Tomaz Curk, Janez Demsar, Gregor Leban, Blaz Zupan

Installation for Mac OS X

Orange is a comprehensive data mining suite that includes components in C++, interface to scripting language Python, a number of data mining modules developed in Python, and a set of graphical user interface components called Orange widgets. The guiding ideas behind Orange are flexibility through scripting and visual programming, component-based design and ease of use.

Orange is developed as an Open Source platform. Thanks to portable code and usage of portable solutions such as Python, Qt, PyQt, Qwt, Numerical, and other can be run on Windows, Linux and Mac OS X. This document is about the installation of Orange under Mac OS X, which was beta tested under Mac OS X 10.2 and 10.3.

Installation

There are two installation files (disk images: Orange Install.dmg and Orange Doc.dmg). Put them on the desktop, double click them and drag the contents to whatever location you wish. In this way, you will copy an Orange application (bundle) and a documentation folder containing also several data sets.

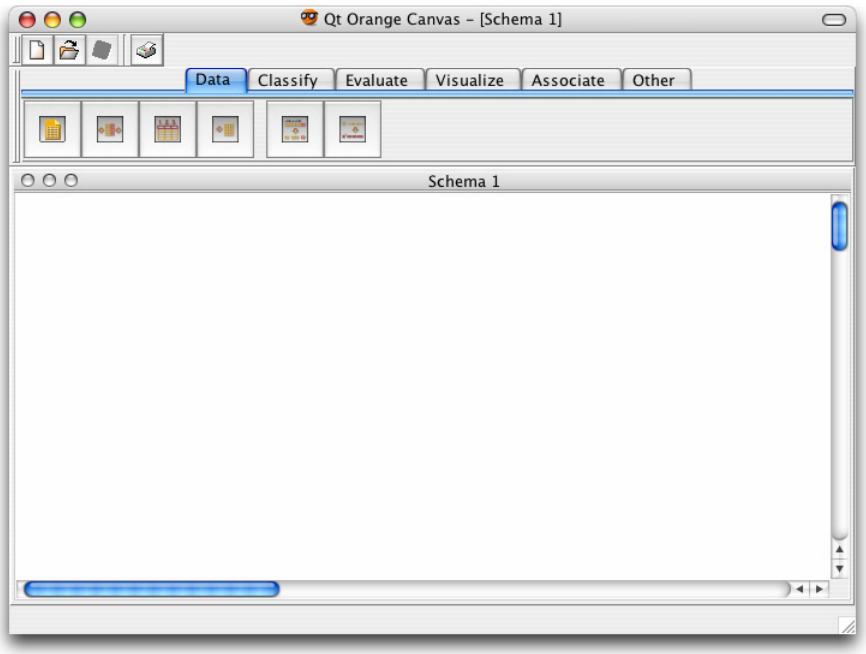
Orange installation is a fairly large file (about 15 MBytes) and besides Orange application and visual programming environment Orange Canvas contains a number of precompiled libraries and programs, including Python, Qt, Qwt, Numeric, and other.

The installation package also contains three white papers that, besides this document, include a white paper on Orange and a white paper on Orange Widgets.

Starting With Orange: Few Examples

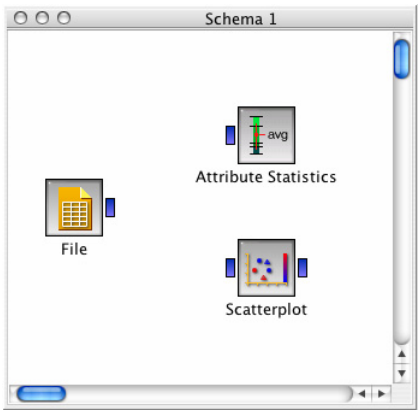
Orange Widgets most often include a part where user sets the parameters for particular data analysis or visualization method, and a part with visualization.

To run Orange, double click on Orange icon. This should open the following window.

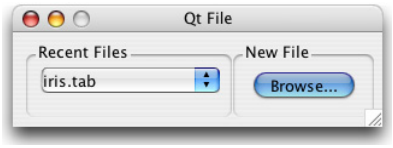


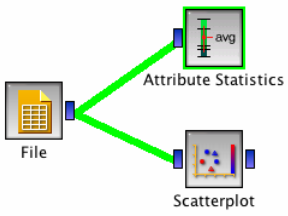
Some Simple Data Exploration

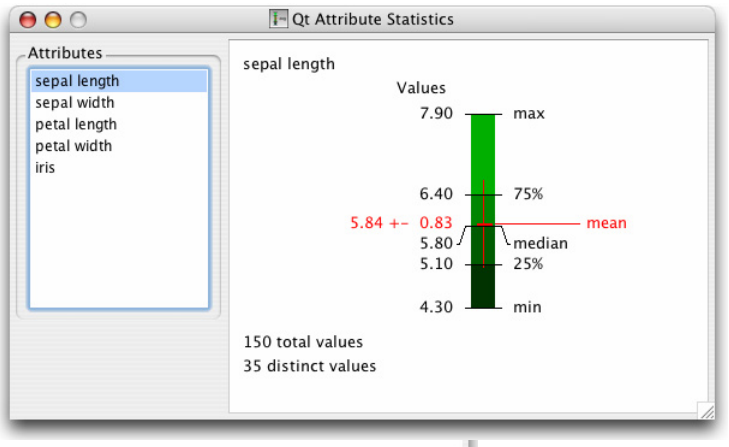
Let us now create an example schema. Put a File, Attribute Statistics, and Scatterplot widget on a canvas (finding these widgets in Data and Visualize tabs and clicking on them once will do the job). With some appropriate placing (click-and-drag) you should get something similar to what we show here.

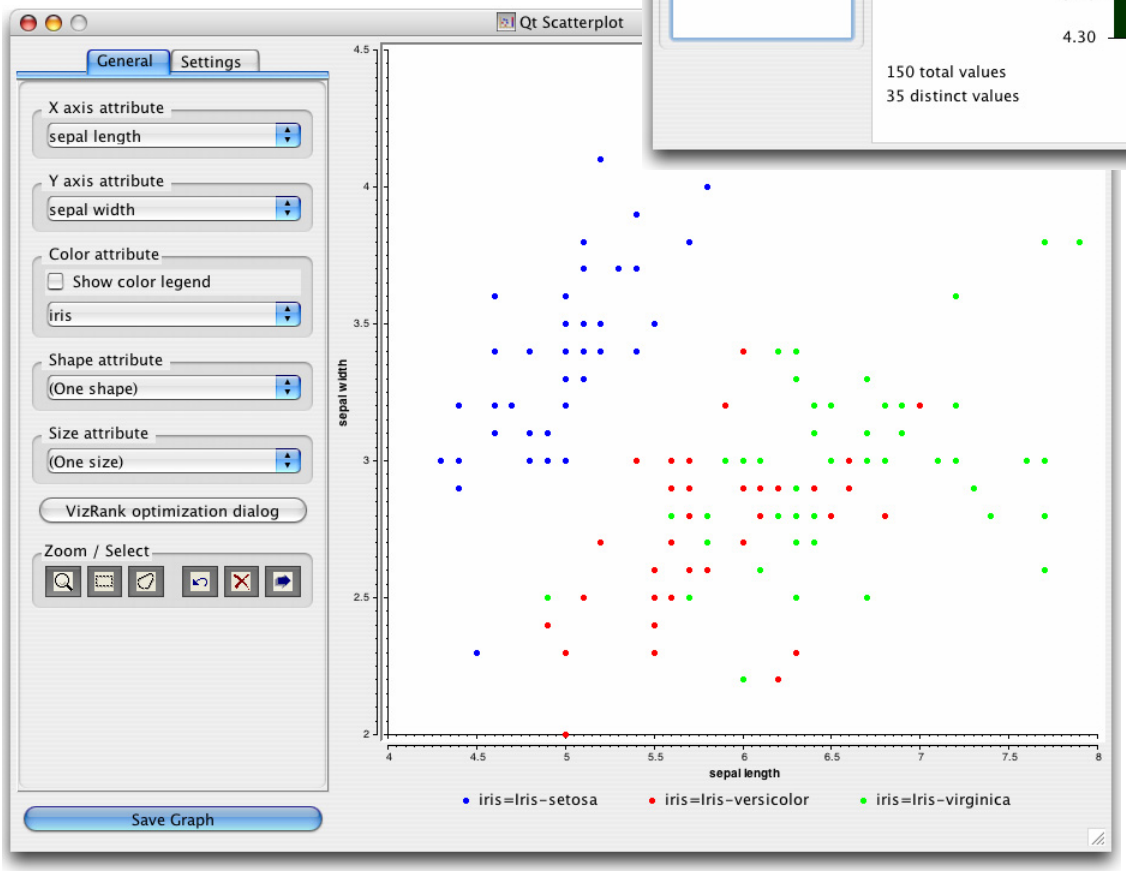


Now double click the File widget, and choose the data file called iris.tab. You will find it in Orange Doc/datasets folder. This is the data set about classification of Iris flower to three different species. Now, to see some basic statistics and visualizations, connect the three widgets as shown below (click on right blue rectangle of File widget and draw the line to each of the other two widgets). Double click the widgets to obtain the visualizations.





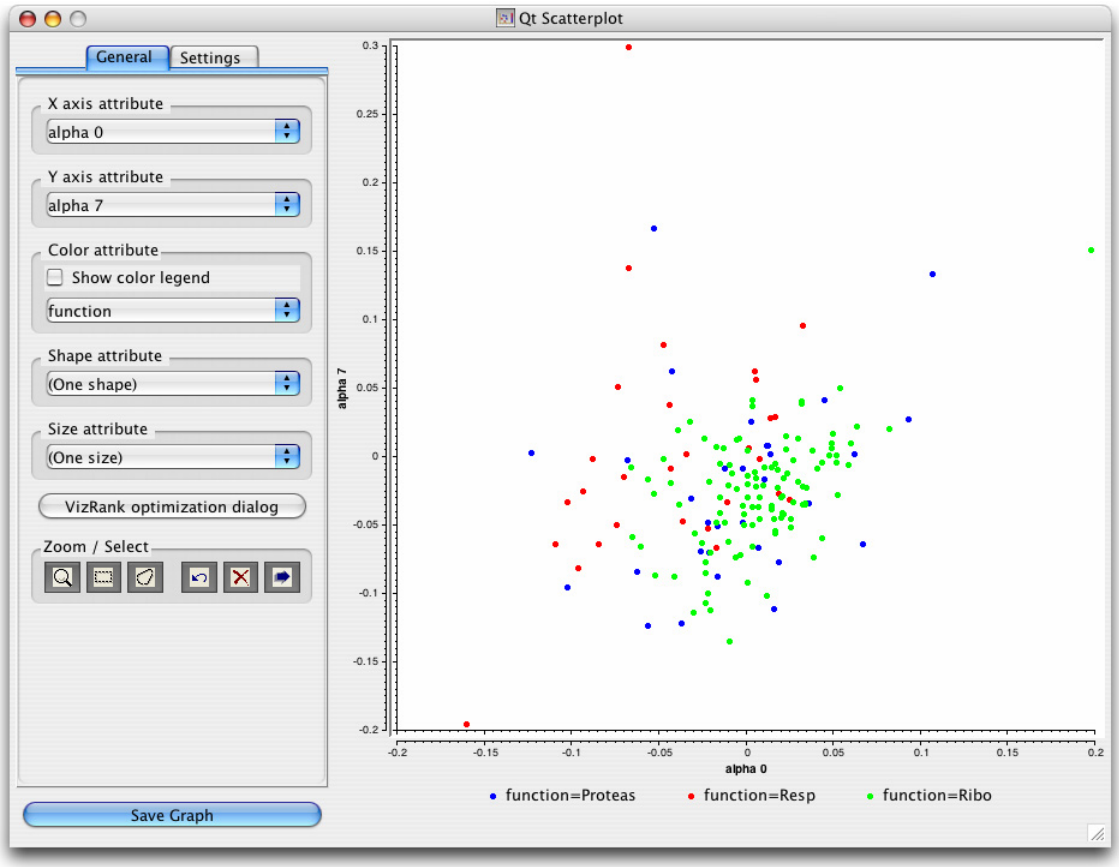




Play with the widgets and settings. To see how tokens propagate through the schema, select a different file in the File widget and observe the update of the other two widgets.

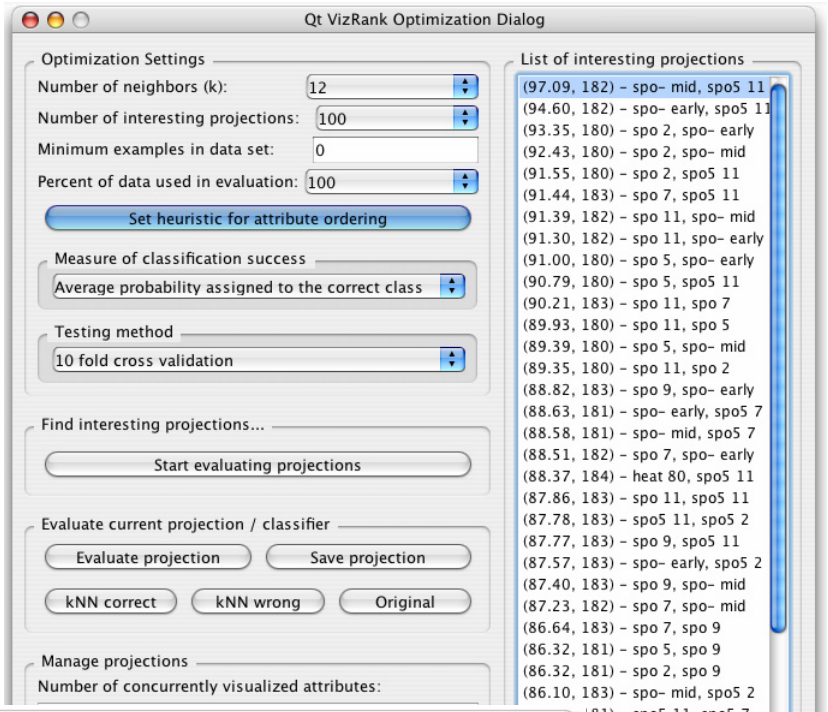
Finding the Best Scatterplot

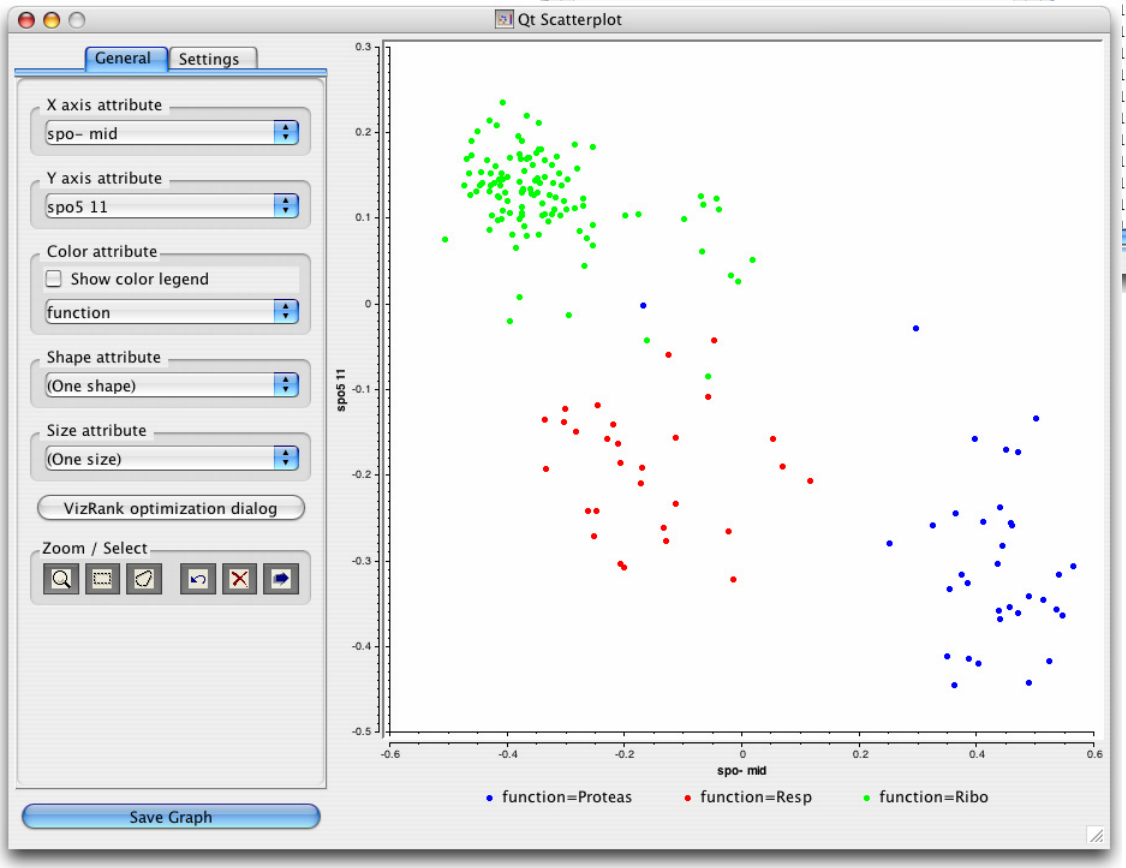
For another exercise with widgets and the same schema, choose yeast-class-RPR.tab data set. This is a data set on expressions of yeast genes, which contains 79 measuremens (attributes) and where each gene is labeled with one of three classes. The number of different scatterplots is large, and the one you will first see is rather messy. How to find a good one? In the Scatterplot widget choose VizRank optimization dialog, click on "Start evaluating projections" button, and then after few projections are proposed (list box on the right), stop the process and click on the first suggested projection. Better then the one on this page, right?



orange
DATA MINING
FRUITFUL&FUN

In Orange, Scatterplot goes beyond simple 2D visualization of the data as it includes a wizard called VizRank that can suggest which attribute pair would define an interesting projection for visualization.





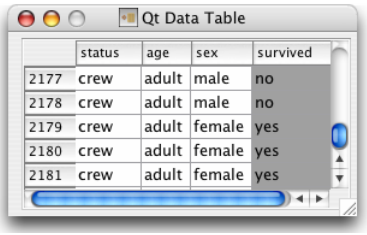
orange
DATA MINING
FRUITFUL&FUN

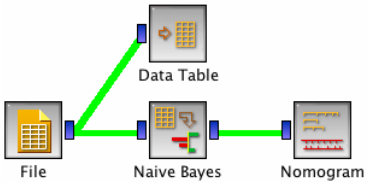
Survival on Titanic

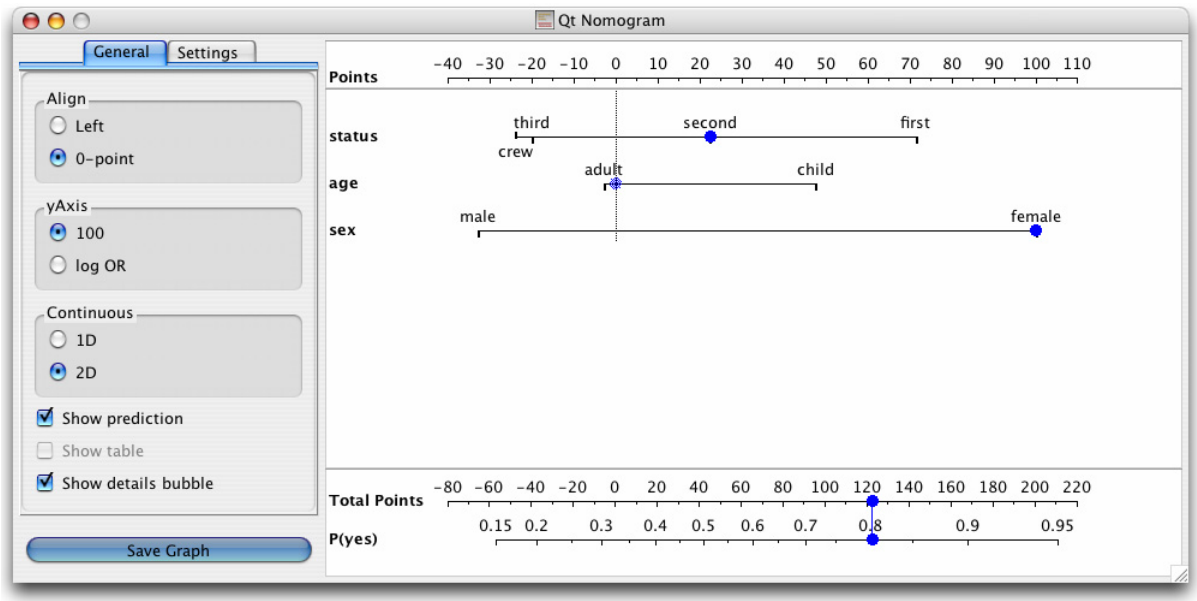
The driving idea in Orange is to complement computationally intensive widgets, like those that build predictive models from data, with widget that can visualize the results of such data analysis.

Orange includes a number of widgets that are about induction of classification rules. In essence, these widgets take a data set where each example (data instance) is labeled with one of the classes, and build a model that is able to predict the class given some information about an example.

Here is an example. Consider a data set on H.M.S. Titanic (titanic.tab included in the distribution). There were 2201 passengers, each described with three attributes: status (first, second, third class or crew member), age (child or adult) and gender. The data also records whether a particular passenger survived Titanic's accident. Orange offers various methods to build models that can predict the probability of the class values. Perhaps the simplest of them is naive Bayesian classifier, which also has a nice visualization in the form of the nomogram. To explore the Titanic data, we have used the schema on the left. Notice that the data is first sent to naive Bayesian classifier, which builds the predictive model that is then sent to a Nomogram widget. To use the nomogram for prediction of survival on Titanic, move the blue dots on the nomogram around. The snapshot shows, for instance, that women traveling in the second class had about 80% chance of survival.







What Next?

These were just a few examples to get you started. Playing with other widgets requires the users to be knowledgeable in data mining. If you are, you should go ahead: widgets are fairly easy to understand, and combining them is just like thinking about what to do in any data analysis task.

Notice that Orange is work in progress, and that this document together with an installation is in beta and prepared for Trolltech Mac Qt competition. We thank the organizers of competition for providing the Qt, which is an excellent environment, and for setting a firm deadline which pushed us towards speeding our Orange port to Mac OS X. The version submitted to the Trolltech should become available at Orange's web site (http://magix.fri.uni-lj.si, to become http://www.ailab.si/orange in the next few days when the package is released to general public).