Database Laboratory

Privacy-Preserving Data Mining: An Overview

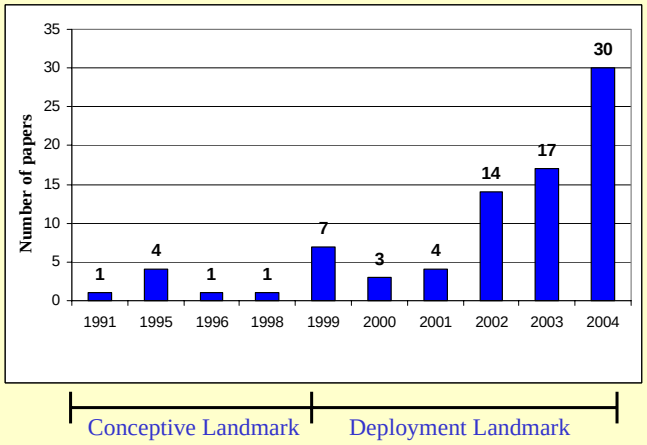
Stanley R. M. Oliveira

Database Systems Laboratory
Computing Science Department
University of Alberta, Canada

CMPUT 695

November 4th, 2004

Introduction



Outline

- Privacy-Preserving Data Mining (PPDM)
  - ✓ The Landmarks
  - ✓ Problems in defining privacy
  - ✓ Privacy violation
  - ✓ Some scenarios in PPDM
- Privacy-Preserving Clustering
  - ✓ Object Similarity-Based Representation (OSBR)
  - ✓ Dimensionality Reduction-Based Transformation (DRBT)

PPDM: Landmarks

- Conceptive Landmark: Preliminary discussions on security and privacy implications of data mining.
- Deployment Landmark: Proliferation of PPDM techniques and Workshops/Conferences.
- Prospective Landmark: The need for foundations for further research in PPDM.

PPDM: Problems in Defining Privacy

- "The right to be alone" (Warren & Brandeis).
- "The right to determine what (personal) information is communicated to others" (Schoeman).
- Privacy has become a digital problem.
- "Getting valid data mining results without learning the underlying data values" (Clifton et al.).
- "PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results" (Oliveira and Zaïane).

PPDM: Privacy Violation

- Changes in technology are making privacy harder.
  - ✓ reduced cost for data storage
  - ✓ increased ability to store and process large amounts of data
- Privacy violation in data mining: misuse of data.
- Defining privacy preservation in data mining:
  - ✓ Individual privacy preservation: protection of personally identifiable information.
  - ✓ Collective privacy preservation: protection of users' collective activity.

Some Scenarios in PPDM

- Characterizing Scenarios in PPDM:
  - ✓ **Scenario 1**: A hospital shares some data for research purposes.
  - ✓ **Scenario 2**: A collaboration between an Internet marketing company and an on-line retail company.
- General parameters:
  - ✓ **Outcome**: refers to the desired data mining results.
  - ✓ **Data distribution**: How are the data available for mining?
  - ✓ **Privacy preservation**: What are the privacy preservation requirements?

Outline

- Privacy-Preserving Data Mining (PPDM)
  - ✓ The Landmarks
  - ✓ Problems in defining privacy
  - ✓ Privacy violation
  - ✓ Some scenarios in PPDM
- ➡ Privacy-Preserving Clustering (PPC)
  - ✓ Object Similarity-Based Representation (OSBR)
  - ✓ Dimensionality Reduction-Based Transformation (DRBT)

Motivation for PPC

- Clustering plays an outstanding role in data mining:
  - ✓ Scientific data exploration;
  - ✓ Marketing;
  - ✓ Medical diagnosis;
  - ✓ Computational biology.
- Dual-goal: Protecting the underlying data values and achieving valid clustering results.
- Challenge: How can organizations protect personal data subjected to clustering and meet their needs to support decision making and to promote social benefits?

The Basics of Clustering Analysis

- Data Matrix ($m \times n$ matrix D)

$$D = \begin{bmatrix} a_{11} & \cdots & a_{1k} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2k} & \cdots & a_{2n} \\ a_{31} & \cdots & a_{3k} & \cdots & a_{3n} \\ \vdots & & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mk} & \cdots & a_{mn} \end{bmatrix}$$

- Dissimilarity Matrix ($m \times m$ matrix DM)

$$DM = \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(m,1) & d(m,2) & \cdots & \cdots & 0 \end{bmatrix}$$

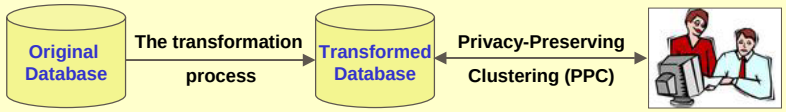
The Basics of Clustering Analysis (cont.)

- Distance between any two data points:
  - ✓ Given two n-dimensional vectors $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$

$$d(i, j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$ ← Euclidean Distance

- Euclidean distance satisfies the constraints:
  - ✓ $d(i,j) \geq 0$
  - ✓ $d(i,i) = 0$
  - ✓ $d(i,j) \leq d(i,k) + d(k,j)$

Problem Definition



- Problem: Given a data matrix $D_{m \times n}$, the goal is to transform D into D' so that the following restrictions hold:
  - ✓ A transformation $T: D \rightarrow D'$ must preserve the privacy of individual records.
  - ✓ The similarity between objects in D and D' must be the same or slightly altered by the transformation process.

Privacy-Preserving Clustering (PPC)

- **PPC over Centralized Data**:
  - ✓ The attribute values subjected to clustering are available in a central repository.
- **PPC over Vertically Partitioned Data**:
  - ✓ There are k parties sharing data for clustering, where $k \geq 2$;
  - ✓ The attribute values of the objects are split across the k parties.
  - ✓ Objects IDs are revealed for join purposes only. The values of the associated attributes are private.

Object Similarity-Based Representation (OSBR)

- **General Assumptions**:
  - ✓ The attributes could contain either binary, numerical, or categorical attributes, or even mixed types.
  - ✓ Object IDs should be replaced by fictitious identifiers.
- **PPC over Centralized Data**:
  - ✓ **Step 1** - Suppressing identifiers (e.g., address, phone number, etc.)
  - ✓ **Step 2** - Normalizing numerical attributes.
  - ✓ **Step 3** - Computing the dissimilarity matrix.

Object Similarity-Based Representation (OSBR)

Example 1: Sharing data for research purposes (OSBR).

Original Data

ID	age	weight	heart rate	Int_def	QRS	PR_int
123	75	80	63	32	91	193
342	56	64	53	24	81	174
254	40	52	70	24	77	129
446	28	58	76	40	83	251
286	44	90	68	44	109	128

A sample of the cardiac arrhythmia database (UCI Machine Learning Repository)

Transformed Data

$$DM = \begin{bmatrix} 0 & & & & \\ 2.243 & 0 & & & \\ 3.348 & 2.477 & 0 & & \\ 3.690 & 3.884 & 3.176 & 0 & \\ 3.020 & 4.082 & 4.130 & 3.995 & 0 \end{bmatrix}$$

The corresponding dissimilarity matrix

Object Similarity-Based Representation (OSBR)

- **Security of the OSBR:**
  - ✓ Lemma 1: Let $DM_{m \times m}$ be a dissimilarity matrix, where m is the number of objects. It is impossible to determine the coordinates of the two objects by knowing only the distance between them.
- **Complexity of the OSBR:**
  - ✓ Communication cost is of order $O(m^2)$, where m is the number of objects under analysis.

Object Similarity-Based Representation (OSBR)

- **Limitations of the OSBR:**
  - ✓Lemma 2: Knowing the coordinates of a particular object *i* and the distance *r* between *i* and any other object *j*, it is possible to estimate the attribute values of *j*.
  - ✓Vulnerable to attacks (Lemma 2).
  - ✓Expensive in terms of communication cost.
  - ✓Conclusion ⇒ The OSBR is not effective for PPC over Vertically Partitioned Data.

Dimensionality Reduction Transformation (DRBT)

- **General Assumptions**:
  - ✓The attribute values subjected to clustering are numerical only.
  - ✓In PPC over centralized data, object IDs should be replaced by fictitious identifiers;
  - ✓In PPC over vertically partitioned data, object IDs are used for the join purposes between the parties involved in the solution..
  - ✓The transformation (random projection) applied to the data might slightly modify the distances between data points.

Dimensionality Reduction Transformation (DRBT)

- **Random projection** from *d* to *k* dimensions:
  - ✓$D'_{n\times k} = D_{n\times d} \bullet R_{d\times k}$ (linear transformation), where

    D is the original data, *D'* is the reduced data, and *R* is a random matrix.
- *R* is generated by first setting each entry, as follows:
  - ✓**(R_1):** r_{ij} is drawn from an i.i.d. $N(0,1)$ and then normalizing the columns to unit length;
  - ✓**(R_2):** $r_{ij} = \sqrt{3} \times \begin{cases} +1 \text{ with probability } 1/6 \\ 0 \text{ with probability } 2/3 \\ -1 \text{ with probability } 1/6 \end{cases}$

Dimensionality Reduction Transformation (DRBT)

- **PPC over Centralized Data (General Approach)**:
  - ✓ **Step 1** - Suppressing identifiers (e.g., address, phone number, etc.)
  - ✓ **Step 2** - Normalizing attribute values subjected to clustering.
  - ✓ **Step 3** - Reducing the dimension of the original dataset by using random projection.
  - ✓ **Step 4** – Computing the error that the distances in *k*-*d* space suffer from:

$$Error^2 = \left(\sum_{i,j} (\hat{d}_{ij} - d_{ij})^2\right) / \left(\sum_{i,j} d_{ij}^{\ 2}\right)$$

Dimensionality Reduction Transformation (DRBT)

ID	age	weight	heart rate	Int_def	QRS	PR_int
123	75	80	63	32	91	193
342	56	64	53	24	81	174
254	40	52	70	24	77	129
446	28	58	76	40	83	251
286	44	90	68	44	109	128

Original Data

A sample of the cardiac arrhythmia database (UCI Machine Learning Repository)

	RP_1			RP_2		
ID	Att1	Att2	Att3	Att1	Att2	Att3
123	-50.40	17.33	12.31	91.0	-125.0	-97.58
342	-37.08	6.27	12.22	81.0	-98.50	-77.07
254	-55.86	20.69	-0.66	77.0	-93.0	-77.78
446	-37.61	-31.66	-17.58	83.0	-101.0	-73.53
286	-62.72	37.64	18.16	109.0	-123.0	-79.19

Transformed Data

RP_1: The random matrix is based on the Normal distribution.

RP_2: The random matrix is based on the much simpler distribution.

Dimensionality Reduction Transformation (DRBT)

- **PPC over Vertically Partitioned Data**:
  - ✓It is a generalization of the solution for PPC over centralized data.
  - ✓Any of the *k* parties can be the central one.
  - ✓Step 1 - Individual transformation (dimensionality reduction).
  - ✓Step 2 - Data exchanging or sharing.
  - ✓Step 3 - The central party mines the data and shares the clustering results with the other parties.

Dimensionality Reduction Transformation (DRBT)

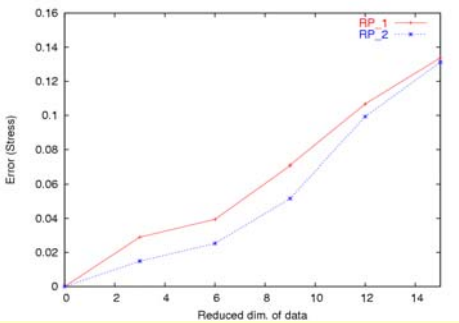
- **Security of the DRBT:**
  - ✓Lemma 3: A random projection from *d* to *k* dimensions, where $k << d$, is a non-invertible linear transformation.
- **Complexity of the OSBR:**
  - ✓The complexity of space requirements is of order $O(m)$, where *m* is the number of objects.
  - ✓The communication cost is of order $O(mlk)$, where *l* represents the size (in bits) required to transmit a dataset from one party to a central or third party.

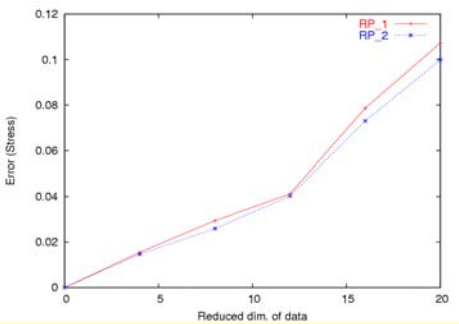
Experimental Results

- Datasets (UCI Repository of ML Databases):
  - ✓ chess (3196 x 37);
  - ✓ mushroom (8124 x 23)
- Methodology (DRBT only):
  - ✓PPC over centralized data:
    - (a) we applied dimensionality reduction to the datasets; (b) we computed the error produced on reduced datasets.
  - ✓PPC over vertically partitioned data:
    - (a) We split the datasets up to 4 parties; (b) we applied dimensionality reduction to the sub-datasets in each party; (c) we computed the error produced on the merged datasets (central party).

Experimental Results (cont.)

PPC over Centralized Data



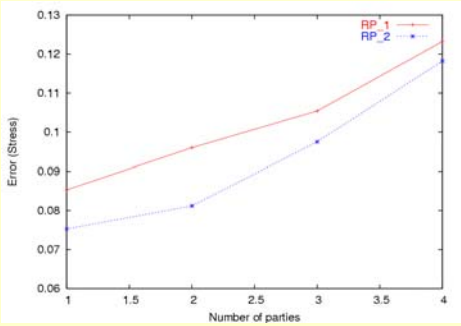


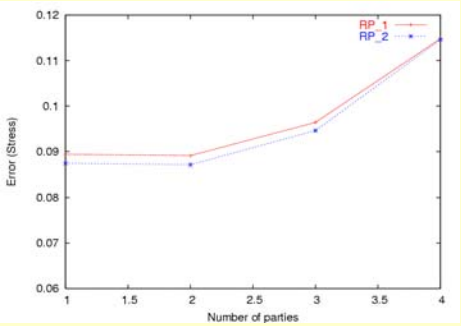
(a) The error produced on the the Mushroom dataset **(b) The error produced on the Chess dataset**

- **The Mushroom dataset = 23 dimensions**
- **The Chess dataset = 37 dimensions**

Experimental Results (cont.)

PPC over Vertically Partitioned Data





(a) The error produced on the the Mushroom dataset **(b) The error produced on the Chess dataset**

- **We reduced 50% of the dimensions of each sub-dataset, in each party. Then we combined the sub-datasets and computed the error on the merged datasets.**

Summary

- Privacy-Preserving Data Mining (PPDM)
  - ✓ The PPDM landmarks
  - ✓ Problems in defining privacy
  - ✓ Privacy violation
  - ✓ Some scenarios in PPDM
- Privacy-Preserving Clustering
  - ✓ Object Similarity-Based Representation (OSBR)
  - ✓ Dimensionality Reduction-Based Transformation (DRBT)

Summary (cont.)

- Highlights of the OSBR/DRBT Solutions:
  - ✓ They are independent of distance-based clustering algorithms.
  - ✓ They do not require CPU-intensive operations.
  - ✓ OSBR is effective to address PPC over centralized data.
  - ✓ DRBT is effective to address both PPC over centralized data and PPC over vertically partitioned data.

Thank You!