Privacy preserving distributed data mining based on secure multi-party computation

Authors of the presentation: Mihailov Daniel and Malgin Dimitri

Objectives

- To introduce what data mining is
- To define two problems to be solved
- Explain each task in detail
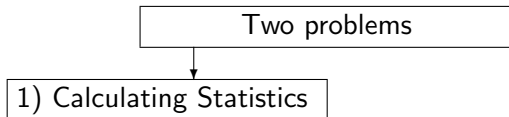- Provide algorithms 1,2,3 as solutions

Introduction

Text Data Mining

In recent years, we have witnessed an unprecedented big data explosion than ever before, especially for private data in business environments.

Problem definition In this section, we will first introduce two real-world tasks of privacy preserving distributed mining.

Two problems

Problem definition In this section, we will first introduce two real-world
tasks of privacy preserving distributed mining.

```
┌─────────────────────────────┐
│        Two problems         │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│   1) Calculating Statistics │
└─────────────────────────────┘
```

Problem definition In this section, we will first introduce two real-world
tasks of privacy preserving distributed mining.
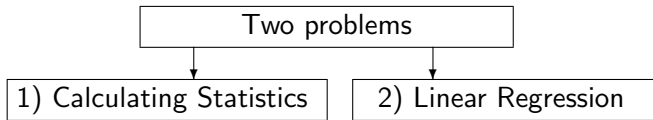
Task 1

Here, we first define
the problem of calculating
statistics of different types of data under privacy preserving distributed
data mining condition as:
Given data in the form of $\langle$type, value$\rangle$ tuples, the problem is to calculate
the statistical indicators of given data, without privacy leak risk for
both type and value in original data.

Task 2

In general, we can use the
linear regression to reveal
the relationship between variables. The same concern of business privacy
leakage exists if these companies are asked to disclose their trading records
to perform linear regression analysis. Hence, we define the problem of
performing linear regression under privacy preserving distributed data
mining condition.

Conclusions

Privacy preserving distributed data mining is an important task for big data. In this paper, we have proposed to utilize the MPC and SPDZ protocol to perform this task. To accomplish these data mining tasks, we have proposed two solutions based on matrix computation with one-hot encoding and LU decomposition.