UNIVERSITY OF MANNHEIM

Data and Web Science Group
Prof. Dr. Christian Bizer
B6 – B1.15
68159 Mannheim

Data Mining – FSS 2020

Exercise 5: Classification

5.1. Parameter optimization

In Exercise 4.2 we have used the German credit data set from the UCI data set library (http://archive.ics.uci.edu/ ml/index.html), which describes the customers of a bank with respect to whether they should get a bank credit or not. The data set is provided as *credit-g.arff* file in ILIAS. You need to use the RapidMiner *ARFF reader* operator to import the data set. Please also have a look at the data set documentation that is included in the file.

1. (recap) Go back to the results of exercise 4.2.4, in which you have compared Rule Induction, k-NN (k=5) and Decision Tree classifiers. In that exercise you
    a. Used the 10-fold X-Validation approach.
    b. Balanced the training multiplying the "bad customer" examples by using *Filter Examples* and *Append* operators.
    c. Used the *Performance (Costs)* operator to evaluate the results, setting up your cost matrix to ((0,100)(1,0)) – that is, you assumed you will lose 1 Unit if you refuse a credit to a good customer, but that you lose 100 Units if you give a bad customer a credit.

    Rerun your process to get the performance results. What were the default parameters of the *Decision Tree* operator?

2. Now try to find a more optimal configuration for the Decision Tree operator. Use Nested X-Validation and the *Optimize Parameter* Operator with an inner X-Validation. Then use the "Edit Parameter Settings" option of this operator to let RapidMiner test different combinations of parameters. Try the following parameters of the *Decision Tree* operator:
    - CRITERION (information_gain, gain_ratio, gini_index, accuracy)
    - APPLY_PRUNING (true, false)
    - MAXIMAL_DEPTH (try the range from -1 to 50 with 10 steps). What does -1 mean?

    You should come up with 88 (4 x 2 x 11) combinations.

    What is the best configuration for the data set and the classification approach?

3. What is the misclassification cost for this configuration? Can the same cost be obtained with other setups? Use the *Log* operator to find out.
4. How does the optimal decision tree differ from the one you have learned in 4.2.4?

5.2. Open Competition: Finding rich Americans

The Adult data set from the UCI data set library (http://archive.ics.uci.edu/ml/datasets/Adult) describes 48842 persons from the 1994 US Census. The data set is provided as *adult.arff* file on the website of this course.

Your task is to find a good classifier for determining whether a person earns over 50.000 $ a year. Beside of being accurate, your classifier should also have balanced precision and recall.

To evaluate your classifiers use split validation (split ratio=0.8, linear sampling).

In order to find the best classifier, you may experiment with:

1. different algorithms
2. different parameter settings
3. the balance of the two classes in the data set
4. the set of attributes that are used or not used
5. other preprocessing techniques

People are described by the following 14 attributes:

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	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
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.

In order to increase your understanding of the data set, you might want to visualize different attributes or attribute combinations.

5.3. Have RapidMiner help you!

RapidMiner offers operators to automate the process of

- feature selection (automatically try different attribute combinations in order to find the attribute combination that works best for learning) as well as
- to automatically try out different parameter settings of the learning algorithms in order to find the parameter setting that results in the best performance of the learned model.

- Rapidminer Tutorial on Accidental Contamination through Feature Selection and Parameter Optimization
    - https://rapidminer.com/blog/learn-right-way-validate-models-part-4-accidental-contamination/