Computer Networks 133 (2018) 59–72

Contents lists available at ScienceDirect

Computer Networks

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

Privacy-preserving sparse representation classification in cloud-enabled mobile applications

Yiran Shen[a], Chengwen Luo[b], Dan Yin[a,*], Hongkai Wen[c], Rus Daniela[d], Wen Hu[e]

[a] College of Computer Science and Technology, Harbin Engineering University, China
[b] College of Computer Science and Software Engineering, Shenzhen University, China
[c] Department of Computer Science, University of Warwick, UK
[d] Computer Science and Artificial Intelligence Laboratory, MIT, USA
[e] School of Computer Science and Engineering, UNSW Australia, Australia

ARTICLE INFO

Article history:
Received 25 June 2017
Revised 13 November 2017
Accepted 18 January 2018

Keywords:
Sparse representation classification
Mobile computing
Privacy preserving

ABSTRACT

Mobile devices are now pervasive to provide prolific services to the users meanwhile collect the information derived from the activities of the individuals with the onboard sensors. Classification and authentication are popular provided services of many mobile applications which have high probability to involve the sensitive information of the users. In this paper, we propose a new cloud-enabled and Privacy-preserving sparse representation classification ($\mathcal{P}^2$-SRC) system to protect the privacy of both the "data contributors" and "application users" when cloud server is *untrusted*. Different from the state-of-the-art approaches which only consider the attacks on data values, our proposed system, $\mathcal{P}^2$-*SRC*, addresses multiple types of privacy attacks including *Content Privacy Attacks, Source Privacy Attacks* and *Label Privacy Attacks*. As a result, besides the data values, in $\mathcal{P}^2$-SRC, the identities and activities of the users are also protected. According to our evaluations on two different classification applications (face recognition and activity recognition), $\mathcal{P}^2$-SRC achieves almost the same classification accuracy compared with traditional SRC approach which indicates the security add-ons do not affect the accuracy of the SRC classifier. We also demonstrate that it outperforms the most related work, Pickle, significantly on recognition accuracy and privacy protections. Meanwhile the implementation of $\mathcal{P}^2$-SRC in a face recognition application on smartphones demonstrates that $\mathcal{P}^2$-SRC based authentication system accounts for only 0.000041% of the total energy supply of a normal smartphone and the average responding time is around 1.1 s for each recognition request.

1. Introduction

Cloud-enabled mobile applications (CMAs) are booming with the pervasive availability of smart mobile devices (e.g., smartphones, tablets, wearable devices), high speed networks (e.g., Wifi and 4G) and high performance cloud services [19,28]. As CMAs process large amount of crowdsourced data in centralised manner, most of the CMAs utilise the resources of cloud servers to store overwhelming amount of collected data and undertake computationally intensive tasks.

Signals classification is popular in CMAs and it has been studied in the literature for decades [24]. It is the basis of the authentication [9], medical diagnosis [5] and environment awareness systems [25]. Crowdsourcing benefits the classification system especially for learning the classification model (i.e., the classifier). For example, to develop a classification model, the application publisher needs to collect sufficient training samples from certain group of people where crowdsourcing saves the efforts on the data collection. In this paper we define two risky groups of subjects whose privacy can be attacked in the CMAs: the *Data Contributors* and *Application Users*. Data contributors are recruited by application publishers to upload their sensor data of their mobile devices to the cloud for building the training set. The application users make use of the built classification models for recognition/classification.

Innovate CMAs have created many possibilities [4,26,27,30]; however, security issues arise when cloud server is not trusted [46,55]. The uploaded sensor data of mobile devices may contain personal information or be used to infer individuals' private information. To prevent the privacy leakage, many researchers have proposed new privacy-preserving methods [12,25,53]. For instance, Pickle was proposed by Liu et al. [25] to provide certain

* Corresponding Author.
E-mail addresses: shenyiran@hrbeu.edu.cn (Y. Shen), yindan@hrbeu.edu.cn (D. Yin).

https://doi.org/10.1016/j.comnet.2018.01.035
1389-1286/

privacy protection in the classification system when learning the classifier from the encrypted crowdsourced training set. However, Pickle only considered the *content privacy* attacks (sensor values) but ignored the possible *label privacy* and *source privacy* attacks (see definitions next paragraph) though they admitted labels may leak important information.

Three types of attacks In this paper we introduce three types of privacy attacks, i.e., *Content Privacy Attacks, Source Privacy Attacks* and *Label Privacy Attacks*.

- Content privacy attacks are the attacks on drawing actual values of the sensor data of mobile devices.
- Source privacy attacks aim to find the sources or related identities where the sensor data is derived;
- Label privacy attacks are the attacks on obtaining class labels of the training set which allows the adversaries to interpret the classification results and users' activities.

The sources of the sensor data are sensitive information. For example, in a cloud-enabled authentication system, some of the computational burden is shifted to the cloud. The authentication decision is made according to the results computed on the cloud. If the sources are not protected and the cloud is compromised, the adversaries may collude with the cloud server to deduce the possible results from the historical observations and send back the "fake" results to the mobile devices to help the adversaries to break into the authentication systems. **The class labels** can be used to interpret the classification activities. In label privacy attacks, if the class labels are not protected, the adversary is able to obtain the specific physical meaning of the class labels. Some private information of the users or data contributors can be drawn from the classification results. For examples, the attackers are able to deduce the health status of the patients in the medical diagnosis system or track the daily activities of the individuals in the activity recognition system. However, health status and daily activities are extremely sensitive and most of the users are not willing to disclose them to the third party or to the public. Meanwhile, They have substantial commercial values so that are very likely to be targeted by the attackers.

To solve the above mentioned three types of privacy issues meanwhile providing reliable classification services, we propose a new Privacy-Preserving and cloud-enabled classification system, $\mathcal{P}^2$-SRC, based on Sparse Representation Classification (SRC), for CMAs. SRC is an emerging classification method and it has demonstrated superior performance on recognition accuracy compared with other traditional classification methods and is successfully used in face recognition [39,48], wildlife sound recognition [45] and activities classification [44]. Besides the accurate classification services, $\mathcal{P}^2$-SRC uses random projection matrices to compress sensor data meanwhile protect the data content. Then a Tor-like network is incorporated in the SRC-based classification framework to protect users from label and source privacy attacks. The contributions of this paper are as follows:

- Overall, we propose a new privacy-preserving and cloud-enabled classification framework, $\mathcal{P}^2$-SRC. As our evaluations on different classification applications, its recognition accuracy is almost the same to the traditional SRC method which indicates the privacy add-ons do not deteriorate the recognition accuracy of SRC.
- $\mathcal{P}^2$-SRC addresses different types of privacy attacks including content privacy attacks, source privacy attacks and label privacy attacks. To the best of our knowledge, it is the first privacy-preserving classification system that can address all of the three types of privacy attacks. Meanwhile it achieves significant improvement on accuracy-privacy trade-off according to our evaluations on two classification applications.
- We conduct two user studies to demonstrate, in intuitive approach, 1) $\mathcal{P}^2$- SRC provides reliable protections for users' data values, and 2) most of the users concern more about the protections of label and source information than sensor data values.
- At last, we implement a face recognition system based on $\mathcal{P}^2$-SRC on smartphones. The results show that the system cost of $\mathcal{P}^2$-SRC is negligible, and it provides real-time responses.

The organization of the rest of this paper is as follows. We first review the related literature in Section 2. Then we provide a brief introduction of SRC in Section 3. In Section 4, we discuss the system architecture, present two application examples and provide privacy analyses of $\mathcal{P}^2$-SRC. Section 5 evaluates the performance of two classification applications of $\mathcal{P}^2$-SRC with two publicly available datasets. Section 6 evaluates the system cost of the implementation of $\mathcal{P}^2$-SRC face recognition application on mobile devices. Finally we conclude the whole paper in Section 7.

2. Related work

In this section, we will give a literature review on the state of the arts in relevant research area. As our proposed system aims to protect the privacy of the users in mobile classification applications and the classification engine is the Sparse Representation Classifier (SRC), we discuss the recent advances in privacy protection for CMAs and applications of SRC.

2.1. Privacy protections for mobile applications

Mobile app development is a hotspot and the security of the mobile devices has become one of the most recent major concerns in mobile system research community [2,21,31,35]. For examples, Herberst et al. [17] designed privacy capsule to avoid the private information leakage via the untrusted third party mobile apps. While Zhu et al. [56] studied the private information leakage in code level; they addressed the 'module-level attacks' of the mobile app to prevent the third-party code stealing the private information on the COTS mobile devices. To protect the mobile devices from illegal usage meanwhile provide non-intrusive authentications, touch input implicit authentication (touch IA) was proposed and studied on different mobile devices; however, as the evaluations by Khan [22], touch IA was easy to be mimicked and not suitable from a security standpoint.

Private information can be vulnerable when sharing or uploading data via wireless channel. Xu et al. [49] proposed Walkie–talkie to generate encryption key based on gait. It enabled the mobile devices get paired automatically and prevents the devices from eavesdropping radio communications. Chakraborty et al. [7] proposed Ipshield a new context-aware privacy protection scheme to estimate the risks of sharing data in the cloud-enabled mobile applications. It provides the user with a list of the inferences that can be drawn from the data so the user can be aware of potential risks when sharing the data. To deal with the problem of reliability of the sensory data provided by the data contributors, Miao et al. [29] proposed a cloud-enabled privacy-preserving truth discovery framework which solved the problem of private information protection existed in the previous truth discovery approaches. Poolview [15] studied the problem of reconstruction attack and proposed a synthesis model which introduced correlated noise perturbation to protect the privacy of data sharing. Liu et al. proposed Pickle [25] which is the most related work to $\mathcal{P}^2$-SRC. Pickle enabled privacy-preserving collaborative learning for SVM using a linear regression based approach. However, it only considered the problem of content privacy attack. Other solutions [18,33] allowed users to control their resources and created shadow to prevent the