Data Mining, Lecture 15

Privacy preserving data mining

S. Nõmm

[1] Department of Software Science, Tallinn University of Technology

09.12.2021

Introduction

- Nowadays significant amount of data may be considered sensitive from the viewpoint of personal privacy.
- Leakage of such data may have harmful consequences for the particular individual and for the society.
- To overcome this problem privacy techniques are proposed.
    - Methods applied on the stage of data acquisition and publication.
        - Assumption: Data collector is not to be trusted.
        - Anonymous publication: the data is available to a trusted entity which responsibility is to remove sensitive attributes.
    - Methods applied to provide output privacy of the data mining algorithms. Output of the data mining algorithms may contain private information.
- Majority of the methods for privacy-preserving data mining reduce the representation accuracy of the data.

Privacy preserving on a data acquisition stage

- Random noise is added to the data while collecting data from users, with the use of specialized software or plugin. Acquired data may be publicly published together with the probability distribution function (and parameters) used to add the random noise.
- Distribution reconstruction. Leads histograms for each attribute.
- Data mining algorithms are applied to the reconstructed distributions.

Main drawback of this approach that one is required to work with distributions instead of the data record.

Reconstruction of the aggregate distributions

- Let the original values $x_1, \ldots, x_n$ are drawn from the probability distribution X. For each original (acquired) value x_i value y_i is added by the data acquisition software. This yield perturbed value z_i. The value y_i is drawn from the probability distribution Y (known publicly) and independent of X.
- Denote perturbed distribution Z then

$$\begin{aligned} Z &= X + Y \\ X &= Z - Y \end{aligned}$$

  If Y is known publicly and Z may be reconstructed on the basis of particular values then distribution X may be reconstructed for further analysis.

- Drawbacks: If variance of Y is large and the number of samples in Z is small then Z will have large variance. Then X should be reconstructed directly from the discrete samples of Z and Y.

Privacy preserving data publishing

- Some scholars consider privacy preserving data publishing to be a sub-type of privacy preserving data collection.
- Remove attributes which identify individual directly (*explicit identifiers*). (Not enough because linkage attacks are possible using other attributes).
- Remove pseudo and quasi identifiers. Each of this attributes is not explicit identifier but combined together may lead identification of the individual.
- Remove sensitive attributes.
- Group-based anonymization

k - anonymity model

- Suppression.
- Generalization.
- Synthetic data generation.
- Specification as probabilistic and uncertain databases (up to the date this approach was not studied extensively).
- Formal definition by Agarwal: (k-anonymity) A data set is said to be k-anonymized, if the attributes of each record in the anonymized data set cannot be distinguished from at least $k-1$ other data records.
- Algorithms: Samaratis, Incognito, Mondrian Multidimensional k-Anonymity.

ℓ- diversity model

- ℓ–diversity Principle: An equivalence class is said to be ℓ diverse, if it contains ℓ "well-represented" values for the sensitive attribute. An anonymized table is said to be ℓ-diverse, if each equivalence class in it is ℓ-diverse.
- (Entropy ℓ-diversity) Let $p_1, \dots, p_r$ be the fraction of the data records belonging to different values of the sensitive attribute in an equivalence class. The equivalence class is said to be entropy lℓ -diverse, if the entropy of its sensitive attribute value distribution is at least $log(\ell)$.

$$-\sum_{i=1}^{r} p_i \cdot \log(p_i) \geq \log(\ell)$$

An anonymized table is said to satisfy entropy ℓ-diversity, if each equivalence class in it satisfies entropy ℓ-diversity.

ℓ- diversity model

- In some cases ℓ- diversity model may be too restrictive.
- (Recursive (c, ℓ)-diversity) Let $p_1, \ldots, p_r$ be the fraction of the data records belonging to the r different values of the sensitive attribute in an equivalence class, such that $p_1 \geq p_2 \ldots \geq p_r$. The equivalence class satisfies recursive (c, ℓ)-diversity, if the following is true:

  $$p_1 < \sum_{i=l}^{r} p_i$$

  An anonymized table is said to satisfy recursive (c, ℓ)-diversity, if each equivalence class in it satisfies entropy (c, ℓ)-diversity.

Later definition ensures hat the most frequent attribute value in an equivalence class does not dominate the less frequent sensitive values in it.

The t-closeness Model

- ℓ- diversity model does not take into account the distribution of the sensitive attribute values in the original data set.
- (t-closeness Principle) Let $P = (p_1, \ldots, p_r)$ be a vector representing the fraction of the data records belonging to ther$r$ different values of the sensitive attribute in an equivalence class. Let $Q = (q_1 \ldots q_r)$ be the corresponding fractional distributions in the full data set. Then, the equivalence class is said to satisfy t-closeness, if the following is true, for an appropriately chosen distance function $s(\cdot, \cdot)$:

  $$s(P, Q) \leq t$$

  An anonymized table is said to satisfy t-closeness, if all equivalence classes in it satisfy t-closeness.
- As an appropriate distance function one may chose Variational (Manhattan divided by 2) or Kullback-Leibler distance functions.

$$s_{KL} = \sum_{i=1}^{r} \Big(p_i \log(p_i) - p_i \log(q_i) \Big)$$

Output privacy

- Output of any data mining algorithm may provide some private information.
- Frequent itemset mining is frequently mentioned with respect to these types of problems.

Distributed privacy

Partition the dataset between different entities.

- Different types of partitions.
- Connection to cryptography.