Mayday Machine Learning HowTo

Stephan Symons

Contents

1 Data Prerequisites

Using the Mayday machine learning plugins requires a microarray data set that can be read by Mayday, see Mayday Documentation for suitable formats and a partition of the experiments to two or more classes, preferably in a file with associations of classes and experiments (in a plain tab-separated format, see figure 1).

```
Experiment1 class1
Experiment2 class2
...
ExperimentX class1
```

Figure 1: Example of a class label file

2 Class Selection

For selecting class assignments, the Mayday Class Selection Dialog is used. The class Selection Dialog allows to manually enter class labels, to automatically assign

labels and to load labels from files.

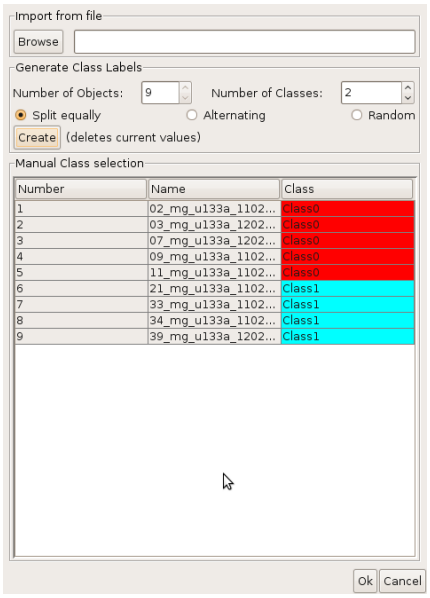

Figure 2: Mayday Class Selection Dialog

Manual class assignment Class names can be entered into each row of the field labeled "Class" in the table (see figure). To enter a class label, double-click on the field in the respective row and type in the name. After a class label is entered, to assign this label to other experiments, right-click on one ore selected experiments and choose the required label from the menu .

Automatic class label assignment Three ways of automatic assignment are available: "Split equally' assigns the same number of consecutive objects to each class. "Alternating" assigns an equal number of experiments rotationally to each class. "Random" assigns class labels randomly. See figure 2 Then number of class labels can be edited, and at any time, class labels can be manually adjusted.

Class labels from file Click on browse to select a file. The class labels are automatically loaded from the file. Afterwards, class labels can be manually adjusted.

```
Number Name Split equally  Alternating Random
1 Experiment1 class1 class1  class2
2 Experiment2 class1 class2 class1
3 Experiment3 class1 class1 class1
4 Experiment4 class2 class2 class2
5 Experiment5 class2 class1 class2
6 Experiment6 class2 class2 class1
```

Figure 3: Example of automatic class assignment modes

3 Overview of Machine Learning Plugins

Mayday features several plugins for supervised and unsupervised machine learning tasks. All are based on Weka [2]. In this document, a brief overview of the plugins is given. Then, the usage of the classification and classifier evaluation tools is discussed.

3.1 Association Mining

Mayday's Association Mining plugin allows to find associations between probes. For this purpose, probes are first discretized. Several discretization methods are available: Equal width Binning (each bin covers the same range of values), Equal Frequency Binning (each bin has the same size), Mid-Ranged Discretization and TSD Discretization. For details on these methods, cf [1]). Binary or Ternary binning can be used. For
As association mining methods, Apriori, Predictive Apriori and Tertius can be used (see [2] for details on those).
The result of association mining is presented as a graph, or as raw output of the found rules. The result s can also be exported to the Graphviz dot format.

3.2 Feature Selection

The Mayday Feature selection plugin allows to reduce a dataset's dimensionality. Two general strategies are available:

- In the **Filtering** approach, one probe is evaluated at a time, allowing to reduce the probes to a predefined number of probes with the highest figure of merit. For this purpose, one or more predefined size can be entered in the "Feature Selection Dialog" (see figure 4). The following methods are available:
  - Information Gain
  - Gain Ratio
  - X^2-Statistic
  - One Rule
  - ReliefF
  - SVM
  - Symmetrical Uncertainty
- The Wrapper approach employs a search strategy for finding an optimal subset of the probes. Optimality criterion and search strategy are exchangeable. Mayday offers the following methods available from WEKA:
  - Search strategies: Exhaustive, Best First, Generic, Greedy, Race, Random, Rank Search.

– Subset evaluation methods: CFS, Consistency, Unsupervised, Classificator Wrapper.

3.3 Training

Reusable Classification models can be trained using Mayday's training plugin. It allows to select and configure a vast number of classification models (via the classifier tree, see figure 7). The training method can be chosen between resampling test, test set evaluation, percentage split, n-fold cross valid and leave one out cross validation. Available classification models include: SVM, Decision Trees, kNN, Naïve Bayes, Rule-based methods and several meta models, combining other methods. Trained models can be saved to disk, and inspected using a multitude of performance measures.

3.4 Classification

The classification plugin uses prepared modes (from the Training Plugin) to classify new data. Results are displayed and can be exported to Excel spreadsheet format.

3.5 Batch Training

The Mayday Batch Training Plugin is run on several probelists. It allows to train and evaluate several classification models on the selected ProbeLists at a time.

3.6 Evaluation

The Mayday Classifier Evaluation plugin is an extension of the batch training plugin. It includes the process feature selection and allows to use the "honest evaluation" strategy, which attempts to give more realistic performance measures if cross-validation is used. All classifiers and all filter-style feature selection methods can be applied, combined with the above-mentioned evaluation methods. Testing can be repeated several times. A large number of evaluation measurements can be used to estimate classifier performance.

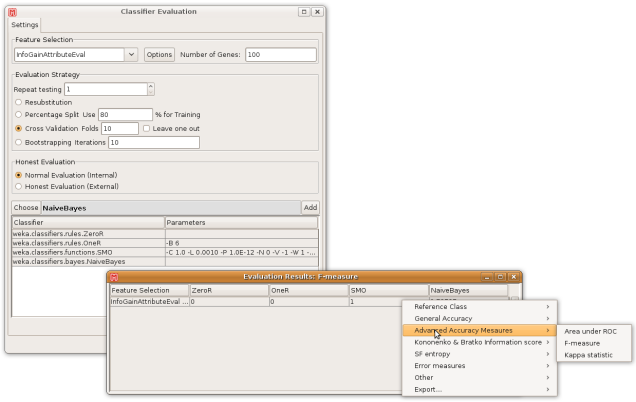

Figure 4: Left: Classifier evaluation dialog. From top to bottom: Feature Selection settings, Evaluation Strategy settings, Honest Evaluation setting (decides when to apply feature selection), Classifier Chooser and list of selected classifiers. Bottom: Evaluation result showing the selected measurement for the classifier.

4 Feature Selection

A common first task in machine learning is reducing the dimensionality of the dataset. Using any open dataset in Mayday, select a probe list, preferably the "global" probe list and right-click it. Run the "Feature Selection" plugin (Category Data Mining Classification). It is has a magnifying glass as an icon.
Assign class labels as appropriate from this dialog and click on "ok".
Now the Mayday dataset will be converted to a Weka dataset in order to prepare it for feature selection. When this is done, the "Feature Selection" dialog shows (→ figure 4). In this dialog, the user can access all feature selection tools available. We will use a filter strategy for feature selection, and try for ease of exposition only the Information Gain evaluator. Select the "Evaluators" radio button, and choose the "InfoGainAttrEval" from the drop down menu. Also, check the "Select best attributes" check box and enter the value "50" in the text field right to it. Feature selections are handled as Mayday probe lists. You might therefore want to edit the name of the new probe list under General Settings, for example to InfoGain 50. You can also set the color to anything other than the default black. Multiple selections can be created by entering additional values into the "Multiple Selections" field. Now, click on the Ok button. You will notice, that a new probe list emerged in Mayday, with name and color according to your settings and containing the best 50 probes according to Information Gain.

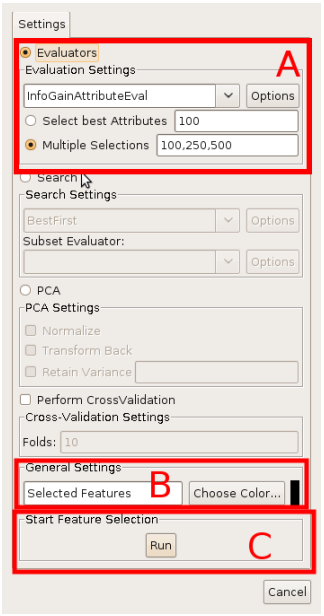

Figure 5: Mayday Feature Selection Dialog. In Panel A, the probe evaluation method can be selected, as well as the target probe size. In B, the resulting ProbeList name can be set. Button C starts the feature selection process.

5 Batch Evaluation

Now that we have some feature selections, we should see what classifier is the best on this dataset. The mass training plugin is the best for this purpose. Select all probe list you created during feature selection and run the "Weka Batch Training" plugin. You will again be queried for the class labels. Click on the "Ok" button to continue.
Now, the Batch Training dialog is visible (→ figure 6). In this dialog, you can create a list of classifiers with different settings, We will test four classifiers in the following: a linear SVM, a kNN classifier with k = 3, a C4.5 decision tree and a One Rule. First, we will add the linear SVM. Locate the SVM in the tree structure on the left ("Functions" branch). The name and default setting of the classifier is displayed in the "Current classifier" field on the upper right. We do not need to change the settings of this classifier, and therefore can just click on the "Add" button. Then, the SVM is added to the list at the bottom of the dialog. Note that classifiers in the list can be edited and removed. The list can be saved to a file for later use. Next, a kNN classifier (in the "Lazy" branch) should be added. The default parameter for k is 1. Edit this by clicking on the Options button and set the "KNN" parameter to 3. Also add this classifier to the list. Next add a C4.5 ("Trees") and a One Rule ("Rules") classifier to the list. The classifier list is now ready. The default evaluation procedure is ten-fold cross validation. The absolute number of errors will be reported. These settings are reasonable for now. Click on "Run" to start the batch training. Now, each classifier is trained and evaluated on each selected probe list using 10-fold cross validation. Eventually, a "Batch training result" dialog (→ figure 6) emerges. The results are presented in the table.

6 Classifier Training

In the previous section, we learned which classifiers should work good on the dataset. Now, we want to produce a classifier which we can use to classify new data. To do this, we need to use the "Weka Training" plugin. Based on the results of the previous chapter, it is best to run it on the "InfoGain 250" probe list. Doing so, the user is queried again for class labels.
In the "Training" dialog, single classifiers can be trained. In order to demonstrate the diagnostic tools offered by this plugin, it is best to start with a C4.5 decision tree. Select this classifier from the tree and click on "Start Training". When the training is completed, a trained classifier of type C4.5 with all properties as set before, is listed in the "Previously trained classifiers" list and the "Details..." dialog shows.
This dialog provides a variety of valuable information about the classifier. From this diagnostic plots (! figure 5) we can inspect the quality of the classifier. However, an SVM may be better. Select the SVM classifier from the tree and train it. The diagnostics shown in the are "Details..." dialog are indeed better for the SVM than for the C4.5. We now want to use the SVM classifier to classify a new dataset.

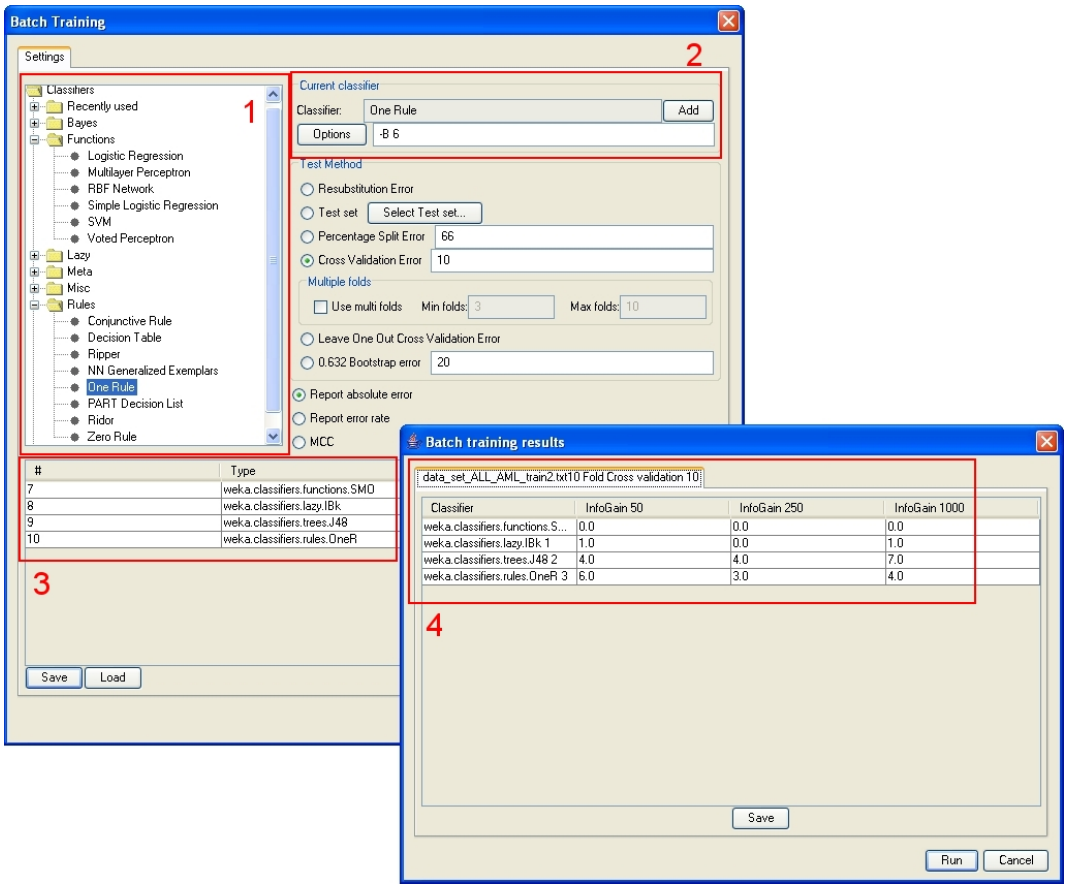

Figure 6: "Batch Training" dialog and "Batch Training Results". (1): Select classiers from this tree; (2): Edit options and add the classier to the list; (3): List of classiers to test (4): Result of the tests. The lists show the Java class names of the classiers.

Select the SVM in the "Previously trained classifiers" and choose "Save as Meta Information". Then, close the dialog.

Open another dataset on the same probes in Mayday. Remember that the trained classifiers know what genes they are trained on and select their genes automatically when used for classification. Therefore, run the Classification plugin on the "global" probe list. When asked for the class labels, of course none are required (simply click on "Ok"), but can be provided if available as they are useful to compare the results of the classification with the actual classes.

All classifiers saved as MIOs are automatically available in the "Classification" dialog. Additional classifiers can be loaded via the "Load" button or via drag and drop. To classify the dataset with a classifier, select it from the list: the details of the classifier are displayed in the "Classifier" field. Click on the "Classify" button to start the classification. The classification results are displayed in tabs at the bottom of the dialog.

The results can be saved as a spreadsheet file by right-clicking on the results table and choose "Export predictions to Excel" button. You now have successfully classified a new dataset and exported the predictions for further use.

References

[1] S.C. Madeira and A.L. Oliveira. An Evaluation of Discretization Methods for Non-Supervised Analysis of Time-Series Gene Expression Data.

[2] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmannn, San Francisco, 2nd edition edition, 2005.

This Mayday How-To was written and edited by Stephan Symons. If you have comments or questions please contact the author via email, *symons@informatik.uni-tuebingen.de*. The latest version of this document can be found at *http://www.zbit.uni-tuebingen.de/pas/mayday*.

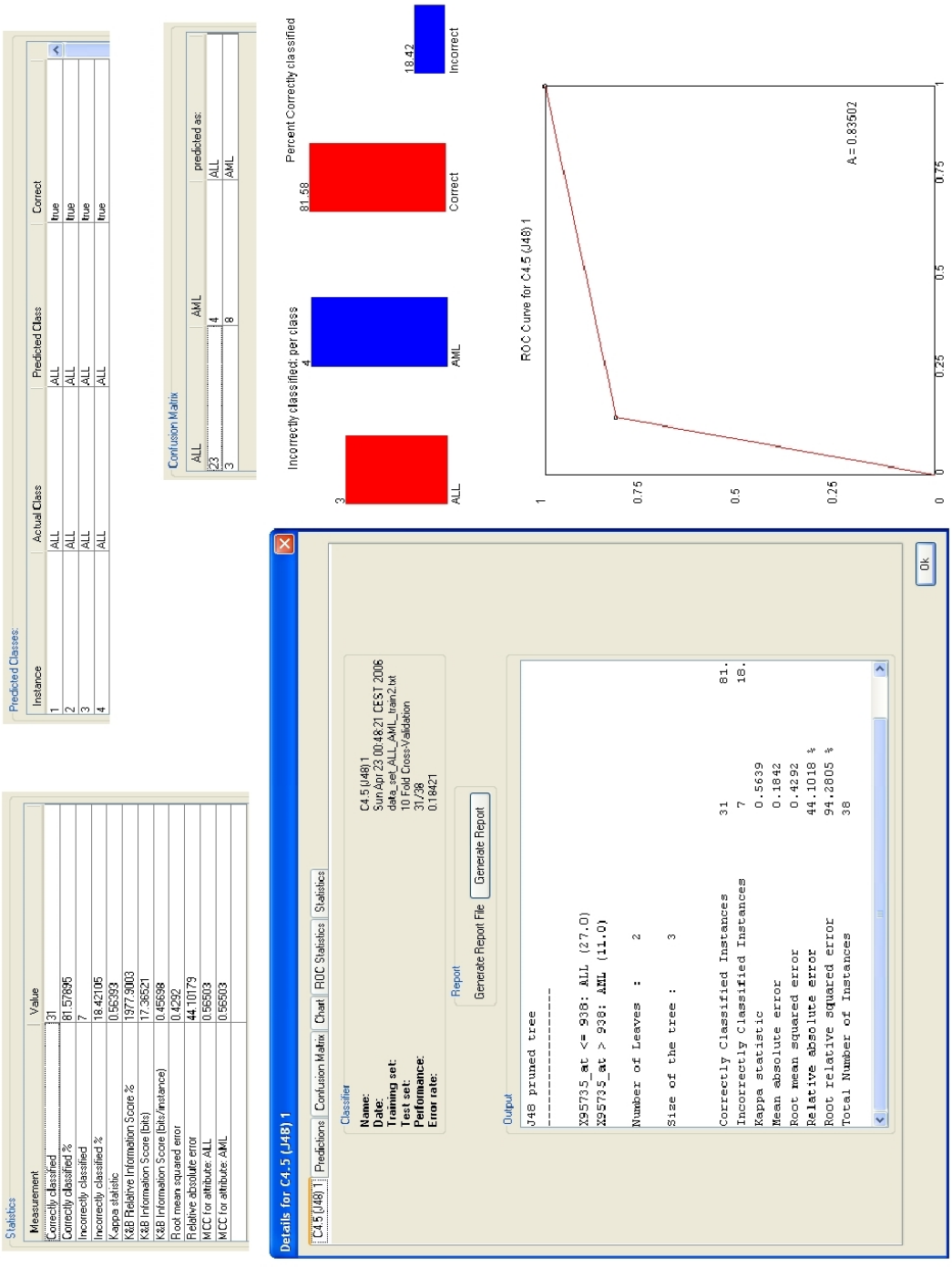

Figure 7: Classier evaluation of the C4.5 classier. The gure shows all diagnostic output for the classier: (in clockwise order) statistics, predictions on the training set, confusion matrix, charts, ROC curve, overview.