Data Analysis and Knowledge Discovery

Data Preparation

Jukka Heikkonen

University of Turku
Department of Information Technology

Jukka.Heikkonen@utu.fi

Outline

Data preparation

Data understanding provides general information about the data like

- the existence and partly also about the character of missing values,
- outliers,
- the character of attributes and
- dependencies between attribute.

Data preparation

Data preparation uses this information to

- select attributes,
- reduce the dimension of the data set,
- select records,
- treat missing values,
- treat outliers,
- integrate, unify and transform data and
- improve data quality.

Feature extraction

Feature extraction refers to construct (new) features from the given attributes.

Example. We are interested in finding the best workers in a company.

- Attributes like
  - the tasks, a worker has finished within each month,
  - the number of hours he has worked each month,
  - the number of hours that are normally needed to finish each task.
- In principle, these attributes contain information about the efficiency of the worker.
- But instead using these three "raw" attributes, it might be more useful to define a new attribute *efficiency* which is the hours actually spent to finish the task divided by the hours normally needed to finish the tasks.

Feature extraction and dimensionality reduction

Dimensionality reduction techniques like PCA can also be considered as features extraction methods.

But such automatic feature extraction methods usually lead to features that can no longer be interpreted in a meaningful way.

(How to understand a feature which is a linear combination of 10 attributes?)

Therefore, in most cases either knowledge-based, problem dependent feature extraction methods or feature selection techniques are preferred.

Feature extraction

Especially, for complex data types, feature extraction is required.

Example.

Text data analysis. Frequency of keyword, ...

Time series data analysis. Fourier or wavelet coefficients, ...

Image data analysis. Fourier or wavelet coefficients, ...

Graph data analysis. number of vertices, number of edges, ...

Feature selection

Feature selection refers to techniques to choose a subset of the features (attributes) that is as small as possible and sufficient for the data analysis.

Feature selection includes

- removing (more or less) irrelevant features and
- removing redundant features.

- For removing irrelevant features, a performance measure is needed that indicates how well a feature or subset of features performs w.r.t. the considered data analysis task.
- For removing redundant features, either a performance measure for subsets of features or a correlation measure is needed.

Feature selection

Example. Consider the following classification task that consists of 9 repetitions of the four records in the first table and the four records in the second table.

9 ×

A	B	C	D	target
+	+	+	−	no
+	−	+	−	yes
−	+	+	−	yes
−	−	−	+	no

1 ×

A	B	C	D	target
+	+	+	−	no
+	−	+	−	yes
−	+	+	−	yes
−	−	+	+	no

Performance of the single attributes

A	target no	target yes	B	target no	target yes	C	target no	target yes	D	target no	target yes
+	10	10	+	10	10	+	11	20	+	10	0
−	10	10	−	10	10	−	9	0	−	10	20

Feature selection

- A greedy strategy selecting those attributes with the best performance, would choose attributes C and D first.
- Attributes C and D together cannot perfectly predict the target value.
- Attributes A and B alone provide no information about the target value.
- However, attributes A and B together are sufficient to perfectly predict the target value.

Evaluating the performance of isolated attributes does usually not provide proper information about their performance in combination.

Feature selection techniques

Selecting the top-ranked features. Choose the features with the best evaluation when single features are evaluated.

Selecting the top-ranked subset. Choose the subset of features with the best performance.
This requires exhaustive search and is impossible for larger numbers of features.
(For 20 features there are already more than one million possible subsets.)

Forward selection. Start with the empty set of features and add features one by one. In each step, add the feature that yields the best improvement of the performance.

Backward elimination. Start with the full set of features and remove features one by one. In each step, remove the feature that yields to the least decrease in performance.

Record selection

Timeliness. Some of the older data might be outdated and might not be useful or even misleading for the data analysis task. Then only the recent data should be selected.

Representativeness. The sample in the database might not be representative for the whole population. When we have information about the distribution of the population, we can draw a representative subsample from our database.

Rare events. When we are interest in predicting rare events (e.g. stock market crashes, failures of a production line), it can be helpful

- to incorporate this in the cost function or
- to artificially increase the proportion of these rare events in the data set by adding copies of them or
- to choose only a subset of the other data.

Data cleansing

Data cleansing or data scrubbing refers to detecting and correcting or removing

- inaccurate,
- incorrect or
- incomplete

records from a data set.

Improve data quality

- Turn all characters into capital letters to level case sensitivity.
- Remove spaces and nonprinting characters.
- Fix the format of numbers, date and time (including decimal point).
- Split fields that carry mixed information into two separate attributes, e.g. *"Chocolate, 100g"* into *"Chocolate"* and *"100.0"*. This is known as field overloading.
- Use spell-checker or stemming to normalize spelling in free text entries.
- Replace abbreviations by their long form (with the help of a dictionary).

Improve data quality

- Normalize the writing of adresses and names, possibly ignoring the order of title, surname, forename, etc. to ease their re-identification
- Convert numerical values into standard units, especially if data from different sources (and different countries) are used.
- Use dictionaries containing all possible values of an attribute, if available, to assure that all values comply with the domain knowledge.

Missing value

Ignorance/Deletion. If only a few records have missing values and it can be assumed that the values are missing completely at random (MCAR) (observed at random (OAR)), these records can be deleted for the following data analysis steps.

Imputation. The missing values may be replaced by some estimate.

- The mean, the median or the mode of the attribute. (MCAR/OAR required!)
- By an estimation based on the other attributes. (MAR required!)

Explicit value. Missing values are characterized by a specific value, say `MISSING` or `?`. The chosen model in the modelling steps must be able to handle missing values. (Most models assume MCAR/OAR!)

Transformation of data

Some models can only handle numerical attributes, other models only categorical attributes.

In such cases, categorical attributes must be transformed into numerical ones or vice versa.

Transforming a categorical attribute into a numerical attribute.

- A binary attribute can be turned into a numerical attribute with the values 0 and 1.
- A categorical attribute with more than two values, say $a_1, \ldots, a_k$, should not be turned into a single numerical attribute with the values $1, \ldots, k$, unless the attribute is an ordinal attribute. It should be turned into k attributes $A_1, \ldots, A_k$ with values 0 and 1. a_i is represented by $A_i = 1$ and $A_j = 0$ for $i \neq j$.

Transformation of data: Discretization

Discretization techniques refer to splitting a numerical range into a number of finite bins.

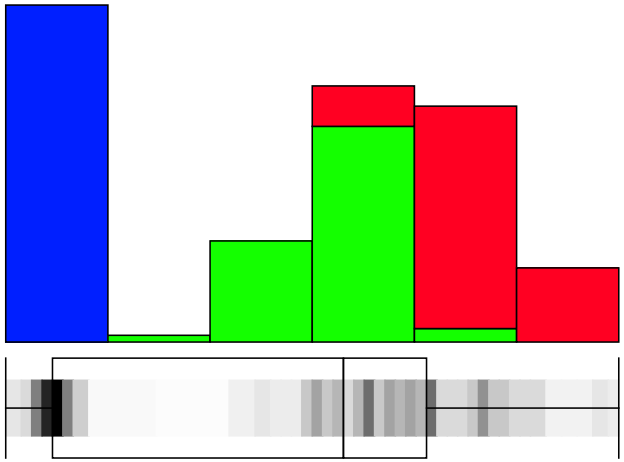
Equi-width discretization. Splits the range into intervals (bins) of the same length.

Equi-frequency discretization. Splits the range into intervals such that each interval (bin) contains (roughly) the same number of records.

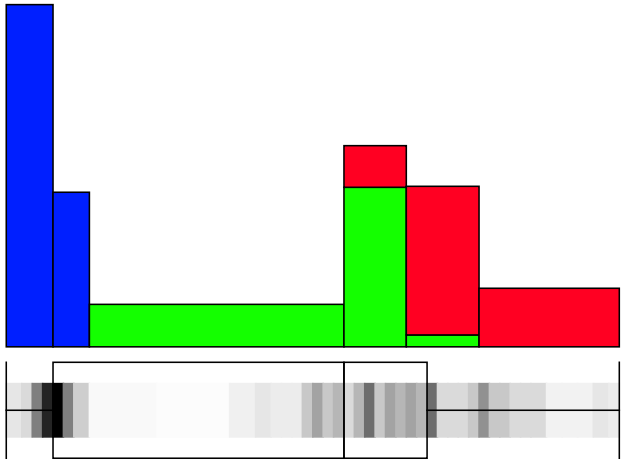
V-optimal discretization. Minimises $\sum_i n_i V_i$ where n_i is the number of data objects in the ith interval and V_i is the sample variance of the data in this interval.

Minimal entropy discretization. Minimises the entropy. (Only applicable in the case of classification problems.)

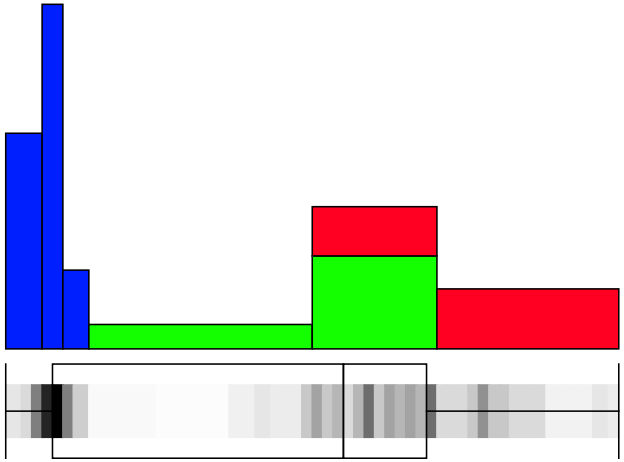
Transformation of data: Discretization



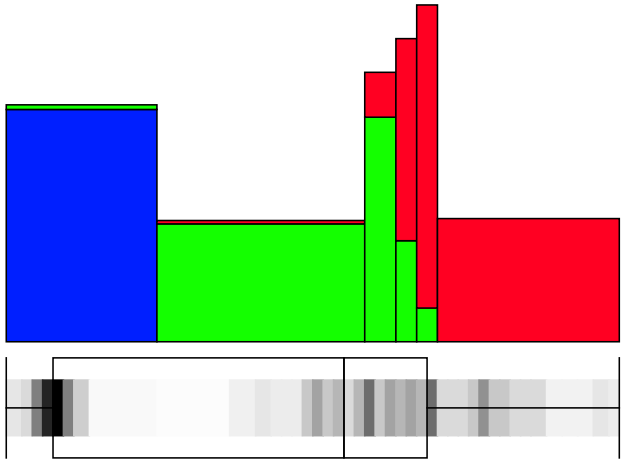
Equi-width



Equi-frequency



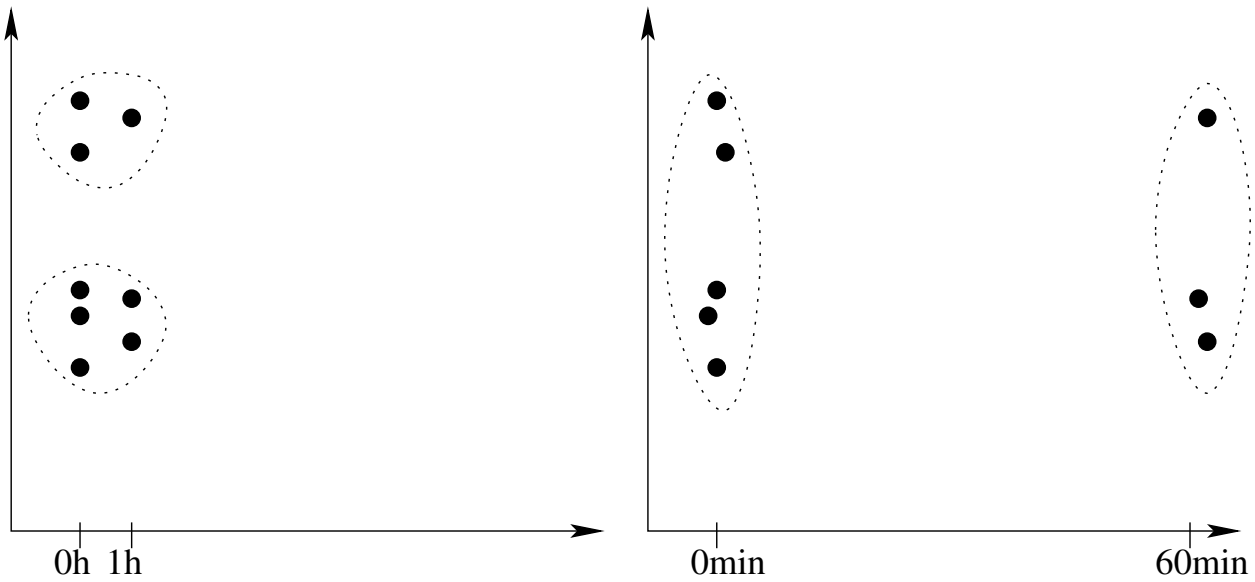
V-optimal



Minimal entropy

Normalisation/Standardisation

For some data analysis techniques (e.g. PCA, MDS; cluster analysis) the influence of an attribute depends on the scale or measurement unit.



To guarantee impartiality, some kind of standardisation or normalisation should be applied.

Normalisation/Standardisation

min-max normalization. For a numerical attribute X with $\min_X$ and $\max_X$ being the minimum and maximum value in the sample, the min-max normalization is defined as

$$n : \mathrm{dom}X \to [0, 1], \qquad x \mapsto \frac{x - \min_X}{\max_X - \min_X}$$

z-score standardization. For a numerical attribute X with sample mean $\hat{\mu}_X$ and empirical standard deviation $\hat{\sigma}_X$, the z-score standardization is defined as

$$s : \mathrm{dom}X \to \mathbb{R}, \qquad x \mapsto \frac{x - \hat{\mu}_X}{\hat{\sigma}_X}$$

robust z-score standardization. The sample mean and empirical standard deviation are easily affected by outliers. A more robust alternative is (see also boxplots):

$$s : \mathrm{dom}X \to \mathbb{R}, \qquad x \mapsto \frac{x - \tilde{x}}{IQR_X}$$

Normalisation/Standardisation

decimal scaling. For a numerical attribute X and the smallest integer value s larger than $\log_{10}(\max_X)$, the decimal scaling is defined as

$$d : \mathrm{dom}X \to [0, 1], \qquad x \mapsto \frac{x}{10^s}$$