Data Mining Lectures - Case Study

Piotr Wąsiewicz

Institute of Computer Science

pwasiewi@elka.pw.edu.pl

22 maja 2017

Case study

Data preparation

- Attributes not related to the class target have to be removed.
- Attributes with too many missing values (NA) may be removed or go through imputation - filling missing values with means or medians for continuous attributes or with modes (most frequent values) for discrete ones.
- Attributes with many outliers may be removed or another way is to get rid of these outliers or to use median.
- For some algorithms attributes can be standardized $(x - mean)/sd$ or/and normalized $(x - min)/(max - min)$.
- Attributes strongly correlated should be removed except the best one e.g. with less missing values (NA) or with a fewer outliers.
- Attributes with too many discrete values can aggregate them to max number of 32 (it is suitable for many random forest algorithms).
- The given dataset is divided into three sets: training, validating and testing data.

Case study

Training phase

- Training of your model based on the training dataset.

Validation phase

- Estimation how well your model is trained and how to find model best properties, training algorithm parameters.

Testing phase

- At the end of the process checking quality of the trained and validated model using the testing dataset.