Privacy-preserving collaborative fuzzy clustering

Lingjuan Lyu, James C. Bezdek, Yee Wei Law, Xuanli He, Marimuthu Palaniswami

PII: S0169-023X(16)30321-4
DOI: 10.1016/j.datak.2018.05.002
Reference: DATAK 1651

To appear in: *Data & Knowledge Engineering*

Received Date: 17 November 2016
Revised Date: 18 November 2017
Accepted Date: 5 May 2018

Please cite this article as: L. Lyu, J.C. Bezdek, Y.W. Law, X. He, M. Palaniswami, Privacy-preserving collaborative fuzzy clustering, *Data & Knowledge Engineering* (2018), doi: 10.1016/j.datak.2018.05.002.

Privacy-Preserving Collaborative Fuzzy Clustering

Lingjuan Lyu[a,*], James C. Bezdek[b], Yee Wei Law[c], Xuanli He[b], Marimuthu Palaniswami[a]

[a] *Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, Australia*
[b] *Department of Computing and Information Systems, The University of Melbourne, Parkville, Australia*
[c] *School of Engineering, University of South Australia, Mawson Lakes, Australia*

Abstract

The proliferation of Internet of Things devices has contributed to the emergence of *participatory sensing* (PS), where multiple individuals collect and report their data to a third-party data mining cloud service for analysis. The need for the participants to collaborate with each other for this analysis gives rise to the concept of *collaborative learning*. However, the possibility of the cloud service being semi-honest poses a key challenge: preserving the participants' privacy.

In this paper, we address this challenge with a two-stage scheme called RG+RP: in the first stage, each participant perturbs his/her data by passing the data through a nonlinear function called *repeated Gompertz* (RG); in the second stage, he/she then projects his/her perturbed data to a lower dimension in an (almost) distance-preserving manner, using a specific *random projection* (RP) matrix. The nonlinear RG function is designed to mitigate *maximum a posteriori* (MAP) estimation attacks, while random projection resists *independent component analysis* (ICA) attacks and ensures clustering accuracy. The proposed two-stage randomisation scheme is assessed in terms of its recovery resistance to MAP estimation attacks. Preliminary theoretical analysis as well as experimental results on synthetic and real-world datasets indicate that RG+RP has better recovery resistance to MAP estimation attacks than most state-of-the-art techniques. For clustering, fuzzy c-means (FCM) is used. Results using seven cluster validity indices, root mean squared error (RMSE) and accuracy ratio show that clustering results based on two-stage-perturbed data are comparable to the clustering results based on raw data — this confirms the utility of our privacy-preserving scheme when used with either FCM or HCM.

Keywords: Participatory sensing, collaborative learning, clustering, privacy-preserving, randomisation

1. Introduction

The ubiquity of mobile sensing devices gave birth to *participatory sensing* (PS), a data crowdsourcing paradigm where participants "use evermore capable mobile phones and cloud services to collect and analyse systematic data for use in discovery" [1]. A closely related trend is *collaborative learning*, a data crowdsourcing paradigm where the participants not only "contribute individually collected training samples", but also "collaboratively construct statistical models for tasks in pattern recognition" [2]. Essentially, PS is participant-oriented, sensing-focused and cloud-assisted collaborative learning.

In this work, we are concerned with the PS scenario where the participants are data owners that contribute their data, but rely on a third-party data mining cloud service to perform clustering-based analysis on their joint data (see Fig. 1). For the clustering operation, we consider specifically *fuzzy c-means* [3], because it is widely used and well established. In this scenario, the participants' privacy becomes an issue because the cloud service might be *semi-honest*, i.e., although the cloud service does not deviate arbitrarily from the protocol, it might try to gain private information about the participants from their data. For

[*] Corresponding author
Email address: `llv@student.unimelb.edu.au` (Lingjuan Lyu)