Lab 1: Data preprocessing

13.03.2014

Laboratories are aimed at practicing the basic pre-processing methods and analysis of data: data cleaning, descriptive statistics, the histograms, the identification of outliers and finding the relationship between variables.

1. Download and upload churn-error-data.csv to Excel.
2. Identify and clean the bugs planted in the data (there are seven). These are examples of errors that may occur when entering data into the form.
3. Download and upload churn-data.csv to Excel.
4. Calculate simple statistics: mean, median, standard deviation, minimum and maximum for numerical attributes (note: not all statistics always make sense). What we can tell when the average and median are equal or different from each other?
5. Create histograms of account length, the number of voice mail messages and the number of customer service calls. What is the difference between these histograms?
6. Identify (using information found in the previous paragraphs) and discard the three outliers. How the descriptive statistics and histograms have changed?
7. Determine the correlation coefficients between the following pairs of attributes:

   - total day minutes vs. total day charge
   - total intl minutes vs. total intl charge
   - total day charge vs. total intl charge
   - account length vs. number of customer service calls

   What conclusions can be drawn?
8. Create an array of contingency and the value of statistics χ^2 between the following pairs of attributes:

   - voice mail plan vs. churn
   - international plan vs. churn
   - state vs.churn

   Which attribute is more suitable for prediction of churn attribute?
9. Download and upload churn-orig-data.arff to WEKA.
10. Go to 'Select attributes' tab and run ChiSquaredAttributeEval with default settings. Check which of the attributes are the best to describe the churn class attribute.