b] School of Information Systems, Singapore Management University, Singapore

ARTICLE INFO

Article history:
Received 23 October 2012
Received in revised form 24 July 2013
Accepted 26 January 2014
Available online 1 February 2014

Keywords:
Association rule mining
Outsourcing
Semantic security
Privacy
Soundness

ABSTRACT

When outsourcing association rule mining to cloud, it is critical for data owners to protect both sensitive raw data and valuable mining results from being snooped at cloud servers. Previous solutions addressing this concern add random noise to the raw data and/or encrypt the raw data with a substitution mapping. However, these solutions do not provide semantic security; partial information about raw data or mining results can be potentially discovered by an adversary at cloud servers under a reasonable assumption that the adversary knows some plaintext–ciphertext pairs. In this paper, we propose the first semantically secure solution for outsourcing association rule mining with both data privacy and mining privacy. In our solution, we assume that the data is categorical. Additionally, our solution is sound, which enables data owners to verify whether there exists any false data in the mining results returned by a cloud server. Experimental study shows that our solution is feasible and efficient.

1. Introduction

Data mining is a very popular research area in the field of computer science. The objective of data mining is to extract comprehensible, useful, and non-trivial knowledge from large datasets. Association rule mining, introduced in [1], which aims at finding close relationships between items in transactional data, has been the major focus in data mining research with enormous applications such as market basket analysis, network intrusion detection, and inventory control [2,17,20,26,36,48].

In the cloud computing era, data owners who lack expertise or computational resources may outsource association rule mining to cloud service providers. While outsourcing has the potential of reducing computation and software cost, it is critical to protect both sensitive raw data, which is the database of transactions, and valuable mining results from being snooped at cloud servers. To address this concern, data owners need to encrypt the original data using a suitable algorithm such that the mining tasks can be performed by cloud servers directly on the encrypted data. Though this topic has been intensively researched since the introduction of privacy preserving data mining [3], none of the previous works on outsourcing of association rule mining [14,33,41,50] ensure semantic security [16]. Informally, semantic security states that an adversary (at cloud servers) can gain no partial information about the target data even the adversary has the capability of

* Corresponding author at: Department of Computer Science, Jinan University, China.
E-mail addresses: junzuolai@smu.edu.sg (J. Lai), yjli@smu.edu.sg (Y. Li), robertdeng@smu.edu.sg (R.H. Deng), cryptjweng@gmail.com (J. Weng), cwguan@smu.edu.sg (C. Guan), qiang.yan.2008@phdis.smu.edu.sg (Q. Yan).

http://dx.doi.org/10.1016/j.ins.2014.01.040
0020-0255/

obtaining adaptively plaintext–ciphertext pairs. In this paper, we investigate semantically secure solutions for efficient outsourcing of association rule mining. To the best of our knowledge, this is the first attempt in this area.

Initial researches in association rule mining [1,2,18,19] have assumed that the data is categorical. We also start the study of semantically secure outsourcing of association rule mining from the categorical data. In practice, databases may contain much quantitative data and are not limited to categorical items only. Different evolutionary algorithms and especially genetic algorithms [4–6,8,21,25,30,35,43–45] have been proposed to extract association rules in quantitative data, and how to extend these algorithms to construct semantically secure outsourcing of association rule mining on quantitative data will be left as our future work. As most association rule mining algorithms [1,2,18,19] on categorical data, we split the problem of association rule mining into two subproblems. The first is to find all of the frequent itemsets in database. The second is to find the association rules from the discovered frequent itemsets. As showned in [1], it is straightforward to solve the second subproblem after the first subproblem is resolved. In this paper, we focus on the first subproblem.

It had been difficult to design a semantically secure solution for outsourcing association rule mining until the emergence of fully homomorphic encryption [12]. Since fully homomorphic encryption [12] enables arbitrary functions to be computed on encrypted data, it can be directly applied to association rule mining. However, we will show that this solution is extremely inefficient and thus impractical. In addition, it requires that most of the data mining task be performed by data owners rather than by cloud servers, which is contrary to the purpose of outsourcing.

In our method, the dominant task of association rule mining is to discover all frequent itemsets (i.e., itemsets whose frequencies appearing in transactions are higher than a pre-defined threshold). The basic operation for discovering all frequent itemsets is to decide whether or not a transaction contains an itemset. We discover that this basic operation can be transformed to an inner product operation, and it is possible to adopt the notion of symmetric-key predicate-only encryption for inner products [40] in the design of semantically secure and efficient schemes for outsourcing association rule mining. We notice that the existing symmetric-key predicate-only encryption scheme for inner products [40] is only selectively secure, meaning that the security is proven in a weaker model where part of the challenge strings must be revealed before the attacker receives the public parameters. To remove this constraint, based on the predicate encryption scheme for inner products proposed by [27] and the dual system encryption methodology introduced by [49], we propose a fully secure symmetric-key predicate-only encryption scheme for inner products, which enables us to construct a semantically secure and efficient scheme for outsourcing association rule mining.

Besides addressing the privacy concern for protecting both raw data and data mining results, we also consider the situation in which a cloud server may have incentive to insert false data into true mining results (e.g., for charging more fees). To address this concern, we introduce the soundness requirement so that data owners are able to verify whether the data mining results returned by the cloud server contain any itemsets that are not frequent and if so, the data owners are able to filter out such false data.

1.1. Our contribution

In this paper, we define a formal model of outsourcing association rule mining. Our model consists of four algorithms: SysSetup, AddEncTransRecord, RetrieveTransRecord, and GetFreqItemSets. Algorithm SysSetup is used to output public parameters and data owner's secret key. Algorithm AddEncTransRecord is used to encrypt transaction records before outsourcing data to a cloud server. Algorithm RetrieveTransRecord is not necessarily part of outsourcing association rule mining, and it is for the convenience of data owner to retrieve transaction records from their encrypted forms. Algorithm GetFreqItemSets is a 2-round protocol executed between data owner and a cloud server. In this protocol, data owner sends data mining requests together with necessary token keys to the cloud server, which returns to data owner with encrypted data mining results after performing major data mining tasks on encrypted data. At the end of this protocol, data owner retrieves all frequent itemsets from the encrypted data mining results.

We define three security requirements for outsourcing association rule mining. In addition to protecting raw data (*data privacy*) and mining results (*mining privacy*) under chosen plaintext attacks, we identify another important security requirement: *soundness*, which has not been considered in previous works. In brief, soundness enables data owner to check and filter out any false data from the data mining results returned by a cloud server.

Based on the predicate encryption scheme for inner products proposed by Lewko et al. [27], we propose an efficient solution for outsourcing association rule mining in our model. We prove that our solution achieves data privacy, mining privacy and soundness under chosen plaintext attacks. To the best of our knowledge, it is the first semantically secure solution for outsourcing association rule mining. We also implement our solution and show that it is efficient in experiments.

1.2. Organization

The rest of this paper is organized as follows. Section 2 introduces the related work. In Section 3, we give the preliminary knowledge for designing our solution. In Section 4, we define the model of outsourcing association rule mining and propose a concrete construction under our model with security proofs. In Section 5, we show performance results. Finally, in Section 6, we conclude this paper and point out some future directions.