A Secure Data Sharing Mechanism Based on Principal Component Analysis

Yeon-Ji Lee, Na-Yeon Shin, and Il-Gu Lee*

Sungshin Women's University, Seongbuk, Seoul, Korea
{220226036, 220216135, iglee}@sungshin.ac.kr

Abstract

Due to the varying environments of nodes and the characteristics of users, the information generated and collected can differ, leading to class imbalance issues. This can be addressed by sharing data; however, sharing data brings up concerns about privacy and potential interception by intermediaries. Therefore, this study proposes a secure data sharing mechanism using Principal Component Analysis (PCA). Nodes share data with each other through PCA, and each node reconstructs the shared data based on the PCA model they previously established. As a result, we address the trade-off issues among accuracy, memory usage, and privacy protection.

Keywords— PCA, data imbalance problem, unsupervised learning

1 Introduction

When collecting data from distributed nodes, the information gathered can vary depending on the environment and user characteristics, leading to class imbalance issues[1]. To address label-biased problems, it's essential to share data collected by other nodes[2]. However, the act of sharing data can raise concerns about data privacy breaches and potential interception by intermediaries. Therefore, a mechanism to securely share data between nodes is needed.

In this study, we propose a method that utilizes the properties of Principal Component Analysis (PCA) to securely share data between nodes in order to address class imbalance.

- By proposing a secure data sharing mechanism using PCA in a node-to-node data sharing environment, we address the issue of class imbalance.
- We evaluated assuming a data theft situation, dividing perspectives into attackers and legitimate users.
- We address the trade-off issue between accuracy, memory usage, and privacy.

2 Proposed Method

The proposed mechanism operates as follows: Initially, each node generates data and identifies the features they lack. Subsequently, they send a data transmission request to nodes that have an abundance of the missing features. The nodes that receive the request use Principal Component Analysis (PCA) to reduce the dimensionality of their data before transmitting it. The node that made the transmission request also conducts a PCA to create a model for data reconstruction. Upon receiving the data, it utilizes its PCA model to restore the data's dimensionality. Data compressed using PCA represents specific dimensional coordinates, reducing

The 7th International Conference on Mobile Internet Security (MobiSec'23), Dec. 19-21, 2023, Okinawa, Japan, Article No.P-10
*Corresponding author: Sungshin Women's University, Seoul, 02844, Republic of Korea, Email: iglee@sungshin.ac.kr

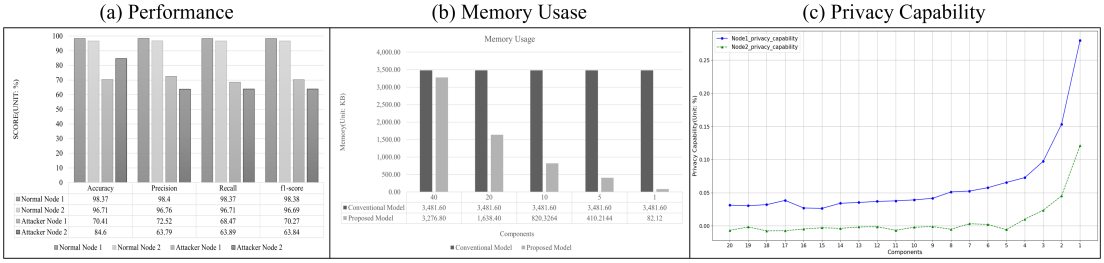

Figure 1: Experimental results

concerns about privacy breaches when sharing the original dataset. Additionally, sharing dimensionally compressed data requires less energy and memory than sharing the original dataset. However, to prove the performance of this mechanism, two validations are necessary. The first is to confirm that the data reduced in dimension by the sender's PCA model maintains its performance when reconstructed with the receiver's PCA. The second is to determine the optimal component value.

3 Conclusion

The data being shared is dimensionally reduced, and this data should remain safe from a privacy standpoint even if intercepted by a middleman. For the evaluation, the conventional model simply shared data without the use of PCA. And we assumed that an attacker would attempt label classification after acquiring the data. The experimental results can be seen in Figure 1. This study, through experiments, showed that the proposed mechanism does not compromise performance from a legitimate user's perspective, but its performance decreases from an attacker's standpoint. Moreover, at the optimal components value, the proposed mechanism was 42 times more memory efficient than the conventional sharing methods and exhibited the highest privacy metric.

Acknowledgement: This work was partly supported by grants of the Korea Institute for Advancement of Technology (KIAT) funded by the Korean Government (MOTIE) (P0008703, The Competency Development Program for Industry Specialist) and the MSIT under the ICAN (ICT Challenge and Advanced Network of HRD) program (No. IITP-2022-RS2022-00156310) supervised by the Institute of Information & Communication Technology Planning & Evaluation (IITP).

References

[1] Zheng, X., and Cai, Z. Privacy-preserved data sharing towards multiple parties in industrial IoTs. https://ieeexplore.ieee.org/document/9037362, 2020.

[2] Cheng Fan, Weilin He, Yichen Liu, Peng Xue and Yangping Zhao. A novel image-based transfer learning framework for cross-domain HVAC fault diagnosis: From multi-source data integration to knowledge sharing strategies. https://doi.org/10.1016/j.enbuild.2022.111995, 2022.