Lab 4: Classification Case Study

17.04.2014

Laboratories are aimed at practicing the ability to solve a classification problem.

1. Given is a data set with the description of wealth of USA citizens:
   data set.
2. The main goal of this task is to discover which attributes affect the prediction of wealth. This means that we are interested in classification whether the income is higher or lower than 50000$.
3. This data set consists of 25000 learning examples described with 14 conditional attributes (age, workclass, demogweight, education, education-num, maritalstatus, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, nativecountry) and 1 decision attribute (income).
4. The domains of conditional attributes are as follows:
   - age: continuous
   - workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
   - fnlwgt: continuous
   - education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
   - education-num: continuous
   - marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
   - occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
   - relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
   - race: Amer-Indian-Eskimo, Asian-Pac-Islander, Black, White, Other
   - sex: Female, Male
   - capital-gain: continuous
   - capital-loss: continuous
   - hours-per-week: continuous
   - native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
5. Create an arff file with the data describing the wealth of USA citizens.
6. Perform the data preprocessing: check the quality of the data and clean the data if necessary.
7. Check whether all conditional attributes are necessary to describe the decision attributes (perform all necessary calculations).
8. Build the model for the given classification problem.
9. Pay attention to the level of imbalance in the decision attribute. Calculate sensitivity and specificity measures, where positive class is income higher than 50000$.

10. Write a report containing:

    - introduction with the short description what is a classification problem and what is the main goal of this case study
    - description of performed preprocessing techniques
    - description, which conditional attributes are chosen and why (attach all performed calculations and their results)
    - full descriptions of tested method: J48, PART, JRip, NaiveBayes, IBk and Bagging
    - description of the chosen evaluation method
    - experimental results with calculated additional evaluation measures (sensitivity and specificity, where positive class is income higher than 50000$)
    - discussion of obtained results
    - conclusions