



Data Mining Fruitful and Fun

Orange includes procedures and modules for:

- **Data input and preprocessing** (sampling, filtering, scaling, discretization, ...)
- **Classification techniques**, like trees, linear classifiers, instance-based approaches, support vector machines, ...
- **Popular regression methods**
- **Ensemble approaches**, like boosting and bagging
- **Wrappers for feature subset selection and discretization**
- **Clustering**
- **Constructive induction**
- **Model validation**, including a wide range of discrimination and calibration scoring methods

What is Orange?

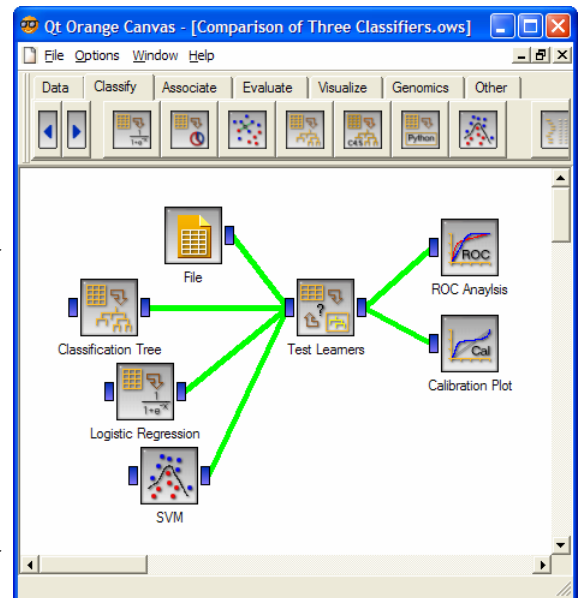
Orange is a library of C++ core objects and routines that includes a large variety of standard and not-so-standard machine learning and data mining algorithms, plus routines for data input and manipulation.

Orange is also a scriptable environment for fast prototyping of new algorithms and testing schemes. It is a collection of Python-based modules that sit over the core library and implement some functionality for which execution time is not crucial and which is easier done in Python than in C++. This includes a variety of tasks such as pretty-print of decision trees, attribute subset, bagging and boosting, and alike.

Orange also includes a set of graphical widgets that

use methods from core library and Orange modules. Through visual programming, widgets can be assembled together into an application by a visual programming tool called Orange Canvas.

All these together make an Orange, a comprehensive, component-based framework for machine learning and data mining, intended for both experienced users and researchers in machine



learning who want to develop and test their own algorithms while reusing as much of the code as possible, and for those just entering who can enjoy in powerful while easy-to-use visual programming environment.

Components, Scripting and Visual Programming

Orange provides a versatile environment for developers, researchers and data mining practitioners.

Thanks to Python, a new generation scripting language and programming environment, your data mining scripts may be simple but powerful. To further allow for fast prototyping, Orange employs a component-based approach: you can implement your analysis method just like putting together the LEGO bricks, or even use an existing al-

gorithm and replace some of its standard components with your own ones. What are Orange components to scripting are Orange widgets to visual programming. Widgets employ a specially designed communication mechanism for passing objects like data sets, attribute lists, learners, classifiers, and alike, allowing to easily build rather complex data mining schemes that use state-of-the-art approaches and techniques.

The guiding principle in Orange is not to cover just about any method and aspect in machine learning and data mining (although through years of development quite a few have been build up), but to cover those that are implemented deeply and thoroughly, building them from reusable components that expert users can change or replace with the newly prototyped ones.

Recent papers on Orange and its Components

Demsar J, Zupan B (2004) Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.

Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, Shaulsky G, Zupan B (2004) Microarray Data Mining with Visual Programming, Bioinformatics, in press (bioinformatics.oupjournals.org).

Leban G, Bratko I, Petrovic U, Curk T, Zupan B (2004) VizRank: Finding Informative Data Projections in Functional Genomics by Machine Learning, Bioinformatics, in press (bioinformatics.oupjournals.org).

Demsar J, Zupan B, Leban G, Curk T (2004) Orange: From Experimental Machine Learning to Interactive Data Mining, In Proc. ECML/PKDD.

Visit our web site:

www.ailab.si/orange



Origins of Orange

Quite a few years ago, we were (each!) writing our own code for attribute scoring, decision tree induction, ten-fold cross validation and alike (not to mention routines for loading the data and pretty-printing). We got bored. Knowing that coding of the basic set of tools from the ground up was within the job description of just about any researcher in machine learning did not help. At the time quite a few machine learning programs like C4.5 and CN2 were available, but they were coded separately, used different data

file formats, and were incompatible in every other respect. There were very few machine learning suites available, which did not offer much in terms of easy prototyping and flexibility in experimenting.

Then, thanks to Donald Michie, in 1997 came a meeting called WebLab. Taking place at a romantic site (lake Bled), it called for at a time rather rule-breaking initiative to build a flexible experimental benchmark where one could easily add his own algorithms, record the ex-

periments through scripts, and do all sorts of data analysis and machine learning. The benchmark would support both scripting and graphical user's interface. WebLab meeting generated a number of good ideas, but never took on a project it was aiming for. Nevertheless, though, it inspired us, and in that year we have started to work on Orange, a machine learning and data mining suite that had occupied us ever since.