Decision Tree and Subgroup Discovery based on Distributed Privacy Preservation

Nan Meng
Chen Yuechen
Fung Divine

University of Hong Kong

u3003637@connect.hku.hk

March 1, 2016

Content

- Motivation
- Distributed Data Mining
- Privacy Preserving Data Mining
- Decision Tree(ID3)
- Subgroup Discovery
- Related Works
- Problem Definition

Motivation

- Use of **technology for data collection** has seen an unprecedented **growth** in the last couple of decades. Individuals and organizations generate huge amount of data through everyday activities.

- **Decreasing storage and computation costs** have enabled us to collect data on different aspects of people's lives such as their **credit card transaction records**, **phone call** and **email lists**, **personal health information** and **web browsing habits**

Objective

The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process.

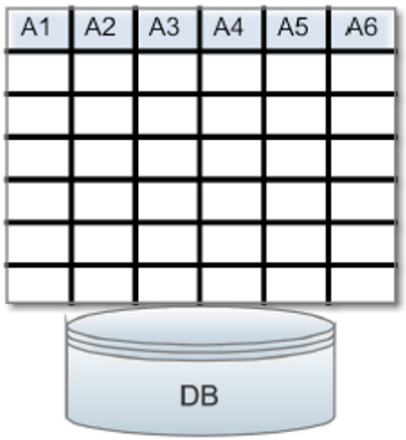
Distributed Data Mining

- Distributed data mining deals with the problem of data analysis in environments with **distributed** data, computing nodes, and users.
- Distributed data mining is a field of research that concentrates on **developing efficient algorithms** for mining of information from distributed data without centralizing it.
  - **Algorithms for homogeneous data distribution:**
    Also known as the **horizontally partitioned scenario**, all attributes or features are observed at every site. However, the set of observations or tuples across the different sites differ.
  - **Algorithms for heterogeneous data distribution:**
    Also known as the **vertically partitioned scenario**, each site has all tuples or rows, but only for a subset of the attributes for the overall data set.

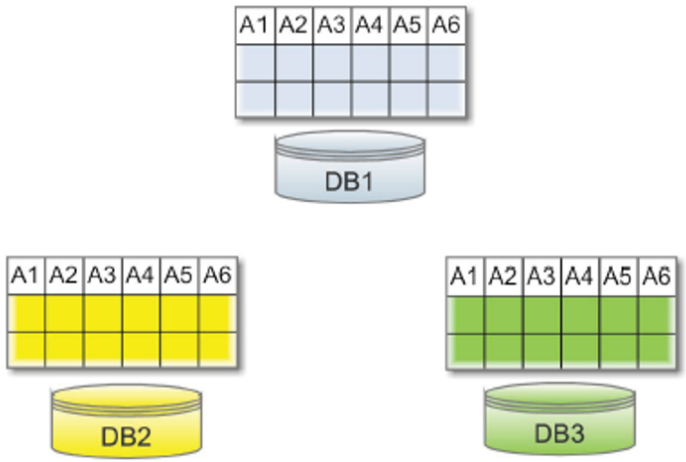
Distributed Data Mining(Intuitive View)

(b) **horizontally** partitioned scenario(homogeneous)

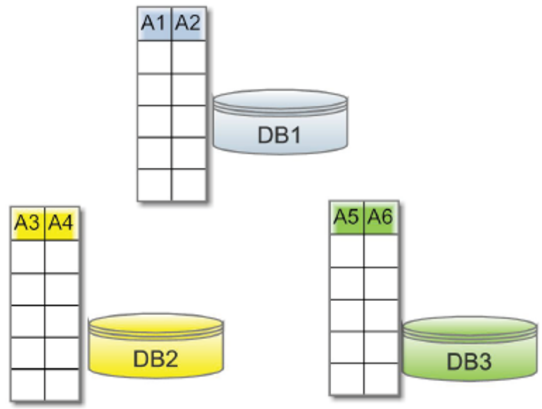
(c) **vertically** partitioned scenario(heterogeneous)



(a)



(b)



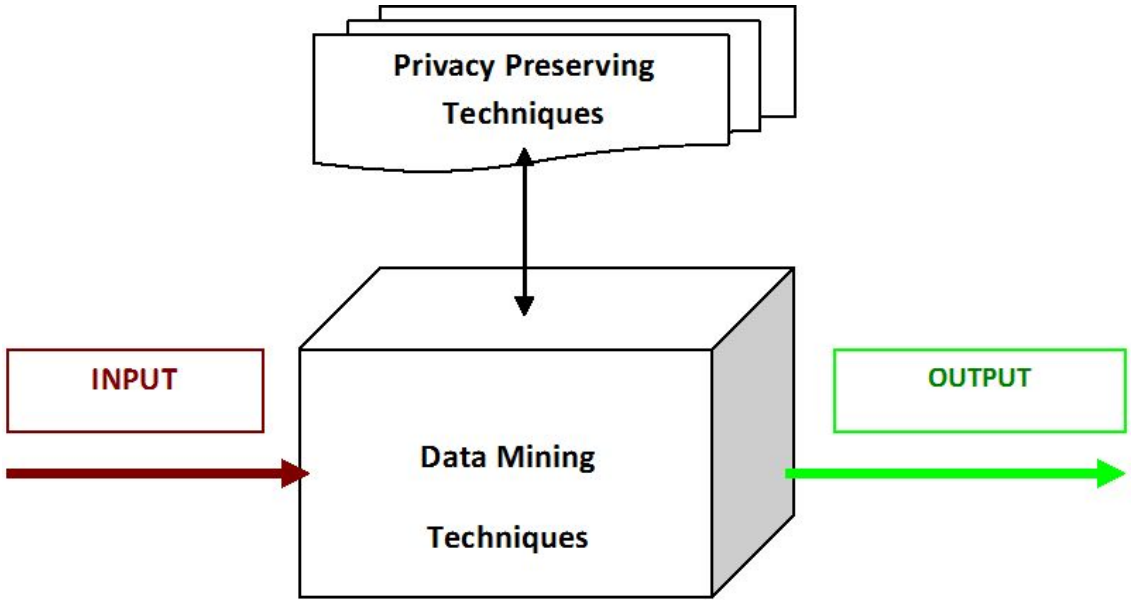
(c)

Privacy Preserving Data Mining

- Privacy preserving data mining, is a novel research direction in data mining and statistical databases, where data mining algorithms are **analyzed for the side-effects** they incur in data privacy.
- privacy preserving data mining algorithms can be classified into three categories:
  - **Data distortion based privacy:** Aim at **distorting** the **original private data**, when released, do not divulge any individually identifiable information.
  - **Cryptography based privacy:** Cryptographic protocols are called private when their execution does **not reveal any additional information** about the involved parties' data, other than what is computed as a result of the protocol execution.
  - **Output perturbation based privacy:** Output perturbation techniques discuss privacy with respect to the **information released as a result of querying** a statistical database by some external entity.

Privacy Preserving Data Mining(Intuitive View)

* Safeguard sensitive information
* Preserve data utility



Data distortion (perturbation/blocking)

Encrpytion techniques

Output perturbation

Decision Tree(ID3)

ID3(Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset. The kernel idea of ID3 is to classify samples according to the measurement of "information entropy" calculated by greedy algorithm.

ID3 Algorithm Steps

1. Calculate the **entropy** of **every attribute** using the data set S.
2. Split the set S into subsets using the attribute for which **entropy is minimum** (or, equivalently, information gain is maximum).
3. Make a decision tree node containing that attribute
4. Recurse on subsets using remaining attributes.

Subgroup Discovery

Rule Learning

* An approach to **predictive induction** (or supervised learning), aimed at constructing a set of rules to be used for **classification** and/or **prediction**

Association Rule Learning

* A form of **descriptive induction** (non-classificatory induction or unsupervised learning), aimed at the discovery of individual rules which define **interesting patterns** in data.

Subgroup Discovery

* A Task at the **Intersection** of **Predictive** and **Descriptive** Induction
* **Definition:** Given a **population** of individuals and a **property** of those individuals we are interested in, find population subgroups that are statistically '**most interesting**', e.g., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest

Related Works

Secure Multi-party Computation(SMC)

1. Distributed Association Rules Mining Based on SMC
   - **VAIDYA J, CLIFTON C. ACM SIGMOD 2002**
     Privacy preserving **association rules mining** in **vertically partitioned** data
   - **KANTARCIOGLU M,CLIFFTON C. IEEE Trans on Knowledge and Data Engineering, 2004**
     Privacy preserving distributed mining of **association rules** on **horizontally partitioned** data
   - **Rathore B.S. et al.** proposed **secure-sum** based technique to do Privacy Preservation Association Rule Mining task on Horizontally Partitioned Data **2016**
2. Distributed Clustering Mining Based on SMC
3. Distributed Classification Mining Based on SMC

Related Works

Secure Multi-party Computation(SMC)

1. Distributed Association Rules Mining Based on SMC
2. Distributed Clustering Mining Based on SMC
   - **Zhou et al.** proposed two distributed models based on SMC **(Naive model & Muilt-clustering model)**(**2009**)
   - **Rathore B.S. et al.** proposed **secure-sum** based technique to do Privacy Preservation Association Rule Mining task on Horizontally Partitioned Data(**2016**)
3. Distributed Classification Mining Based on SMC

Related Works

Secure Multi-party Computation(SMC)

1. Distributed Association Rules Mining Based on SMC
2. Distributed Clustering Mining Based on SMC
3. Distributed Classification Mining Based on SMC
   - **DU Wen-liang, ZHANG Zhi-jun** proposed the **Privacy Preserving ID3** algorithm(based on SMC) focus on **vertically partitioned data**.(**2002**)
   - **XIAO Ming-jun et al.** proposed the **Privacy Preserving C4.5** algorithm(based on SMC) focus on **horizontally partitioned data**.(**2006**)

Problem Definition

Consider a **distributed computing environment** consisting of nodes(parties) and connected via an underlying communication infrastructure.

- Each node has some data which is known only to itself.
- The nodes can exchange messages with any other node in the network.

This project aims at answering the following question:

Q : how can data mining tasks for extracting and utilizing useful knowledge from the union of all the data be executed in the system such that different nodes participating in the collaborative computation.

More specifically ...

$Q1$: Can ensure that the **required privacy** is actually achieved when mining the knowledge of data?

$Q2$: Can compute the privacy preserving data mining results with an **efficient** use of resources?

$Q3$: Can find some **subgroups** in the context of Privacy Preservation

$Q4$: Can **realize** the system and test it using algorithms?

Q & A

The End