The Search for the Higgs Boson at the ATLAS Experiment using Multivariate Techniques

J. Ebke, J. Elmsheuser, T. Langer, B. Ruckert, M.P. Sanders, and D. Schaile

One of the most fundamental questions addressed by particle physicists today is the origin of the mass of fundamental particles. The most popular theory holds that particles acquire mass by interacting with a scalar field which permeates space - the Higgs field. This field can itself be excited, giving rise to a massive, unstable scalar boson, which can be produced in particle collisions at the LHC.

If the mass of the Higgs particle is in the range of 135 to 200 GeV, it will predominantly decay to W boson pairs. In the analysis [1] presented here, we observe the W particles from their decay into muons and neutrinos. We have studied simulated data from the ATLAS detector at a centre of mass energy of 14 TeV using a full Geant4 simulation and in addition events from a fast simulation using Atlfast II. Most of the simulation and analysis has been performed on the LHC Community Grid using the distributed analysis tool Ganga.

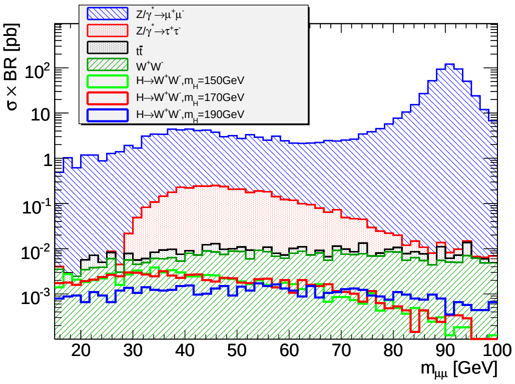

Fig. 1: Distribution of the invariant mass of the two leading muons after preselection cuts.

In Figure 1, the relative magnitude of signal and background shows that the separation of signal and background can be very difficult. As a baseline for this separation we use a traditional cut based analysis, which already yields a good signal significance. We then compare several multivariate methods to this baseline analysis, using the TMVA package [2]: neural networks, boosted decision trees, bagged randomised trees and the Fisher classifier have been studied.

Firstly, on a training dataset, the parameters for the multivariate methods are determined using training algorithms. Secondly, the efficiency of the selection is determined on a testing dataset - this datasets must be statistically independent of the training dataset, or the effect of "overtraining" can fake extremely optimistic results. For example, one can use a simple decision tree to exactly classify all events in a datasets to "signal" and "background" - the performance on independent data is the completely unknown.

We also applied this separation of training and testing datasets to the optimisation of cut positions. Even though the effect was less pronounced than for the multivariate methods, we observed some overtraining after optimising the cuts, leading to an initial overestimation of the analysis performance.

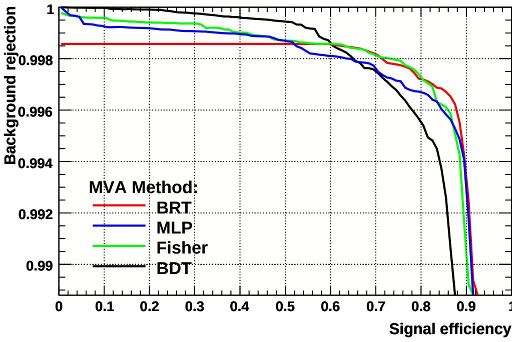

Fig. 2: ROC curves for the four examined classifiers for an Higgs mass of 170 GeV

As an example for the performance of the methods, the signal efficiency is plotted against the background efficiency for the four examined methods in figure 2. This so-called "ROC Curve" can be used to determine the best working point for each method.

Advanced methods to describe the signal region - in contrast to a cut analysis - are most useful in situations where the actual signal region is far from a hypercube. If the cut analysis is close to optimal, the benefits of using multivariate methods are small, and systematic uncertainties begin to dominate. However, if backgrounds with more complex decay topologies are important, the strengths of multivariate analysis show - boosted decision trees, for example, can improve upon cut analysis by a factor of two in significance.

However, the gains from multivariate analysis must always be balanced with the systematic uncertainties introduced. We have shown that the use of multiple variables can increase the systematic shifts by a large factor. If the systematic uncertainties are carefully constrained by using data, multivariate analysis can however increase the sensitivity of an analysis significantly.

In our analysis the boosted decision tree method gave the largest improvement if no systematic uncertainties were considered. However, it takes a long time to train and is also relatively slow in application. In contrast, while the Fisher classifier performs worse it has an extremely low training and application time, and can be useful in situations where speed is essential. Finally, in the presence of strong systematic uncertainties the randomized decision tree method showed best performance and robustness.

References

[1] J. Ebke, The search for the Higgs Boson at the ATLAS Experiment using Multivariate Techniques, LMU München, November 2008

[2] A. Hocker *et al.*, TMVA - Toolkit for Multivariate Data Analysis, *CERN-OPEN-2007-007*, 11 Jun 2007.