



Data preprocessing



Elena Baralis and Tania Cerquitelli
Politecnico di Torino

Data set types

- Record
 - Tables
 - Document Data
 - Transaction Data
- Graph
 - World Wide Web
 - Molecular Structures
- Ordered
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data




From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

2

Tabular Data

- A collection of records
 - Each record is characterized by a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes




From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

3

Document Data

- Each document becomes a `term' vector,
 - each term is a component (attribute) of the vector,
 - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0




From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

4

Transaction Data

- A special type of record data, where
 - each record (transaction) involves a set of items.
 - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

5

Ordered Data

- Sequences of transactions

Items/Events

↓ ↓


(A B) (D) (C E)

(B D) (C) (E)

(C D) (B) (A E)


⏟

An element of the sequence



From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


6



Attribute types

- There are different types of attributes
 - **Nominal**
 - Examples: ID numbers, eye color, zip codes
 - **Ordinal**
 - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
 - **Interval**
 - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
 - **Ratio**
 - Examples: temperature in Kelvin, length, time, counts


DBG 7
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Discrete and Continuous Attributes

- **Discrete Attribute**
 - Has only a finite or countably infinite set of values
 - Examples: zip codes, counts, or the set of words in a collection of documents
 - Often represented as integer variables.
 - Note: binary attributes are a special case of discrete attributes
- **Continuous Attribute**
 - Has real numbers as attribute values
 - Examples: temperature, height, or weight.
 - Practically, real values can only be measured and represented using a finite number of digits.
 - Continuous attributes are typically represented as floating-point variables.


DBG 8
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
 - Noise and outliers
 - missing values
 - duplicate data


DBG 9
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Missing Values

- Reasons for missing values
 - Information is not collected (e.g., people decline to give their age and weight)
 - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Handling missing values
 - Eliminate Data Objects
 - Estimate Missing Values
 - Ignore the Missing Value During Analysis
 - Replace with all possible values (weighted by their probabilities)


DBG 10
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Important Characteristics of Structured Data

- Dimensionality
 - Curse of Dimensionality
- Sparsity
 - Only presence counts
- Resolution
 - Patterns depend on the scale


DBG 11
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation


DBG 12
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
 - Data reduction
 - Reduce the number of attributes or objects
 - Change of scale
 - Cities aggregated into regions, states, countries, etc
 - More "stable" data
 - Aggregated data tends to have less variability


DBG 13
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Data reduction

- It generates a reduced representation of the dataset. This representation is smaller in volume, but it can provide similar analytical results
 - sampling
 - It reduces the cardinality of the set
 - feature selection
 - It reduces the number of attributes
 - discretization
 - It reduces the cardinality of the attribute domain


DBG 14
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Sampling ...

- The key principle for effective sampling is the following:
 - using a sample will work almost as well as using the entire data sets, if the sample is representative
 - A sample is representative if it has approximately the same property (of interest) as the original set of data


DBG 15
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Types of Sampling

- Simple Random Sampling
 - There is an equal probability of selecting any particular item
- Sampling without replacement
 - As each item is selected, it is removed from the population
- Sampling with replacement
 - Objects are not removed from the population as they are selected for the sample.
 - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
 - Split the data into several partitions; then draw random samples from each partition


DBG 16
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Dimensionality Reduction

- Purpose:
 - Reduce amount of time and memory required by data mining algorithms
 - Allow data to be more easily visualized
 - May help to eliminate irrelevant features or reduce noise
- Techniques
 - Principle Component Analysis
 - Singular Value Decomposition
 - Others: supervised and non-linear techniques

DBG 17
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Discretization

- It splits the domain of a continuous attribute in a set of intervals
 - It reduces the cardinality of the attribute domain
- Techniques
 - N intervals with the same width $W=(v_{\max} - v_{\min})/N$
 - Easy to implement
 - It can be badly affected by outliers and sparse data
 - Incremental approach
 - N intervals with (approximately) the same cardinality
 - It better fits sparse data and outliers
 - Non incremental approach
 - clustering
 - It well fits sparse data and outliers

DBG 18
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Discretization

The figure shows four scatter plots of data points on a coordinate system where the x-axis ranges from 0 to 20 and the y-axis from 0.0 to 1.0. The plots are labeled: 'Data' (raw data points), 'Equal interval width' (vertical dashed lines at intervals of 4), 'Equal frequency' (vertical dashed lines at positions where the data is split into four equal groups), and 'K-means' (vertical dashed lines at the mean of each of the four clusters).

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

19

Normalization

- It is a type of data transformation
 - The values of an attribute are scaled so as to fall within a small specified range, typically (-1, +1) or (0, +1)
- Techniques
 - min-max normalization

$$v' = \frac{v - \min_s}{\max_s - \min_s} (new_max_s - new_min_s) + new_min_s$$
 - z-score normalization $v' = \frac{v - mean_s}{stand_dev_s}$
 - decimal scaling

$$v' = \frac{v}{10^j}$$
 j is the smallest integer such that $\max(v') < 1$

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

20

Similarity and Dissimilarity

- Similarity
 - Numerical measure of how alike two data objects are.
 - Is higher when objects are more alike.
 - Often falls in the range [0, 1]
- Dissimilarity
 - Numerical measure of how different are two data objects
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0
 - Upper limit varies
- Proximity refers to a similarity or dissimilarity

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

21

Euclidean Distance

- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q .
- Standardization is necessary, if scales differ.

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

22

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.
 - $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
 - $d(p, q) = d(q, p)$ for all p and q . (Symmetry)
 - $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q , and r . (Triangle Inequality)

where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q .
- A distance that satisfies these properties is a **metric**

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

23

Common Properties of a Similarity

- Similarities, also have some well known properties.
 - $s(p, q) = 1$ (or maximum similarity) only if $p = q$.
 - $s(p, q) = s(q, p)$ for all p and q . (Symmetry)

where $s(p, q)$ is the similarity between points (data objects), p and q .

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

24

Similarity Between Binary Vectors

- Common situation is that objects, p and q , have only binary attributes
- Compute similarities using the following quantities
 - M_{01} = the number of attributes where p was 0 and q was 1
 - M_{10} = the number of attributes where p was 1 and q was 0
 - M_{00} = the number of attributes where p was 0 and q was 0
 - M_{11} = the number of attributes where p was 1 and q was 1
- Simple Matching and Jaccard Coefficients
 - SMC = number of matches / number of attributes
 $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$
 - J = number of 11 matches / number of not-both-zero attributes values
 $= (M_{11}) / (M_{01} + M_{10} + M_{11})$

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006 25

SMC versus Jaccard: Example

$p = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$
 $q = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$

$M_{01} = 2$ (the number of attributes where p was 0 and q was 1)
 $M_{10} = 1$ (the number of attributes where p was 1 and q was 0)
 $M_{00} = 7$ (the number of attributes where p was 0 and q was 0)
 $M_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$SMC = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$

$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006 26

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / (||d_1|| ||d_2||)$$
 where \bullet indicates vector dot product and $||d||$ is the length of vector d .
- Example:

$$d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$$

$$d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$||d_1|| = (3^2+2^2+0^2+5^2+0^2+0^2+0^2+2^2+0^2+0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$||d_2|| = (1^2+0^2+0^2+0^2+0^2+0^2+1^2+0^2+2^2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006 27

Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.
 - For the k^{th} attribute, compute a similarity, s_k , in the range $[0, 1]$.
 - Define an indicator variable, δ_k , for the k_k attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$
 - Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^n \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006 28

Combining Weighted Similarities

- May not want to treat all attributes the same.
 - Use weights w_k which are between 0 and 1 and sum to 1.

$$similarity(p, q) = \frac{\sum_{k=1}^n w_k \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

$$distance(p, q) = \left(\sum_{k=1}^n w_k |p_k - q_k|^r \right)^{1/r}$$

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006 29