Part 2 - Incremental learning.

In this part, we use Forest Covertype dataset. For this exercise, the dataset has been divided into five separate sets, which you should download (covtypeNorm100000.arff , covtypeNorm200000.arff, covtypeNorm300000.arff, covtypeNorm400000.arff, and covtypeNormTest.arff). Data contains the forest cover type for 30 x 30 meter cells. The training set is divided into four files, containing 100 000 instances each. The test set has 181 012 instances. The amount of attributes is 55 (54 + class). There are seven classes: 1, 2, 3, 4, 5, 6, and 7.

What to report?

Answer the following questions and report the answers as well as the required Weka outputs in your exercise report.

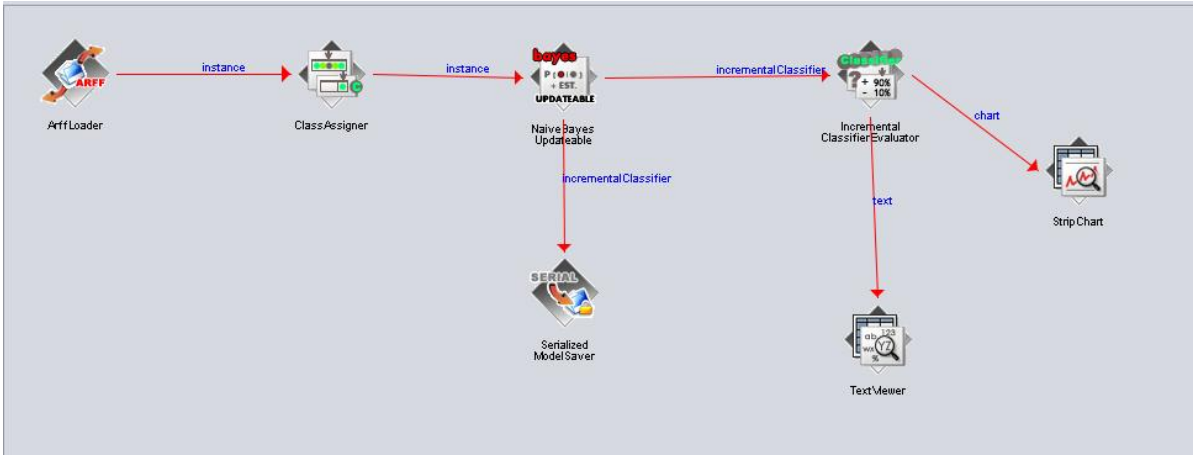
Questions

1. Incremental learning. (Weka manual, section 7.1 "Introduction" and subsection 7.4.3 "Processing data incrementally" and/or other sources)

(a) How does incremental learning differ from conventional batch learning?

(b) Incremental learning algorithms can be evaluated using interleaved test-then-train evaluation procedure. How is it performed?

2. Start Weka and choose KnowledgeFlow as GUI. Construct a knowledge flow structure similar to the one presented in Weka manual, subsection 7.4.3 "Processing data incrementally". Follow the instructions in the manual but do not start the flow yet. (Connect ArffLoader and ClassAssigner using an "instance" connection as in the figure even though textual instructions propose a "dataSet" connection.) That knowledge flow structure allows incremental training and testing of a naive Bayes classifier. Use “SerializedModelSaver” (which can be found in “DataSinks”) to save the incrementally trained model. In the options of the “SerializedModelSaver”, change the output directory to something else than the Weka installation directory. After the first flow, you can configure your model by specifying its location in the field “Classifier model to load” in the “NaiveBayesUpdateable options”, which can be found by right clicking "NaiveBayesUpdateable" and selecting "Configure". The structure should look like this:



Configure "StripChart" by increasing its "refreshFreq" to, for example, 10 000.

Note that the classifier is learning more whenever it is in training mode and some data passes through it. So, please, follow the instructions below carefully in order to avoid learning data that was not meant to be learnt.

(a) Right click "NaiveBayesUpdateable" and select "Configure". Check that the additional option "Update incremental classifier" is selected as True (i.e. the classifier is in training mode). Then use "ArffLoader" and browse the first file for training, "covtypeNorm100000.arff". Start the flow. Check the "Performance information" from "TextViewer". Report the number of correctly and incorrectly classified instances.

(b) Now, set the additional option "Update incremental classifier" as False (i.e. change the classifier into testing mode). Remember to configure the trained classifier model by specifying its location in the field “Classifier model to load” in the “NaiveBayesUpdateable options”. Browse the test file "covtypeNormTest.arff" and start the flow. Report the number of correctly and incorrectly classified instances.

(c) Change the classifier back to training mode in order to continue learning. Browse the second file for training "covtypeNorm200000.arff" (the next 100 000 samples). Start the flow. Report the number of correctly and incorrectly classified instances.

(d) Change the classifier into testing mode in order to test the updated classifier. Browse the test file "covtypeNormTest.arff". Start the flow. Report the number of correctly and incorrectly classified instances.

(e) Change the classifier into training mode and browse the third file for training "covtypeNorm300000.arff". Start the flow. Report the number of correctly and incorrectly classified instances.

(f) Change the classifier into testing mode and browse the test file "covtypeNormTest.arff". Start the flow. Report the number of correctly and incorrectly classified instances.

(g) Change the classifier into training mode and browse the fourth file for training "covtypeNorm400000.arff". Start the flow. Report the number of correctly and incorrectly classified instances.

(h) Change the classifier into testing mode and browse test file "covtypeNormTest.arff". Start the flow. Report the number of correctly and incorrectly classified instances.

3. How does the naive Bayes model evolve in terms of accuracy when the amount of training instances increases?

4. Discussion and conclusions.

Last modified: 10.4.2017 by Risto Vehmas