PROCEEDINGS OF SCIENCE

TMVA - Toolkit for Multivariate Data Analysis in ROOT

Jan Therhaag[*†]
Univ. Bonn, Physikalisches Institut
E-mail: jan.therhaag@cern.ch

Given the ever-increasing complexity of modern HEP data analysis, multivariate analysis techniques have proven an indispensable tool in extracting the most valuable information from the data. TMVA, the Toolkit for Multivariate Data Analysis, provides a large variety of advanced multivariate analysis techniques for both signal/background classification and regression problems. In TMVA, all methods are embedded in a user-friendly framework capable of handling the pre-processing of the data as well as the evaluation of the results, thus allowing for a simple use of even the most sophisticated multivariate techniques. Convenient assessment and comparison of different analysis techniques enable the user to choose the most efficient approach for any particular data analysis task. TMVA is an integral part of the ROOT [2] data analysis framework and is widely-used in the LHC experiments.

35th International Conference of High Energy Physics
July 22-28, 2010
Paris, France

[*]Speaker.
[†]for the TMVA core developer team: A. Hoecker, P. Speckmayer, J. Therhaag, J. Stelzer, E. v. Toerne, H. Voss

PoS(ICHEP 2010)510

1. Data analysis with TMVA

A TMVA analysis consists of two phases: Training the multivariate methods through supervised learning, and application of the most performing methods to the classification or regression problem in question. An overview of the typical code flow for these two phases is sketched in Fig. 1.

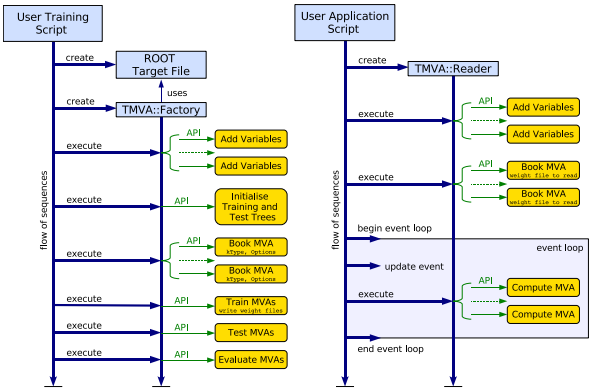

Figure 1: Left: A typical TMVA training sequence. The `Factory` organises the user's interaction with the TMVA modules. After registering the discriminating variables the selected MVA methods are booked and configured. The analysis proceeds by consecutively calling the training, testing and performance evaluation methods of the `Factory`. The training results are then written to custom weight files and the evaluation histograms are written to a ROOT file.
Right: A typical TMVA application sequence. The selected MVA methods are now used to classify data of unknown signal and background composition or to predict a regression target. A `Reader` class object serves as the interface to the methods' responses. The selected MVA methods are booked and fully configured through the weight files produced during the training phase.

Every TMVA analysis begins with the instantiation of a `Factory` object, which then steers the training, testing and evaluation. This transparant factory mode allows for an unbiased comparison of different MVA methods and guarantees that all methods see the same training and testing data with the same preprocessing applied. The `Factory` interface is highly flexible - precuts can be applied to the input data and individual event weights are as well supported as overall weights for entire trees or files. TMVA offers a great variety of MVA methods, among them Boosted Decision Trees (BDT) and different Neural Network implementations. Please consult the official TMVA Users Guide [1] for a detailed description. To guide the user in the method selection process, TMVA computes a variety of benchmark quantities for each method which are either directly printed to screen during evaluation or can be conveniently accessed via a graphical user interface.
In the application phase, the most performing MVA methods are used to classify events in data samples of unknown composition or to predict the value of a regression target. All interactions with the methods are now interfaced by the `Reader`: It manages datasets and methods, takes care of all necessary variable transformations and steers the processing of the data. All information required to configure the MVA methods is automatically retrieved from the weight files produced during the training.

In conclusion, TMVA unifies highly customizable multivariate methods for both classification and regression and convenient evaluation in a single framework. The user interface comprised of the `Factory` and the `Reader` places emphasis on clarity and functionality and will hardly exceed a few lines of code in most applications.

References

[1] A. Hoecker et al. *TMVA - Toolkit for Multivariate Data Analysis*, arXiv:physics/0703039

[2] R. Brun and F. Rademakers *ROOT - an object oriented data analysis framework*, Nucl. Inst. Meth. in Phys. Res., A 389, 81, 1997