Data Protection and Privacy

Alessio Merlo

DIBRIS - University of Genoa

Email: alessio@dibris.unige.it
Google Hangout: alessio.mrl
Mobile Phone: +39 366 6060 815
Office Location (Valletta Puggia): Office 102, 1st floor, Via Dodecaneso 35, Genoa.
Office Location (Villa Bonino): 1st floor, Viale Francesco Causa, 13, Genoa.

Lesson IV: Privacy-Preserving Data Mining

Data Mining in a nutshell I

Privacy-preserving Data Mining

- Data mining is a process where critical business data are analyzed to gain new insights about customers, businesses, and markets. Data can be stored in any of the formats we analyzed before.
- Companies must ensure that the data are anonymized before being used for analytics/mining.

Data Mining in a nutshell II

Key features

- *Clustering*. It means partitioning a data set into clusters of similar data.
- *Classification*. It is used for prediction. In predictive modeling, a model is built to predict a value of a single variable based on the values of the other variables.
- *Association rule/pattern mining*. It is used to find associations between the transactions of a customer.
- *Outliers*. Identifying outlying data, that is, the data whose value is way outside or away from other data values.

Clustering, classification, and association rule mining, generate an output that does not contain any customer data but generalized models → No threats to *de-identification*. However, they should be protected in any case as:

- They can be provided to third parties.
- It is impossible to make assumptions on the background knowledge of the data snooper.
- There are regulatory compliance needs.

Association Rule Mining I

Scenario: Market Basket Analysis → find customers' buying patterns → ARM is used to find pattern or correlations in transaction database.

Problem formalization

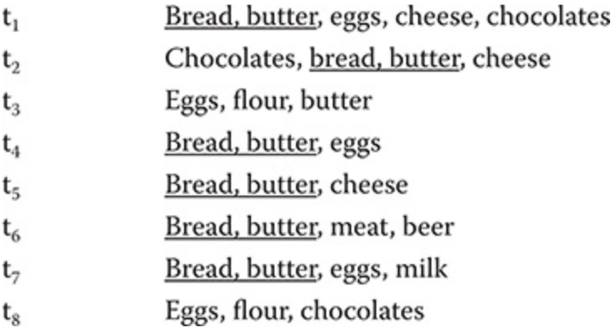
Given a database D, let $I = \{i_1, \ldots, i_m\}$ be a set of items. Let $T = \{t_1, \ldots, t_n\}$ be a set of transactions on the database. Each transition t_i is a set of items s.t. $t_i \subseteq I$.

$$t_i = X \rightarrow Y \text{ where } X \subset I \text{ and } Y \subset I \text{ and } X \cap Y = \varnothing.$$

Example: bread, butter, eggs, where $X = \{bread, butter\}$ and $Y = \{eggs\}$.

- **Support**: the number of transactions containing X. Low support implies that the transaction randomly occurs → a minimum support (*minSup*) should be defined to prune rare transactions.
- **Confidence** is the percentage of transactions in T that contain X and that also contain Y. Low confidence implies that it is impossible to predict Y from X → *minConf* should be defined to remove weak associations.

Association Rule Mining II



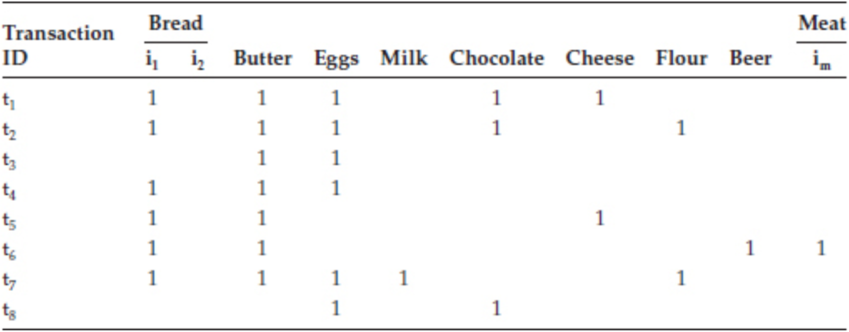
t_1	Bread, butter, eggs, cheese, chocolates
t_2	Chocolates, bread, butter, cheese
t_3	Eggs, flour, butter
t_4	Bread, butter, eggs
t_5	Bread, butter, cheese
t_6	Bread, butter, meat, beer
t_7	Bread, butter, eggs, milk
t_8	Eggs, flour, chocolates

Association Rule Mining III

Transaction Data: Random Perturbation Workshop Topics

- **MASK** [6] uses *probabilistic distortion*, i.e., flip each 0 or 1 with a parametrized probability p or retain as is with a probability $1 - p$.
- **Select-a-size** [2]
  1. For customer transaction t_i of length m, a random integer j from $[1, m]$ is first chosen with probability $p_m[j]$.
  2. Then j items are uniformly and randomly selected from the original transaction and inserted into the randomized transaction.
  3. A uniformly and randomly chosen fraction p_m of the remaining items in the database that are not present in the original transaction is inserted into the randomized transaction.
  4. The final randomized perturbation is composed of a subset of the true items from the original transaction and additional false items from the complementary set of items in the database [3].

Association Rule Mining IV



Transaction	Bread									Meat
ID	i_1	i_2	Butter	Eggs	Milk	Chocolate	Cheese	Flour	Beer	i_m
t_1	1		1	1		1	1			
t_2	1		1	1		1		1		
t_3			1	1						
t_4	1		1	1						
t_5	1		1				1			
t_6	1		1						1	1
t_7	1		1	1	1			1		
t_8				1		1				

Association Rule Mining V

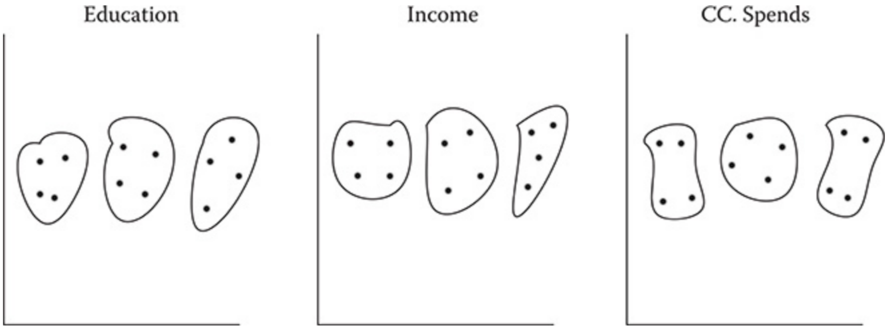
Clustering

- Recap: Data clustering is a method of creating groups of objects in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct;
- clustering is exploratory in nature → there is no right or wrong approach;
- each cluster has a center point;
- the goal of clustering is to find the intrinsic grouping of data for which a distance function is used.
- Example: given m_i the mean of a group, the cluster is made by all the data that has an Euclidean distance less than a given threshold.

The cluster quality is evaluated in terms of:

- Similarity measure → how much close are points in the cluster;
- Center
- Distance measure
- Structure

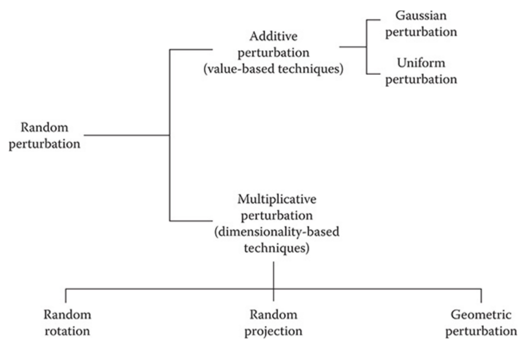
Association Rule Mining VI



Problem: data mining as clustering is generally carried out in outsourcing. Given a data table D that needs to be transformed to D' before outsourcing for cluster analysis, what data anonymization techniques can be applied on D that ensures high cluster quality and at the same time preserves the privacy of customer data?

Association Rule Mining VII

Random perturbation techniques: a survey



Additive Random Perturbation [1, 4]: it perturbs *sensitive values* using a randomized Gaussian or Uniform function. It does not preserve clustering, i.e., distance-based data mining.
Multiplicative Random Perturbation [5]: it allows to preserve distribution across multiple dimensions → is more suitable for data mining.

References I

[1] Rakesh Agrawal and Ramakrishnan Srikant.
Privacy-preserving data mining.
In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 439–450, New York, NY, USA, 2000. ACM.

[2] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke.
Privacy preserving mining of association rules.
Information Systems, 29(4):343 – 364, 2004.
Knowledge Discovery and Data Mining (KDD 2002).

[3] Jayant R. Haritsa.
Mining Association Rules under Privacy Constraints, pages 239–266.
Springer US, Boston, MA, 2008.

[4] X. Li, Z. Yan, and P. Zhang.
A review on privacy-preserving data mining.
In *2014 IEEE International Conference on Computer and Information Technology*, pages 769–774, Sept 2014.

[5] Stanley R. M. Oliveira and Osmar R. Zae.
Privacy preserving clustering by data transformation.
In *IN PROC. OF THE 18TH BRAZILIAN SYMPOSIUM ON DATABASES*, pages 304–318, 2003.

References II

[6] Shariq J. Rizvi and Jayant R. Haritsa.
Maintaining data privacy in association rule mining.
In *Proceedings of the 28th International Conference on Very Large Data Bases*, VLDB '02, pages 682–693. VLDB Endowment, 2002.