

Chapter IV: Preprocessing

Knowledge Discovery in Databases

Luciano Melodia M.A.

Evolutionary Data Management, Friedrich-Alexander University Erlangen-Nürnberg

Summer semester 2021



Chapter IV: Preprocessing

This is our agenda for this lecture:

Data preprocessing: an overview.

Data quality.

Major tasks in data preprocessing.

Data cleaning.

Data integration.

Data reduction.

Data transformation and data discretization.

Summary.

Data quality: why preprocess the data?

This is our agenda for this lecture:

Measures for data quality: A multidimensional view:

Accuracy: correct or wrong, accurate or not.

Completeness: not recorded, unavailable.

Consistency: some modified but some not, dangling refs, etc.

Timeliness: timely updated?

Believability: how trustworthy is it, that the data is correct?

Interpretability: how easily can the data be understood?

And even many more!

Major tasks in data preprocessing

Data cleaning:

- Fill in missing values.
- Smooth noisy data.
- Identify or remove outliers.
- Resolve inconsistencies.

Data integration:

- Integration of multiple databases.
- Data cubes or files.

Data reduction:

- Dimensionality reduction.
- Numerosity reduction.
- Data compression.

Data transformation and data discretization:

- Normalization.
- Concept-hierarchy generation.

Chapter IV: Preprocessing

Data preprocessing: an overview.

Data quality.

Major tasks in data preprocessing.

Data cleaning.

Data integration.

Data reduction.

Data transformation and data discretization.

Summary.

Data cleaning

Data in the real world is **dirty**. Lots of potentially incorrect data:

E.g. instrument faulty, human or computer error, transmission error.

Incomplete: lacking attributes, lacking certain attributes of interest or containing aggregate data.

E.g. occupation = "" (missing data).

Noisy: containing noise, errors or outliers.

Stochastic deviation, imprecision.

E.g. measurements.

Inconsistencies: containing discrepancies in codes or names.

E.g. age = "42", birthday = "03/07/2010".

Was rating "1,2,3" and now it is "A,B,C".

Discrepancy between duplicate records (e.g. address old and new).

Intentional (only default value, e.g. disguised missing data):

Jan. 1 as everyone's birthday?

Incomplete (missing) data

Data is not always available.

E.g. many tuples have no recorded value for several attributes.

Examples are customer income in sales data.

Missing data may be due to:

Equipment malfunction.

Inconsistency with other recorded data and thus deleted.

Data not entered due to misunderstanding.

Certain data may not be considered important at the time of entry.

Not registered history or changes of the data.

Missing data may need to be inferred.

How to handle missing data?

Ignore the tuple:

Usually done when class label is missing (when doing classification).

Not effective when the percentage of missing values per attribute varies considerably.

Fill in the missing value manually.

Tedious or infeasible.

Fill in automatically with:

A global constant, e.g. "unknown", maybe a new class.

The attribute mean.

The attribute mean for all samples belonging to the same class.

The most probable value: Inference-based such as Bayesian formula or decision tree.

Noisy data?

Noise:

- Random error or variance in a measured variable.

- Stored value a little bit off the real value, up or down.

- Leads to (slightly) incorrect attribute values.

May be due to:

- Faulty or imprecise data-collection instruments.

- Data-entry problems.

- Data-transmission problems.

- Technology limitation.

- Inconsistency in naming conventions.

How to handle noisy data?

Beginning:

First sort data and partition into (equal-frequency) bins.

Then smooth by bin mean, by bin median or by bin boundaries.

Regression:

Smooth by fitting the data to regression functions.

Clustering:

Detect and remove outliers.

Combined computer and human inspection:

Detect suspicious values and check by human.

E.g. deal with possible outliers.

Data cleaning as a process

Data-discrepancy detection:

Use **metadata** (e.g. domain, range, dependency, distribution).

Check field overloading.

Check uniqueness rule, consecutive rule and null rule.

Use commercial tools:

Data scrubbing: use simple domain knowledge (e.g. postal code, spell-check) to detect errors and make corrections.

Data auditing: by analyzing data to discover rules and relationships to detect violators (e.g. correlation and clustering to find outliers).

Data migration and integration:

Data-migration tools: allow transformations to be specified.

ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface.

Integration of the two processes.

Iterative and interactive (e.g. the Potter's Wheel tool).

Chapter IV: Preprocessing

Data preprocessing: an overview.

Data quality.

Major tasks in data preprocessing.

Data cleaning.

Data integration.

Data reduction.

Data transformation and data discretization.

Summary.

Data integration

Data integration:

Combine data from multiple sources into a coherent store.

Schema integration:

E.g. $A.cust-id \equiv B.cust-\#$.

Integrate metadata from different sources.

Entity-identification problem:

Identify the same real-world entities from multiple data sources.

E.g. Bill Clinton = William Clinton.

Detecting and resolving data-value conflicts:

For the same real world entity, attribute values from different sources are different.

Possible reasons:

- Different representations (coding).

- Different scales, e.g. metric vs. British units.

Handling redundancy in data integration

Redundant data often occur when integrating multiple databases.

Object (entity) identification:

The same attribute or object may have different names in different databases.

Derivable data:

One attribute may be a "derived" attribute in another table. E.g. annual revenue.

Redundant attributes:

Can be detected by **correlation analysis** and **covariance analysis**.

Careful integration of the data from multiple sources:

Helps to reduce/avoid redundancies and inconsistencies and improve mining speed and quality.

Correlation analysis for nominal data (I)

Two attributes:

A has n distinct values: $A := \{a_1, a_2, \dots, a_n\}$.

B has m distinct values: $B := \{b_1, b_2, \dots, b_m\}$.

Contingency table:

Columns: the n values of A .

Rows: the m values of B .

Cells: counts of records with

$A' = \{a_i \in A : a_i = a_k \text{ for } a_k \in A\}$ and

$B' = \{b_j \in B : b_j = b_l \text{ for } b_l \in B\}$.

Expected count in cell (i, j) :

$$e_{ij} = \frac{\#A' \cdot \#B'}{\#A + \#B}, \quad (1)$$

where $\#A + \#B$ is the total number of records.

Correlation analysis for nominal data (II)

χ^2 -test:

$$\chi^2 = \sum_{i=1}^N \frac{(x_i - \hat{x}_i)^2}{\hat{x}_i}. \quad (2)$$

Summing over all cells of the contingency table.

No correlation (i.e. independence of attributes) yields χ^2 value of zero.

The larger the χ^2 value, the more likely the variables are related.

The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count.

Correlation does not imply causality!

E.g. # of hospitals and # of car-thefts in a city are correlated.

Both are causally linked to the third variable: population.

χ^2 calculation: an example

	Play chess	Not play chess	Sum (row)
Like Science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum (column)	300	1200	1500

Numbers in parenthesis are expected counts calculated based on the data distribution in the two categories.

χ^2 calculation:

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93. \quad (3)$$

It shows that "like science fiction" and "play chess" are correlated in the group.

Correlation analysis of numerical data

Correlation coefficient:

Also called Pearson's product-moment coefficient

$$r_{A,B} = \frac{\sum_{i=1}^{N+M} (a_i - \mu_A)(b_i - \mu_B)}{(N+M-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N+M} (a_i b_i) - (N+M)\mu_A\mu_B}{(N+M-1)\sigma_A\sigma_B}. \quad (4)$$

where $N = \#A$, $M = \#B$, μ_A and μ_B are the means of A and B , respectively. σ_A and σ_B denote the corresponding standard deviations.

If $r_{A,B} > 0$, A and B are positively correlated (A 's values increase with B 's).

The higher, the stronger the correlation.

$r_{A,B} = 0$: independent.

$r_{A,B} < 0$: negatively correlated.

Visually evaluating correlation

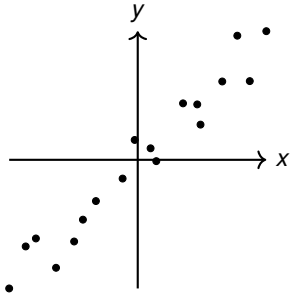


Figure: a) Positive correlation.

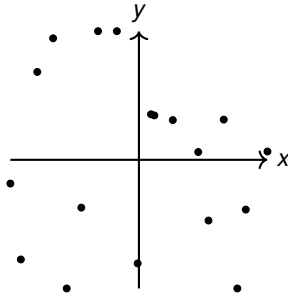


Figure: b) Uncorrelated/no correlation.

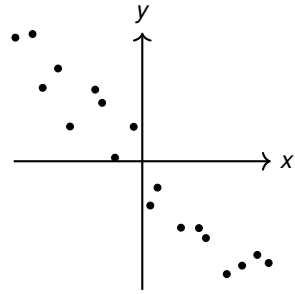



Figure: c) Negative correlation.

Thank you for your attention.
Any questions about the forth chapter?

Ask them now, or again, drop me a line:
 `luciano.melodia@fau.de`.