PROCEEDINGS OF SCIENCE

Tau identification using multivariate techniques in ATLAS

Marcin Wolter[*†]
On behalf of the ATLAS Collaboration
Institute of Nuclear Physics PAN, Cracow, Poland
E-mail: Marcin.Wolter@ifj.edu.pl

Tau leptons will play an important role in the physics program at the LHC. They will not only be used in electroweak measurements and in detector related studies like the determination of the $E_{\mathrm{T}}^{\mathrm{miss}}$ scale, but also in searches for new phenomena like the Higgs boson or Supersymmetry. Due to the overwhelming background from QCD processes, highly efficient algorithms are essential to identify hadronically decaying tau leptons. This can be achieved using modern multivariate techniques which make optimal use of all the information available. They are particularly useful in case the discriminating variables are not independent and no single variable provides good signal and background separation. In ATLAS four algorithms based on multivariate techniques have been applied to identify hadronically decaying tau leptons: Projective Likelihood Estimator (LL), Probability Density Estimator with Range Searches (PDE-RS), Neural Network (NN) and Boosted Decision Trees (BDT). All four multivariate methods applied to the ATLAS simulated data have similar performance, which is significantly better than the baseline cut analysis.

XII Advanced Computing and Analysis Techniques in Physics Research
November 3-7 2008
Erice, Italy

[*]Speaker.
[†]Supported in part by the Polish Government grant N202 006434 (years 2008-2010).

PoS(ACAT08)080

1. Introduction

Tau leptons play an important role in the physics to be observed at the ATLAS experiment [1]. They enter in electroweak measurements, studies of the top quark, and as a signature in searches for new phenomena such as Higgs bosons, Supersymmetry and Extra Dimensions. However, tau reconstruction and identification is not an easy task. The QCD multi jet events dominating the backgrounds have a much larger cross-section [2]. Therefore, efficient selection using multivariate analysis techniques is needed.

Tau leptons decay to hadrons in 64.8% of the cases and to electron or muon the rest of the time. In about 77% of hadronic tau decays only one charged track is produced: $\tau \to \nu_\tau + \pi^\pm + n\pi^0$ and in about 23% there are 3 charged tracks: $\tau \to \nu_\tau + 3\pi^\pm + n\pi^0$ [3]. The tau candidates with a single charged track are called 1-prong, with three tracks 3-prong.

Reconstruction of tau candidates

Hadronically decaying tau candidates are reconstructed using at least one of two possible seed types. The first seed type (“track-seeded" tau candidate) is a track with $p_{\mathrm{T}} > 6$ GeV, which satisfies further quality criteria. The second type of seed (“calo-seeded" tau candidate) consists of jets reconstructed using calorimeter clusters with $E_{\mathrm{T}} > 10$ GeV. These clusters are then grouped into a jet using a seeded cone algorithm [4]. If a match between such a jet and tracks is found, the tau candidate is considered as having two valid seeds. For reconstructed tau leptons in $Z \to \tau\tau$ events, 70% of all tau candidates have two valid seeds, 25% are only “calo-seeded", and 5% only “track-seeded".

Discriminating variables

Discriminating variables used to distinguish tau leptons from QCD jets include: radius and profile of EM calorimeter energy deposits, spread of the associated tracks, isolation variables calculated from calorimetric energy deposits and tracks, impact parameter significance of the leading track, invariant mass of the associated tracks, ratios of energy deposits to the sum of track transverse momenta, and the transverse flight path significance of the candidate vertex.

Eight such discriminating variables are used for tau identification in the case of the “calo-seeded" candidates. For “track-seeded" candidates nine discriminating variables for 1-prong and eleven variables for 3-prong candidates are used. Distributions of four selected variables are shown in Fig. 1.

These variables are not independent and no single variable provides a sufficient signal and background separation.

2. Identification algorithms

Various tau identification algorithms are implemented in the TauDiscriminant package, which is a part of the Atlas reconstruction software. The methods used are: cut analysis, Projective Likelihood Estimator (LL), Probability Density Estimator with Range Searches (PDE-RS), Neural Network (NN) and Boosted Decision Trees (BDT). All algorithms are implemented in the TauDiscriminant package, which is a part of the ATLAS reconstruction software.

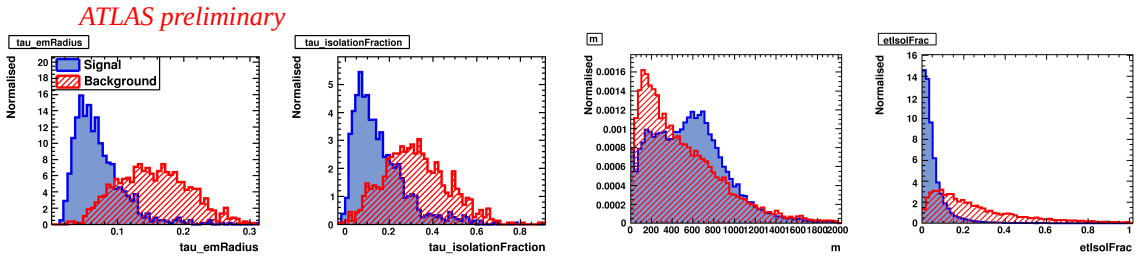

Figure 1: Selected discriminating variables for tau candidates (dark-shaded - signal, hatched - background). From left to right: *emRadius* - radius of the cluster in the EM calorimeter, *isolationFraction* - isolation fraction of transverse energy between $0.1 < \Delta R < 0.2$ around the cluster barycenter (both for “calo-seeded" candidates), *m* - invariant mass, *etisolFrac* - ratio of transverse energy in $0.2 < \Delta R < 0.4$ to total transverse energy (both for “track-seeded" 1-prong candidates).

2.1 Cut analysis

Human optimized cuts are the baseline identification algorithm for “track-seed" candidates. The cuts are optimized separately for 1-prong and 3-prong candidates. Figure 2 shows the reconstruction and identification efficiencies for true and fake candidates as a function of their transverse energy using the base-line cut selection.

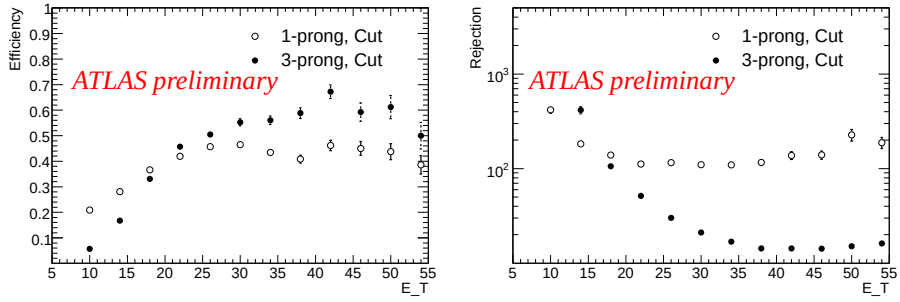

Figure 2: Reconstruction and identification efficiencies with respect to true candidates from signal (left) and rejection for background candidates (right) with human optimized cut selection.

2.2 Projected Likelihood

The likelihood discriminant is constructed as $d = \mathscr{L}_S/(\mathscr{L}_B + \mathscr{L}_S)$, where $\mathscr{L}_S$ and $\mathscr{L}_B$ are the likelihoods that a tau candidate is a real or a fake tau. In this approach correlations between variables are ignored.

The likelihood function uses information from both the calorimeter and the tracker. Different input variables are used for single and multi prong events. The variables from both “track-seeded” and “calo-seeded” candidates are used in the likelihood calculation. This is the base-line method for “calo-seeded" tau candidates.

The projected likelihood method has low memory and CPU consumption together with good performance. It is transparent and insensitive to small changes in the training sample.

Figure 3 (left) shows the performance of the method for three prong tau candidates reconstructed from both the calorimeter and the track seeds.

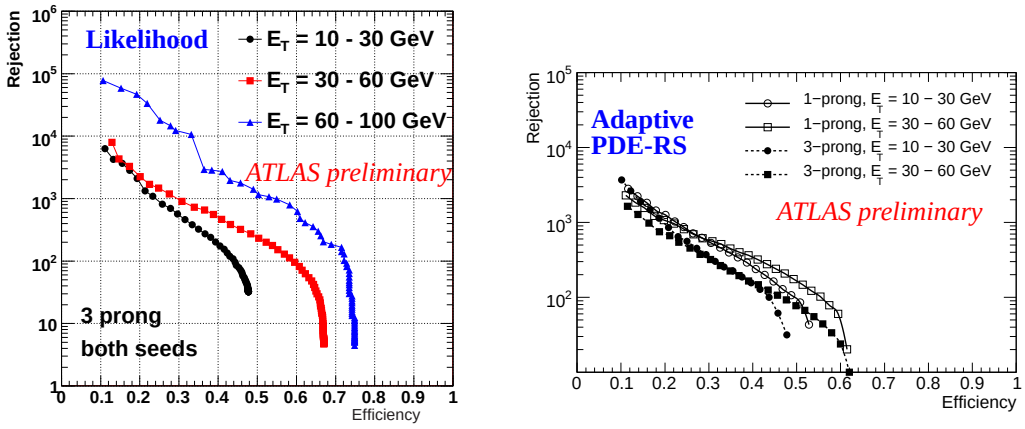

Figure 3: Rejection vs efficiency using projected likelihood (left plot) and using PDE-RS (right plot) for three prong tau candidates. Tau candidates were reconstructed from both track and calorimeter seeds.

2.3 PDE-RS algorithm

Probability Density Estimator with Range Searches (PDE-RS) [5] is based on sampling the signal and background densities in a multidimensional phase space built out of discriminating variables. Taking the number of signal events n_S and the number of background events n_B in a small volume $V(x)$ around point x in the multidimensional space, a discriminant defined as $D(x) = n_S/N_S/(n_S/N_S + n_B/N_B)$ is a good approximation of the probability that a given candidate is a signal. N_S stands for the total number of signal events and N_B for the number of background events. The event counting is done using multidimensional binary trees, which increase the speed of the algorithm. Hypercube dimensions are the only free parameters to be tuned. In the adaptive PDE-RS mode, the volume size is adapted automatically such that the number of events from the training sample enclosed in it is similar for all volumes. Use of adaptive PDE-RS increases background rejection by about 5% compared to standard PDE-RS.

The PDE-RS method has good performance and it takes correlations between variables into account. The disadvantage is the high memory and CPU consumption and relatively big training samples needed for an optimized analysis. The results for the adaptive method are shown in Fig. 3 (right).

2.4 Neural Network

Neural network is a non-linear discriminating method [6]. For tau identification the Stuttgart Neural Network Simulator [7] is used.

In the feedforward network, as used for the tau identification, the information propagates from input to output without any loops. The architecture of the network is optimized to give the proper classification of signal and background and to avoid over-training at the same time. The trained

network is exported to C code which is then included in the ATLAS software. The resulting code is fast and has low memory consumption.

The neural network identification is implemented for “track-seeded" tau candidates. Eight separate networks are trained separately for 1,2,3-prong candidates with and without an additional π^o cluster. For 1-prong candidates separate networks are used depending on the impact parameter significance availability, since including it improves the selection. The rejection of the neural network identification is shown in Fig. 4 (left) as a function of the efficiency.

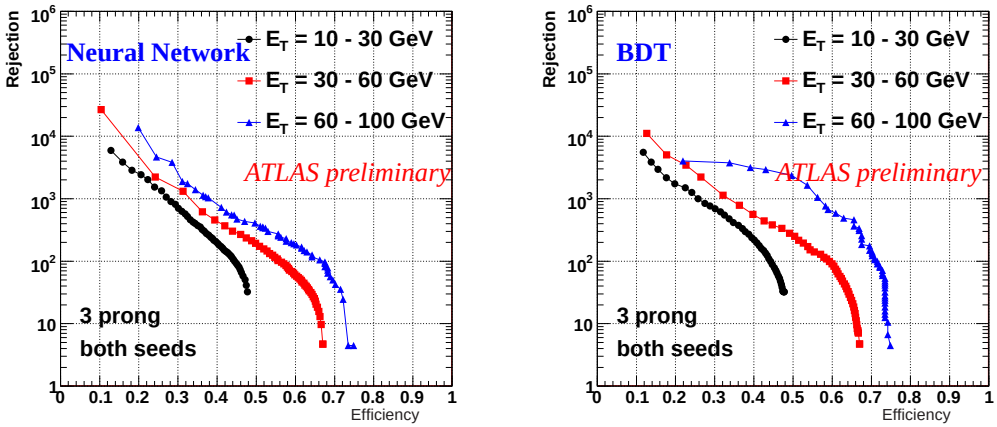

Figure 4: Rejection as a function of the efficiency for neural network (left) and for BDT algorithm (right) for 3-prong tau candidates reconstructed with both seeds.

2.5 Boosted Decision Tree

Well-optimized multivariate algorithms converge to similar signal/background separation, since they all approximate the Bayes discriminant function. An important difference is how easy, fast and robust the optimization is. BDTs have several attractive features in this regard: they are fast to train, they take correlations between variables into account, they can use discrete variables directly, adding well-modelled variables will not degrade performance and the number of tunable parameters is quite small.

A decision tree is a variation of a simple cut-based classifier in which objects failing cuts are not discarded, but are instead subject to further analysis. In this way, a cut-based procedure can be transformed into a multivariate technique with a quasi continuous discriminant output.

A tree is built by training on signal and background samples. To build a tree, the node must be split into a pair of “child” nodes according to some criteria. The algorithm achieves this split by scanning all input variables to find the cut-value which maximizes the decrease in node impurity.

Boosting is a general technique for improving the performance of any weak classifier. It involves a weighted average over many decision trees, which stabilizes the result and improves performance. The boosting algorithm increases the weight of events misclassified by the first tree and repeats training. In effect, this causes the second tree to change its optimization to better classify such events for which the first tree was weak. This procedure continues through a user-

chosen number of trees. In the end, the full set of trees is combined to obtain a final discriminant. The implementation described employs the AdaBoost method [8].

Boosted Decision Tree has been found to perform well in separating taus from jets and electrons. Fig. 4 (right) summarizes the performance of the BDT algorithm.

3. Summary and outlook

All the methods presented are performing well: while the cut analysis is robust, transparent for users and not CPU demanding the use of multivariate techniques leads to performance improvement. Projected Likelihood is a well performing, fast tool which is already popular in HEP. PDE-RS is efficient, but CPU demanding and large samples of reference candidates are needed. Neural Network provides fast classification while converted to the C function after training and BDT besides good performance offers also simple training with not many parameters to be tuned.

Experience shows, that multivariate analysis is necessary, if it is important to extract as much information from the data as possible. However, for classification problems no single “best" method exists. What becomes important is also simplicity of training and fast, robust classification.

We have implemented multivariate algorithms optimized on Monte Carlo samples. Next task is to prepare for real ATLAS data. This requires finding an optimal set of variables by variable ranking and possible reduction. The optimization should be focused on robustness and flexibility. While real data become available the important part is a comparison of Monte Carlo with them, taking into account also correlations between variables. Other important issue is the estimation of systematic uncertainties.

References

[1] **ATLAS** Collaboration, “The ATLAS Experiment at the CERN Large Hadron Collider,” *JINST 3* **S08003** (2008).

[2] **ATLAS** Collaboration, “Expected Performance of the ATLAS Experiment, Detector, Trigger and Physics.” CERN-OPEN-2008-020 (2008), to appear.

[3] **Particle Data Group** Collaboration, C. Amsler, *et al.*, “Review of particle physics,” *Phys. Lett.* **B667** (2008) 1. doi:10.1016/j.physletb.2008.07.018.

[4] S. D. Ellis, *et al.*, “Jets in Hadron-Hadron Collisions,” *Prog. Part. Nucl. Phys.* **60** (2008) 484–551.

[5] T. Carli, B. Koblitz, “A multi-variate discrimination technique based on range-searching.” Nucl. Instrum. Meth. A, (501):576, 2003.

[6] C. M. Bishop, “Neural networks for pattern recognition”. Oxford University Press, Oxford, UK, 1996.

[7] http://www.ra.cs.uni-tuebingen.de/SNNS/.

[8] Y. Freund, R. E. Schapire, “Experiments with a New Boosting Algorithm.” In Proceedings of the 13th International Conference on Machine Learning, pages 146–148, 1996.