Accepted Manuscript

Privacy-preserved big data analysis based on asymmetric imputation kernels and multiside similarities

Bo-Wei Chen, Seungmin Rho, Laurence T. Yang

PII: S0167-739X(16)30562-3
DOI: http://dx.doi.org/10.1016/j.future.2016.11.008
Reference: FUTURE 3213

To appear in: *Future Generation Computer Systems*

Received date: 21 May 2016
Revised date: 27 October 2016
Accepted date: 10 November 2016

Please cite this article as: B.-W. Chen, S. Rho, L.T. Yang, Privacy-preserved big data analysis based on asymmetric imputation kernels and multiside similarities, *Future Generation Computer Systems* (2016), http://dx.doi.org/10.1016/j.future.2016.11.008

ELSEVIER

Available online at www.sciencedirect.com

Future Generation Computer Systems

Privacy-Preserved Big Data Analysis Based on Asymmetric Imputation Kernels and Multiside Similarities

Bo-Wei Chen[a], Seungmin Rho[b], Laurence T. Yang[c]

[a]*School of Information Technology, Monash University, Australia*
[b]*Department of Media Software, Sungkyul University, Korea*
[c]*Huazhong University of Science and Technology, China, and St. Francis Xavier University, Canada*

Abstract

This study presents an efficient approach for incomplete data classification, where the entries of samples are missing or masked due to privacy preservation. To deal with these incomplete data, a new kernel function with asymmetric intrinsic mappings is proposed in this study. Such a new kernel uses three-side similarities for kernel matrix formation. The similarity between a testing instance and a training sample relies not only on their distance but also the relation between the testing sample and the centroid of the class, where the training sample belongs. This reduces biased estimation compared with typical methods when only one training sample is used for kernel matrix formation. Furthermore, centroid generation does not involve any clustering algorithms. The proposed kernel is capable of performing data imputation by using class-dependent averages. This enhances Fisher Discriminant Ratios and data discriminability. Experiments on two open databases were carried out for evaluating the proposed method. The result indicated that the accuracy of the proposed method was higher than that of the baseline. These findings thereby demonstrated the effectiveness of the proposed idea.

Keywords: Incomplete data analysis, partial similarity, multiside similarity, privacy preservation, kernel ridge regression (KRR), missing values, data imputation, kernel method, cloud computing, data analytics

1. Introduction

Incomplete data analysis has become a research hotspot with the recent increasing demand for big data processing, in addition to complexity problems in huge volumes. Take Internet of Things for example. Data collected by large-scale sensor networks could reach trillions in the future. However, when sensors fail, defective data are recorded in the dataset, subsequently resulting in biased estimation. In cloud computing, the same problem arises not merely because of erroneous samples, but because of privacy protection. Sensitive personal data, such as health records, faces, and names, are intentionally removed from the original data to avoid being maliciously manipulated [1-3]. These defective or masked data subsequently form an incomplete dataset. A systematic approach for conquering incomplete data is evitable.