PAPER TITLE: DATA POISON DETECTION SCHEMES FOR DISTRIBUTED MACHINE LEARNING

Name: Habiba Mahrin

ID:20301339

Department: CSE

Course: High Performance computing (Cse449)

ABSTRACT

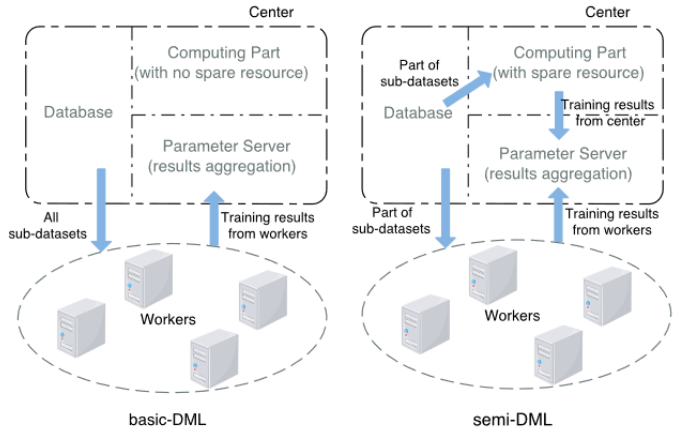
This paper discusses data poison detection schemes for Distributed Machine Learning (DML) and classifies DML into basic-DML and semi-DML. In basic-DML, a novel data poison detection scheme is proposed using a cross-learning mechanism to identify poisoned data, with the optimal number of training loops determined mathematically. In semi-DML, an improved data poison detection scheme is presented to enhance learning protection with central resource support, including an optimal resource allocation approach for efficient resource utilization. Simulation results indicate significant accuracy improvements in basic-DML (up to 20% for support vector machine and 60% for logistic regression) and resource savings in semi-DML (20-100% reduction in wasted resources).

INTRODUCTION

- ✓ Introduction of a data poison detection scheme for basic-DML, utilizing a cross-learning mechanism to generate training loops and establish a mathematical model for optimal security.
- ✓ Presentation of a practical method for identifying abnormal training results, aiding in the detection of poisoned datasets.
- ✓ Classification of DML into basic-DML and semi-DML, depending on the sharing of resources by the central entity in dataset training tasks.
- ✓ Emphasis on the need for a widely applicable DML protection mechanism.
- ✓ Prior mention of specific efforts using game theory to secure distributed support vector machine (DSVM) and collaborative deep learning but limited to specific DML algorithms.
- ✓ Acknowledgment of the urgent requirement for a broadly applicable DML protection mechanism.
- ✓ Validation of the proposed data poison detection schemes through experimental results.

- **BASIC-DML AND SEMI-DML**

This paper introduces a classification of Distributed Machine Learning (DML) into two categories: basic-DML and semi-DML. Both scenarios involve a central system with a database, computing server, and parameter server, but they differ in their functions. In basic-DML, the center lacks spare computing resources and relies on distributed workers for sub-dataset training, while in semi-DML, the center has additional computing resources and can both learn from some sub-datasets itself and integrate results from both central and distributed workers.

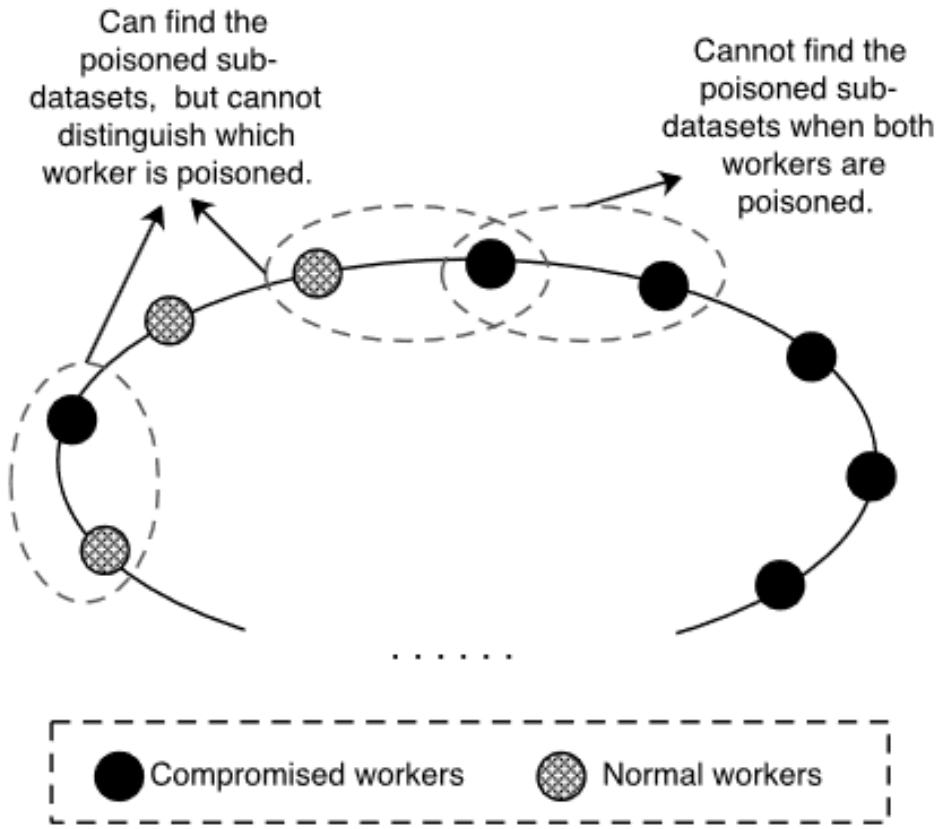


DATA POISON DETECTION SCHEME IN BASIC-DML

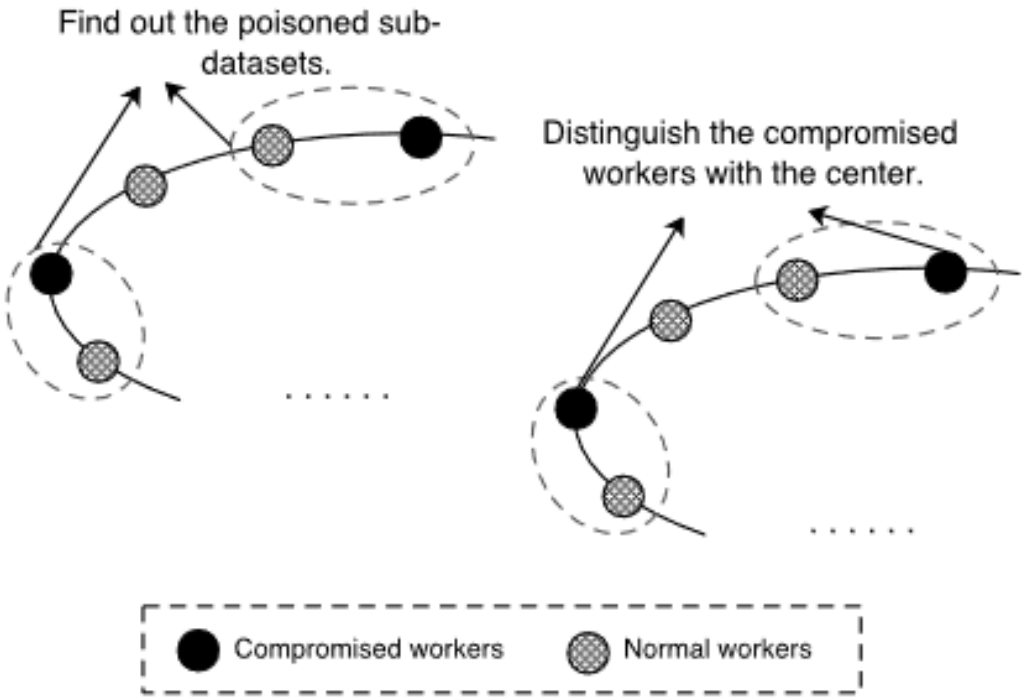
The paper will examine data poison detection in basic-DML, where the center lacks spare computing resources for sub-dataset training. The scheme comprises parameter thresholds, cross-learning, and abnormal result detection.

DATA POISON DETECTION SCHEME IN SEMI-DML

In the paper ,explore the enhanced data poison detection scheme in semi-DML, where the center contributes spare resources to dataset training. This scheme adds central assistance, enabling the center to learn some or verify worker results. Efficient resource utilization is a key concern in this scenario.



Data poison detection scheme in the basic-DML scenario.



Detection scheme in the semi-DML.

SUMMARY AND FUTURE WORK

This paper examines data poison detection in both basic-DML and semi-DML scenarios. It employs parameter thresholds to detect poisoned sub-datasets and presents a mathematical model for threat probability analysis. The improved detection scheme boosts model accuracy by up to 20% (SVM) and 60% (logistic regression) in basic-DML. In semi-DML, the enhanced scheme with optimal resource allocation reduces resource wastage by 20-100%. Future work should consider dynamic patterns for evolving application environments and balance between security and resource cost.

THE END