Data Science and Database Technology

Practice #3 – Data mining

Objective

Exploit data mining classification algorithms to analyze a real dataset using the RapidMiner machine learning tool.

Dataset

The Utenti dataset (Utenti.xls, downloadable at http://dbdmg.polito.it/wordpress/teaching/data-science-and-database-technology/) collects census data about American users of a given company. Users are classified as "basic" or "premium" according to their commonly asked services. Each dataset record corresponds to a different user. The dataset collects around 32,000 different users, including some personal user (e.g., age, sex, workclass) as well as their corresponding class. The class attribute, which will be used as class attribute throughout the practice, is reported as the last record attribute.

The complete list of dataset attributes is reported below.

(1) Age
(2) Workclass
(3) FlnWgt
(4) Education record
(5) Education-num
(6) Marital status
(7) Occupation
(8) Relationship
(9) Race
(10) Sex
(11) Capital Gain
(12) Capital loss
(13) Hours per week
(14) Native country
(15) **class (class attribute)**

Context

Analysts want to predict the class of new users, according to the already classified user characteristics. To this purpose, analysts exploit three different classification algorithms: a decision tree (Decision Tree), a Bayesian classifier (Naïve Bayes), and a distance-based classifier (K-NN). The Utenti dataset is used to train classifiers and to validate their performance.

Goal

The aim of this practice is to generate and analyze different classification models and validate their performance on the Utenti dataset using the Rapid Miner tool. Different Rapid Miner processes have to be developed. To evaluate classification performance, different configuration settings have to be tested and

compared with each other. A 10-fold Stratified Cross-Validation process must be used to validate classifier performance. Results achieved by each algorithm should be analyzed in order to analyze the impact of the main input parameters.

Questions

Answer to the following questions:

1. Learn a Decision Tree using the whole dataset as training data and the default configuration setting for algorithm Decision Tree. (a) Which attribute is deemed to be the most discriminative one for class prediction? (b) What is the height of the generated Decision Tree? (c) Find an example of pure partition in the Decision Tree generated.
2. Analyze the impact of the minimal gain (using the gain ratio splitting criterion) and maximal depth parameters on the characteristics on the Decision Tree model learnt from the whole dataset (keep the default configuration for all the other parameters).
3. What happens if we change the class label from “Service class” to “Native Country”? Answer again to question (1) in this new scenario.
4. Considering again the service class as the class attribute and performing a 10-fold Stratified Cross-Validation, what is the impact of the minimal gain and maximal depth parameters on the average accuracy achieved by Decision Tree? Compare the confusion matrices achieved using different parameter settings (keep the default configuration for all the other parameters).
5. Considering the K-Nearest Neighbor (K-NN) classifier and performing a 10-fold Stratified Cross-Validation, what is the impact of parameter K on the classifier performance? Compare the confusion matrices achieved using different K parameter values. Perform a 10-fold Stratified Cross-Validation with the Naïve Bayes classifier. Does K-NN perform on average better or worse than the Naïve Bayes classifier on the analyzed data?
6. Analyze the Correlation Matrix to discover pairwise correlations between data attributes. In light of the results achieved, does the Naïve independence assumption actually hold for the Utenti dataset?

Practice

Program setup

- Run the Rapid Miner application under Windows XP

Process building and analysis

- Create a new Rapid Miner process.
- Build the data mining flow by dragging the operators available on the left-hand side menu and dropping them into the main process window.

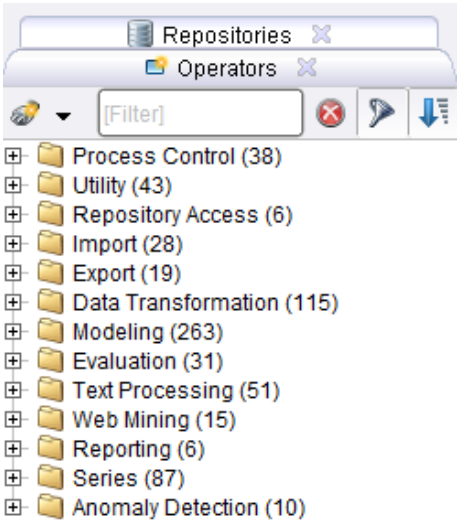

Figure 1. Operators

- To handle process execution, use the Start/Stop/Pause buttons. To view the results, change the perspective from *Design* to *Results*.

Figure 2. Execution/perspective change buttons

- Look into the content of the Utenti dataset, which is available in the Excel format (.xls).
- Import the source data into the Data Mining process by using the operator “Read Excel”. To import data use the *Data Import Wizard* as follows:
    - Select the source file (Step 1).
    - Select all the spreadsheet content (Step 2).
    - Annotate the first row as the attribute name (label “name”), while keeping all the remaining rows unlabeled (“-”) at Step 3.
    - Bind the data import block with the data source. Identify the role of attribute “Service class” as “label” attribute (Step 4).
- Include classifier “Decision Tree” at the end of the data mining flow. The currently generated process looks like the following one:

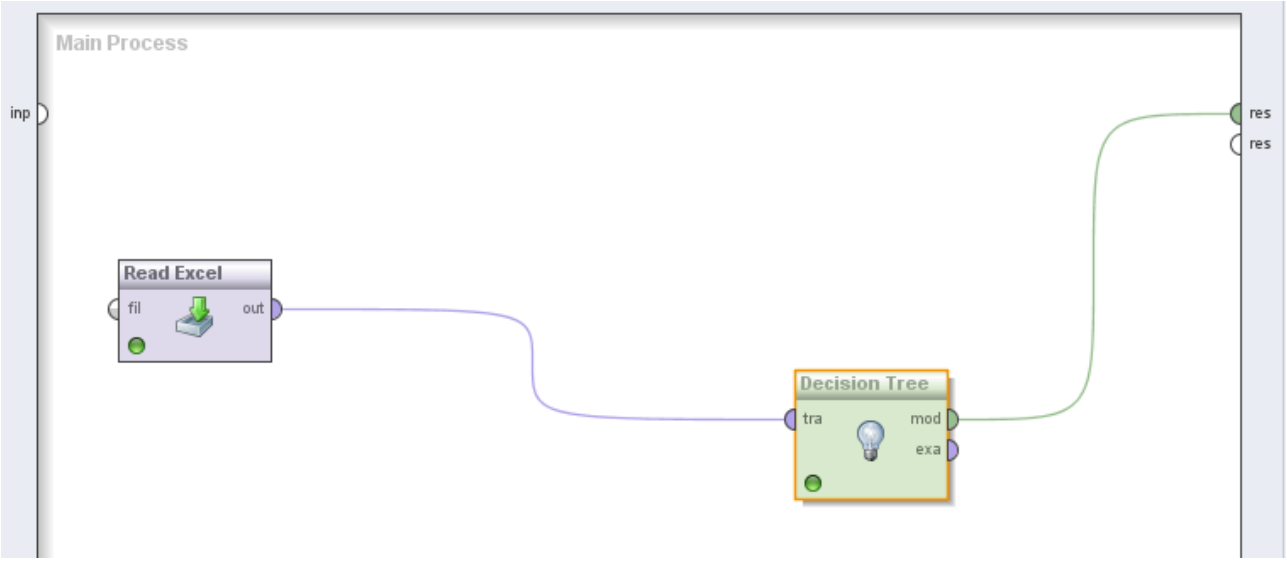

Figure 3. Decision tree classification process

- Execute the process and analyze the Decision Tree generated through the Results perspective.
- Change the configuration setting for algorithm Decision Tree clicking on the corresponding operator and using the right-hand side menu in the Design perspective. Specifically, vary the minimal gain and maximal depth threshold values to analyze their impact on the characteristics of the classification model.
- Click on the "Read Excel" operator to edit its options. Edit the "Data set metadata information" to change the class attribute from "Service class" to "Native Country" (alternatively, re-execute the data import wizard and select the new class attribute at Step 4).
- Re-execute the process to generate the new Decision Tree.
- Modify the process flow in order to perform a 10-fold Stratified Cross-Validation. To this aim, include operator "Validation" in place of Decision Tree into the main process first.

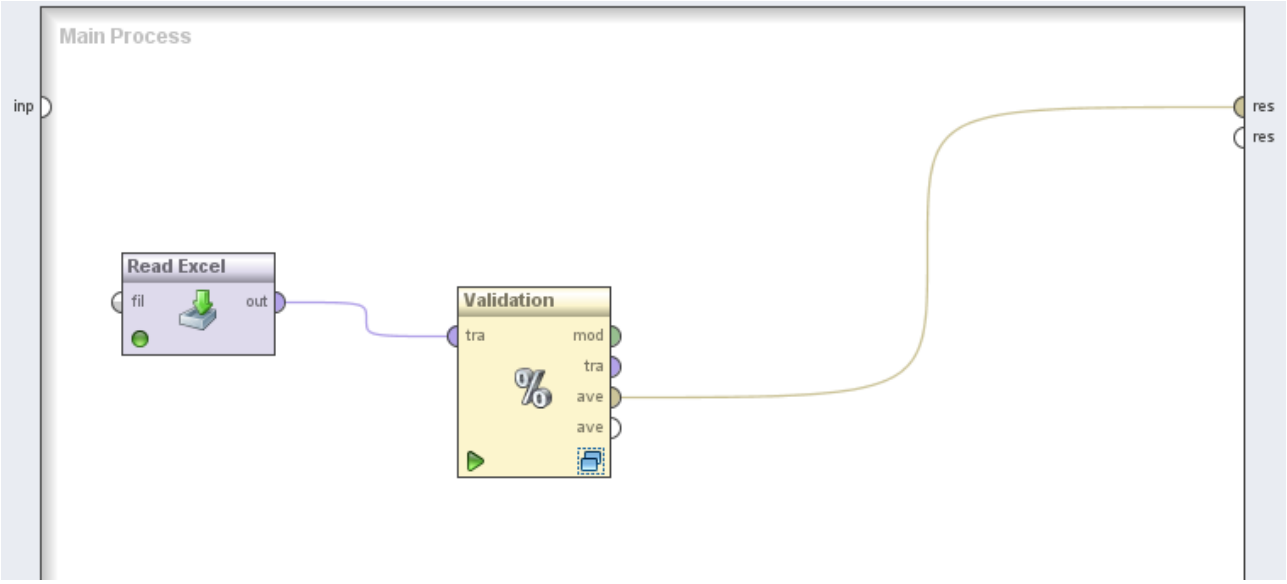

Figure 4. 10-Fold Cross-Validation process.

Next, double-click operator "Validation" and create a nested process as the one reported below:

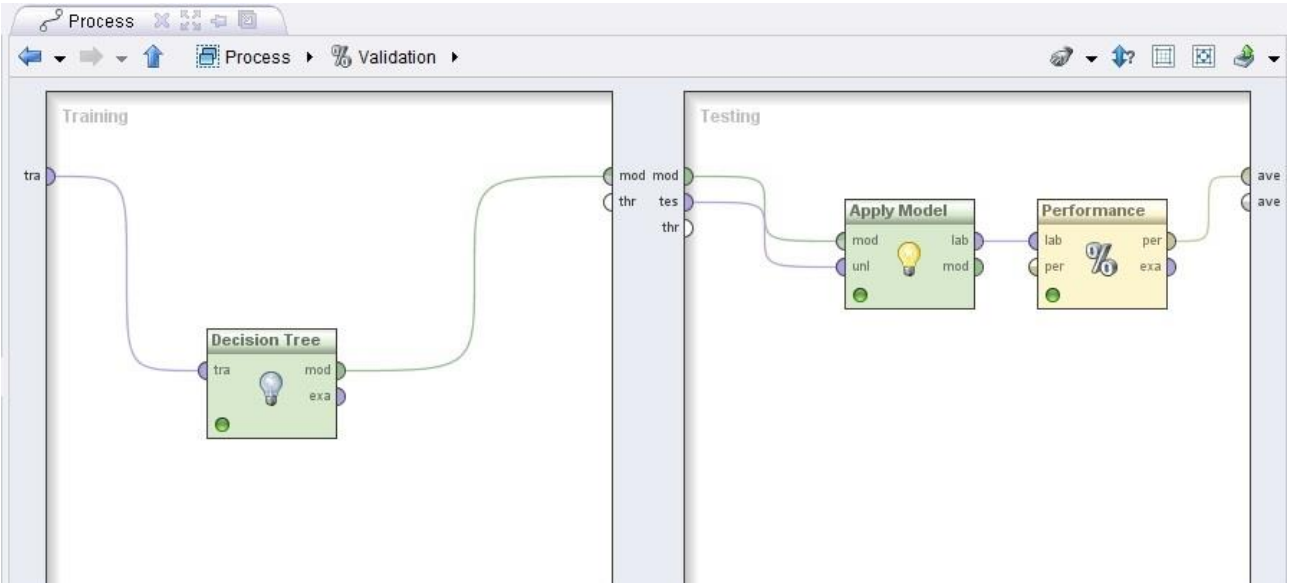

Figure 5. Validation subprocess

- Moving to the Results perspective, analyze the confusion matrix generated by the validation process.
- Temporarily disable the Decision Tree operator (right-click on the operator box and uncheck “Enable Operator”). Substitute the classifier operator with the Naïve Bayes classifier first and with the K-NN classifier next.
- Compare the performance of K-NN and Naïve Bayes performance in terms of average accuracy, precision, and recall, by analyzing the corresponding confusion matrices. For the K-NN classifier, vary parameter K values using the right-hand side menu in the Design perspective.
- To analyze the data correlation matrix associated with the analyzed dataset go back to the main process (click on button “Process”), temporarily disable the Validation operator (right-click the operator box and uncheck “Enable Operator”), insert operator “Correlation Matrix” at the end of the data mining flow, and output the corresponding matrix connecting the block plug-in labeled with “mat” to the process result plug-in. The main process looks like as follows:

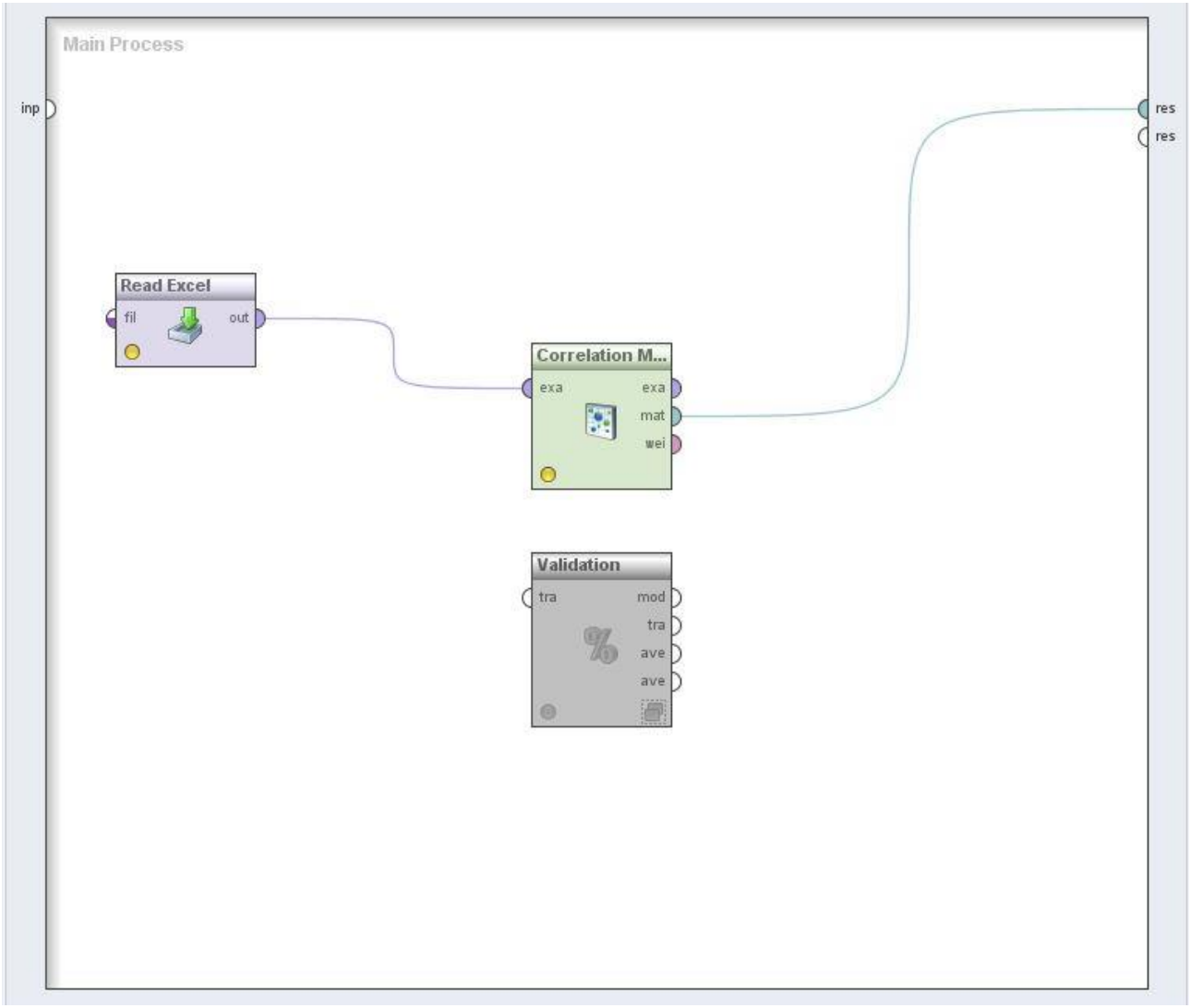

Figure 6. Correlation matrix

Moving to the Results perspective, to sort pairwise attribute correlations in order of decreasing strength select the “Pairwise Table” view and click field “Correlation” of the visualized table.