Information Sciences 444 (2018) 89–104

Contents lists available at ScienceDirect

Information Sciences

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

Differentially private Naive Bayes learning over multiple data sources

Tong Li[a,b], Jin Li[a,*], Zheli Liu[b], Ping Li[a], Chunfu Jia[b,c]

[a] School of Computer Science, Guangzhou University, Guangzhou, China
[b] College of Computer and Control Engineering, Nankai University, Tianjin, China
[c] Information Security Evaluation Center of Civil Aviation, Civil Aviation University of China, Tianjin, China

ARTICLE INFO

Article history:
Received 21 October 2017
Revised 20 February 2018
Accepted 26 February 2018
Available online 27 February 2018

Keywords:
Privacy-preserving
Naive Bayes classification
Differential privacy

ABSTRACT

For meeting diverse requirements of data analysis, the machine learning classifier has been provided as a tool to evaluate data in many applications. Due to privacy concerns of preventing disclosing sensitive information, data owners often suppress their data for an untrusted trainer to train a classifier. Some existing work proposed privacy-preserving solutions for learning algorithms, which allow a trainer to build a classifier over the data from a single owner. However, they cannot be directly used in the multi-owner setting where each owner is not totally trusted for each other. In this paper, we propose a novel privacy-preserving Naive Bayes learning scheme with multiple data sources. The proposed scheme enables a trainer to train a Naive Bayes classifier over the dataset provided jointly by different data owners, without the help of a trusted curator. The training result can achieve ϵ-differential privacy while the training will not break the privacy of each owner. We implement the prototype of the scheme and conduct corresponding experiment.

1. Introduction

Nowadays, the machine learning classifier, as a concrete implementation of classification [5], is becoming an effective tool in data analysis, which has facilitated applications in many areas including economic prediction, risk assessment, and spam detection. Given a queried instance **x**, a trained classifier model **W** can be run to output a prediction that indicates the class which the instance belongs to. To make the prediction accuracy, it is desirable for a trainer to train the classifier over sufficient samples collected from various sources. Unfortunately, privacy and security concerns of personal information arise in recently years, that is, data owners can hard allow untrusted entities to get access to their sensitive data, which restricts trainer's centralization of sample datasets. For example, a medical researcher wants to build a classifier which can be used to classify symptoms according to patients' health records. In this scenario, the researcher is a trainer while health records can be seen as training samples that contains patients' individual sensitive information that should not be revealed by the researcher. It is urgent for this trainer to address a paradox between preserving privacy of patients and keeping the availability of samples.

As a solution, the notion of differential privacy [12,14] has been proposed to provide a privacy guarantee for an analyzed dataset, even in the situation that a trainer hold some prior knowledges about the dataset. Implementing an appropriate

* Corresponding author.
E-mail addresses: litongziyi@mail.nankai.edu.cn (T. Li), lijin@gzhu.edu.cn (J. Li), liuzheli@nankai.edu.cn (Z. Liu), liping26@mail2.sysu.edu.cn (P. Li), cfjia@nankai.edu.cn (C. Jia).

https://doi.org/10.1016/j.ins.2018.02.056
0020-0255/

private mechanism on analysis results of an algorithm can ensure that the algorithm will produce very similar outputs when working on two adjacent datasets. Thus, we can adopting similar methods for the privacy-preserving machine learning to build a classifier over a sample dataset without disclosing any single individual sample in the set. Aiming at a specific learning algorithm, some privacy-preserving schemes [1,48] are developed to achieve the goal of differential privacy.

However, the samples for training are usually collected from multiple data owners rather than a single source. In the multi-owner setting, owners are untrusted for each other, that is, they all try to reveal the data contents of other owners and hope to protect their own privacy. As a result, it is infeasible to directly depute the task of implementing private mechanisms to any owner. A trivial way to solve this training problem is to introduce a trusted curator whose tasks are collecting samples, training a model, and performing the mechanism. Unfortunately, such a curator usually cannot be established in a real world application. This motivates us to design a multi-owner learning scheme without the help of a trusted third party.

Although some powerful cryptographic tools (e.g., fully homomorphic encryption or secure multi-party computation protocol) can be used for these training computations, they are too heavy to be adopted for applications. To carry out multi-owner training results achieving differential privacy in a practical way, we should face challenges from two aspects. On one hand, the sensitivity of the learning function, which is a significant parameter in the mechanism, is related to the whole sample set. That means deriving it without knowing the set is very hard. On the other hand, collecting samples will certainly obtain some of their important characteristics such as counts. So, if there is an untrusted curator, how to prevent it revealing the characteristics is another problem.

1.1. Contribution

To address the problem above, in this paper, we propose a privacy-preserving machine learning scheme in the multi-owner setting for a simple but highly effective classification, the Naive Bayes (NB) classification. The proposed scheme enables a trainer, which is called data receiver, to build a NB classifier over the dataset contributed jointly by different data providers, without the help of a trusted curator. The NB classifier model is able to meet the requirements of ϵ-differential privacy.

In summary, we present the main contributions of this paper as follows.

- To achieve the goal of training, we design novel algorithms for aggregating each data provider's data and implementing differentially private training over these data. These algorithms do not involve heavy cryptographic tools and any trusted curator but can protect each provider's individual privacy.
- Different from existed works, the aggregation method that we design in the proposed scheme can hide some statistics information (e.g., the number of total samples). Furthermore, it makes the scheme guarantee the ownership privacy, which means others cannot know whether a provider holds a sample that contains a specific attribute value.
- We implement the prototype of our scheme and conduct experiments on the LAN server over datasets from UCI Machine Learning Repository [15]. The result shows that the proposed scheme is practical.

1.2. Organization

The rest of this paper proceeds as follows. In Section 2, related works are presented. Some preliminaries of this paper are briefly introduced in Section 3. In Section 4, we state the architecture of our scheme and give the threat model, requirements, and problems. The concrete construction of the scheme are proposed in Section 5. The security analysis for the proposed scheme are presented in Section 6. In Section 7, we present the implementation and evaluate the results of experiments. Finally, we make a conclusion in Section 8.

2. Related work

2.1. Privacy-preserving machine learning

There have been some existing works that address privacy problems for machine learning algorithms, which provide solutions for protecting the privacy of data providers. According to focused problems, the scenarios are mainly divided into two categories. One is classifying a data instance, and the other is training a classifier over the datasets. Our work pays attention on the latter one.

Classification. We briefly describe this category. In the privacy-preserving classification, Barni et al. [4] proposed the secure evaluation based on garbled circuits [22,52] for linear branching programs and neural networks. To deal with a Naive Bayes classifier, some literatures presented schemes for protecting the confidentiality of queried instances [25,46,53]. Based on comparison blocks and argmax blocks, Bos et al. [6] constructed a series of secure classification protocols which contains the NB classification.

Training. The system model of the training usually falls into either *collaborative* or *client-server* style. The addressed problem in the former is how to enable several data providers to collaboratively train a machine learning classifier over all their datasets but not break the privacy of each one. In this setting, each data provider also plays a role of the trainer. The existing works considered this issues for machine learning algorithms, such as the neural network [3,9,40,42,43,54], decision