Data Mining Lab Course WS 2014/15

Data Mining Basics

L. Richter    M. Hecht

Department of Computer Science
Technische Universität München

Wednesday, Nov 5th
Outline

Data Definition

Preparation and Preprocessing
Attribute Types

- an object is described by variables that correspond to certain properties of the object
- in our case these variables are called attributes or features
- for example: people can be described by height, body weight, gender, age, a.s.f.
- the set of feature values describing one object is an instance
- we can distinguish various feature types:
  - nominal
  - ordinal
  - integer
  - interval-scaled
  - ratio-scaled
Categorial Types

- nominal: a variable to put an object into categories: like color, gender, profession, a.s.f. It might come in numerical form, but has no mathematical interpretations! Binary attributes can be seen as a special case with only the categories true/false, male/female, passed/failed, a.s.f. for example.
- ordinal: nominal variables with an order relationship, like small, medium, large or new borne, infant, pupil, student, adult
Continuous Types

- integer: unlike true nominal variables arithmetic is meaningful, even though have only discrete values, e.g. number of children

- interval-scaled: this is a variable that takes numerical values which are measured at equal intervals from an origin. An example is the temperature in °C. A value of 0 does not necessarily mean the absence of temperature! You can define an order on these values.

- ratio-scaled: these are similar to interval-scaled variables, but 0 means an absence of the property. Weight or size is an example for this. A value 0 means not existing.
Instance Related Issues

Why should you do a data inspection/visualization? Data used for data mining usually is not the result of an dedicated experiment but an by-product of other activities. It, e.g., can be noisy and faulty:

- noisy: non-consistent instances
- faulty: recorded values do not match the feature type or is wrong
- outliers: true exceptions or input typos?
- missing values: some feature are not applicable or were not recorded at this time
Feature Related Issues

Since the amount of possible transformations is just too high for brute force exploration you should try get some hints from the inspection:

▶ What is the distribution of an attribute (for all types)
▶ For numerical attributes: equally or normal distributed, concentrations?
▶ Are the attributes independent? Determine correlation resp. the similarity between attributes (Pearson, Manhattan, Cosine, Tanimoto, a.s.f)
▶ If you have labeled data: Dual or multi-class problem, what is the class distribution?
Useful Analysis Methods

For the detection of structure without the consideration of the class label you can use:

▶ Frequent Item Sets / Association Rules: This allows you to detect “interesting” co-occurrences of values between nominal attributes which are missed otherwise if you only look at the data column-wise.

▶ Clustering: This allows you to detect large structures in the data set, i.e. large range similarities between instances.
Possible Data Transformations

The list position of a transformation does not correspond to a superior performance

- convert numeric attributes into nominal one based on the attributes’ distributions
- remove or impute instances with missing values
- filter highly correlated attributes
- filter highly similar instances (only for very large data sets)
- down size very large data sets by sampling
- combine or split binary/nominal attributes