UNIVERSITÄT MANNHEIM

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

Data Mining – FSS 2016

Exercise 4: Classification

General notice: Please use “local seeds” with X-Validation and a value of “1992” to be comparable.

4.1. Rule Learning

Download the glass.arff data set, which is a default data mining data set, distributed e.g. by Weka, from the course website. The dataset was created during a study which was motivated by criminological investigations. At the scene of the crime, the glass left can be used as evidence, if the purpose of the glass can be classified correctly. You can read more about the different attributes within the .arff-File.

1. Import the file and store it into your repository, using the Read ARFF and Store operator.
2. Now use the Rule Induction classifier to learn classification rules based on the glass dataset. Use in a first step the default setup (pureness = 0.9) and have a look at the rules.
3. Change the pureness attribute to 1 and 0.5. How is the change reflected in the set of rules created?
4. Replace the Rule Induction Operator by the Tree to Rules (nested Operator) and use a Decision Tree inside. Compare the results from direct and indirect Rule-Based-Classification Algorithms, based on number of rules created and complexity of the rules.
5. Use the X-Validation (Classification) with a 10-fold setup and compare the accuracy of both rule learning approaches and a k-NN classification algorithm. What does work best on the data set?

4.2. Who should get a bank credit?

The German credit data set from the UCI data set library (http://archive.ics.uci.edu/ ml/index.html) describes the customers of a bank in respect 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. Apply the Compare ROCs Operator to the dataset and include k-NN, Decision Tree and Rules Based classification. Which classification approach looks most promising to you?
2. Include the møst promising classification approaches and try to optimize the results using a 10-fold X-Validation approach. Which level of accuracy do you reach?
3. What does the precision and recall values for the positive class “Bad Customer” tell you? Try to improve the situation by increasing the number of “bad customers” in the training set. For doing this, you first filter all bad customers from the data set and then append these customers to the original set. How does precision and recall change if you apply this procedure twice? Use the Filter Examples and Append Operators.

4. To model a use case specific evaluation, as observed in the previous example, replace the Performance (Classification) operator by the Performance (Costs) operator. Set up your cost matrix by assuming that 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 the experiments from 1 and evaluate the results.
5. As the creation of training data is mostly a manual task as humans tend to be fallible training data might include noise. Simulate this behaviour by using the Add Noise operator and change the parameter “label noise” from 0% to 10% to 20%. Is your preferred classification approach still feasible for this situation? How does the performance of the other classifiers evolve?