Ideal tau tagging with the multivariate data-analysis toolkit TMVA

Heikkinen, Mika Aatos

2010


http://hdl.handle.net/10138/18736
http://dx.doi.org/10.1088/1742-6596/219/3/032010

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.
Ideal tagging with the multivariate data-analysis toolkit TMVA

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2010 J. Phys.: Conf. Ser. 219 032010


View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 128.214.56.242
The article was downloaded on 04/10/2010 at 16:35

Please note that terms and conditions apply.
Ideal $\tau$ tagging with the multivariate data-analysis toolkit TMVA

A Heikkinen, P Kaitaniemi, V Karimäki, M J Kortelainen, T Lampén, S Lehtti, T Lindén and L Wendland
Helsinki Institute of Physics, P.O. Box 64, FIN-00014 University of Helsinki, Finland
E-mail: aatos.heikkinen@cern.ch

Abstract. The experience on using ROOT package TMVA for multivariate data analysis is reported for a problem of $\tau$ tagging in the framework of heavy charged MSSM Higgs boson searches at the LHC. We investigate with a generator level analysis how the $\tau$ tagging could be performed in an ideal case, and hadronic $\tau$ decays separated from the hadronic jets of QCD multi-jet background present in LHC experiments. A successful separation of the Higgs signal from the background requires a rejection factor of $10^5$ or better against the QCD background. The $\tau$ tagging efficiency and background rejection are studied with various MVA classifiers.

1. Introduction
Multivariate analysis methods have become increasingly important in high energy physics. The rare and subtle signals are hidden within voluminous data, and their analysis can benefit from multivariate algorithms, since taking full correlations into account can greatly increase the ability to separate signal from background [1].

The ROOT [2] package TMVA [3] for multivariate data analysis (MVA) was demonstrated to be applicable to b-tagging in Ref. [4]. In this study, the TMVA package is applied to $\tau$ identification in the heavy charged MSSM Higgs boson decay $H^\pm \rightarrow \tau^\pm \nu_\tau \rightarrow$ hadrons. In previous studies conducted in the CMS experiment, this channel has been found to provide an interesting possibility to discover the charged Higgs boson [5], should it exist.

One of the main challenges of finding the heavy charged Higgs boson is, that the cross-section of the largest background, i.e. QCD multi-jet events, which could fake hadronically decaying $\tau$’s, is up to $10^7$ times greater than the signal cross-section at the 14 TeV center of mass collision energy of the LHC. Since the production of such large Monte Carlo samples is not currently feasible with full detector simulation, generator level simulation was used to obtain an estimate for the benefit of the use of multivariate methods to separate the signal and background.

It is estimated that a rejection factor of $10^5$ is needed for a $\tau$ identification in order to make the charged Higgs boson signal visible [5]. Therefore, the performance of the selected MVA classifiers was evaluated for background rejections of $10^5$ and $10^6$, and some programming was required to map the study into the TMVA framework and to analyse the Monte Carlo samples.

2. TMVA - Toolkit for Multivariate Data Analysis with ROOT
ROOT-integrated TMVA is a framework for training, testing and performance evaluation of multivariate classification techniques. TMVA works in transparent factory mode to guarantee
an unbiased performance comparison between the classifiers, such as:

- Rectangular cut optimisation
- Projective likelihood estimation (PDE approach)
- Multidimensional probability density estimation (PDE - range-search approach, PDERS)
- Multidimensional k-nearest neighbour classifier
- Function discriminant analysis (FDA)
- Predictive learning via rule ensembles (RuleFit)

Main characteristics of different TMVA classifiers are summarised in Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuts</td>
<td>Easy to understand</td>
<td>Possibly inefficient</td>
</tr>
<tr>
<td>Likelihood methods</td>
<td>Fast to train and evaluate</td>
<td>Non-linear correlations treated badly</td>
</tr>
<tr>
<td>HMatrix, Fisher</td>
<td>Very fast and transparent</td>
<td>fail if PDFs have same mean, and if non-linear correlations</td>
</tr>
<tr>
<td>PDERS, kNN</td>
<td>Handles well complex class boundaries</td>
<td>Impractical with more than 10 variables</td>
</tr>
<tr>
<td>ANN</td>
<td>Very good with non-linear correlations</td>
<td>Black box, needs tuning</td>
</tr>
<tr>
<td>BDT</td>
<td>Very good out-of-the-box performance</td>
<td>Needs tuning to avoid overtraining</td>
</tr>
<tr>
<td>RuleFit</td>
<td>Like BDT but simpler, fast evaluation</td>
<td>Often needs some tuning</td>
</tr>
<tr>
<td>SVM</td>
<td>Good with non-linear problems, insensitive to overtraining</td>
<td>Not transparent</td>
</tr>
<tr>
<td>FDA</td>
<td>Very good classification if boundary is known</td>
<td>Classification boundary function needed</td>
</tr>
</tbody>
</table>

Of the available classifiers, the following MVA methods were selected for evaluating the TMVA performance:

- Linear discriminant analysis (LDA) based on Fisher discriminants
- Boosted/Bagged decision trees (BDT)
- Support Vector Machine (SVM)
- Artificial neural networks (ANN)

3. Data

The signal was generated with Pythia [7] version 6.4.19 through the process $gg \rightarrow tbH^\pm$, $H^\pm \rightarrow \tau^\pm \nu_{\tau}$ in the maximal $m_H$ SUSY scenario [8] with $m_{H^\pm} = 217 \text{ GeV}/c^2$ and $\tan \beta = 30$.

The $\tau$ leptons were forced to decay hadronically. The decay of the $\tau$ leptons was simulated with the Tauola program [9] version 2.6 to obtain correct polarization for the $\tau$ lepton and its decay products [10]. A total of $10^5$ and $2 \times 10^5$ signal events were produced for training and evaluation of the multivariate analyzers, respectively.

The dominating background for this physics channel is the QCD multi-jet background. This background was generated with Pythia version 8.108. The transverse momentum of the hadronic jets was limited to the bin $120 < p_T < 170 \text{ GeV}/c$, which has been found to be the most difficult $p_T$ range [5]. Training and evaluation samples of $5 \times 10^6$ and $10^8$ QCD multi-jet events were produced, respectively. Another independent sample of $5 \times 10^6$ events was used as a second training sample to estimate the bias caused by the training.

The events were generated with $p$-$p$ collisions at a center of mass energy of 14 TeV. The jets were reconstructed with the PYCELL-method in a cone of 0.5. For the signal, only the jets
corresponding to the $\tau$ decay, i.e. $\tau$ jets, were taken as $\tau$-jet candidates. For the background events, all jets obtained with PYCELL-methods were taken as $\tau$-jet candidates.

In order to save CPU time and disk space, a set of preselection cuts was applied to the $\tau$-jet candidates. These are standard cuts, which are used for $\tau$ identification [11]. Care was taken, that the preselection cuts were loose enough in order not to bias the MVA performance. The following preselections were used:

- jet $E_\mathrm{T} > 100$ GeV
- jet $|\eta| < 2.2$
- matching of the leading track, i.e. the track with the highest $p_T$, to the jet direction within a cone of 0.1
- cut on the $p_T$ of the leading track, $p_T > 20$ GeV/$c$
- charged track isolation, where at most one track was allowed in the isolation annulus between the cones 0.04 and 0.50 around the leading track direction; tracks, which fulfilled the following criteria were counted:
  - $|\eta|$ of tracks, $|\eta| < 2.5$
  - minimum $p_T$ of charged tracks, $p_T > p_{\text{min}} = 0.5$ GeV/$c$
  - track matching to primary vertex along the beam axis, $|P_{\text{track}}^z - z| < 2$ mm

One or two tracks were allowed in a signal cone size of 0.04 around the leading track.

If at least one of the jets in the generated event fulfilled these conditions, all the jets in the event were saved. For training and evaluation of the MVA methods, the jets were required to fulfill $E_\mathrm{T} > 100$ GeV, $|\eta| < 2.4$. Only the leading track was allowed to be within the signal cone in order to select the one-prong final state of $\tau$ decays, which dominate the hadronic $\tau$ decay final states. In the signal samples, the jet was required to be matched to the $\tau$ jet coming from the $H^\pm \rightarrow \tau^\pm \nu_\tau$ decay. The total preselection efficiencies were found to be 17.3 % and 0.7 % for the signal and background samples used for the evaluation of the MVA methods, respectively.

Most of the generally used variables for the $\tau$ identification in the $H^\pm \rightarrow \tau^\pm \nu_\tau$ decay [11] were used as input to the MVA methods. These variables include a cut on the transverse energy and pseudo-rapidity of the $\tau$-jet candidate, maximum track $p_T$ in the isolation annulus between cones of 0.04 and 0.50 around the leading track direction to impose charged track isolation, the electromagnetic energy sum in the region between cones of 0.10 and 0.50 around the jet axis, and matching of the hadronic energy deposition to the leading track momentum to reject electrons. Furthermore, the $R_\tau = p_{\text{track}}/E_\tau$ variable, where $E_\tau$ is the reconstructed $\tau$-jet candidate energy (excluding neutrinos), was used to take advantage of the boost of the $\tau$ due to polarization [10]. The variables are summarized in Table 2. Figure 1 shows the distributions of the jet $E_\mathrm{T}$ and $R_\tau$ variables, which were found to have the best separation power. The input variables were used without transformations as well as with decorrelation or principal component analysis applied.

4. $\tau$ tagging analysis code for TMVA

For $\tau$-tagging it is natural to train and use the MVA methods with jets. However, in order to obtain results which can be compared with other studies, the evaluation of the methods should be done with respect to events. An analysis program was therefore developed to train and evaluate the methods with TMVA in a standard way, and in addition to re-evaluate the methods with events. This analysis code was made for TMVA distributed with ROOT 5.22.

The event based evaluation algorithm is described in Appendix A Listing 1. The algorithm is very similar to the evaluation algorithm in TMVA with the exception of bookkeeping of events. The preselection efficiencies are taken into account, and the signal efficiencies are printed at background efficiency levels $10^{-5}$ and $10^{-6}$. Example Listing 2 in Appendix A demonstrates the output from the analysis program.
Table 2. Variables used in the analysis.

<table>
<thead>
<tr>
<th>ID</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Jet $E_T$</td>
</tr>
<tr>
<td>1</td>
<td>Jet $\eta$</td>
</tr>
<tr>
<td>2</td>
<td>Charged track isolation: no tracks with $p_T &lt; p_T^{\text{max}}$ in isolation annulus of 0.04-0.50</td>
</tr>
<tr>
<td>3</td>
<td>Isolation of electromagnetic energy ($\Delta R=0.10 - 0.50$)</td>
</tr>
<tr>
<td>4</td>
<td>Neutral hadron rejection</td>
</tr>
<tr>
<td></td>
<td>(i.e. track $p$ matching to hadronic energy deposition)</td>
</tr>
<tr>
<td>5</td>
<td>$R_\tau = p(\text{leading track}) / E(\text{jet})$</td>
</tr>
</tbody>
</table>

Figure 1. Example of data used in $\tau$ tagging. Distributions of jet $E_T$ (left) and $R_\tau$ (right) variables are shown after all preselections.

5. Results
In the following the usage and results of selected TMVA classifiers are presented. The amount of data used for training varied depending on the classifier, but for testing all available data was used. The systematic uncertainty was estimated by repeating the full analysis with independent background data.

5.1. Classifying with Fisher discriminant
The method of Fisher discriminants is a computationally easy method, which determines the discriminating function analytically in the multivariate space represented by the input variables. It has been used in analyses of several HEP experiments, for instance in BaBar [12] and Belle [13].

The Fisher method works in a transformed variable space with zero linear correlations, by distinguishing the mean values of the signal and background distributions. An axis in the (correlated) hyperspace of the input variables is chosen so that when projecting the output classes (signal and background) upon this axis, they are pushed as far as possible away from each other compared to the average mutual distance of events belonging to the same class [14].

Fisher discriminants are optimal for Gaussian distributed variables with linear correlations. However, when a variable has the same sample mean for signal and background, no
discrimination is achieved. If the shapes of the distributions are different, a suitable transformation can then be applied [14].

The Fisher classifier was used as an example of linear discriminant analysis in the case of ideal $\tau$ tagging. The following settings were used:

```
Fisher H:!V:!Normalise:CreateMVAPdfs:Fisher:NbinsMVAPdf=50:NsmoothMVAPdf=1
```

The response of TMVA to Fisher classifier is shown in Fig. 2. The signal efficiency was found to be $3.5\pm0.1\%$ at a background rejection of $10^5$. Signal efficiencies are also shown in Table 3 together with other discrimination methods tested.

![Figure 2. TMVA response to Fisher discriminant (left) and the output of the SVM classifier with Gaussian kernel (right).](image)

5.2. Boosted Decision Trees

Recently, the Boosted Decision Trees (BDTs) have been advocated as an alternative to artificial neural networks for particle identification [15].

The BDT method is based on binary decision trees visualized in Fig. 3. Repeated yes/no decisions are made on the variable with the best separation power until the subsamples become small, or until the subsamples are declared as signal or background. The variable phase-space is hence divided into a large number of hypercubes, which is why BDT is effective for both with linear and non-linear samples.

In order to make the decision trees robust against statistical fluctuations of the training sample, boosting, i.e. reweighting, is applied to the training sample. After each reweighting, a new decision tree is constructed. The boosting thus combines iteratively many weak classifiers or hypotheses into a single stronger rule called the combined hypothesis [15]. The outcome of a tested event is carried out by evaluating the decisions of typically hundreds of trees. Overtraining is countered by pruning nodes with insignificant separation.

The BDT was evaluated with the following setup:

```
BDT V:NTrees=400:BoostType=AdaBoost:SeparationType=GiniIndex:nEventsMin=20:
nCuts=20:PruneMethod=CostComplexity:PruneStrength=4.0
```

Increasing the number of trees was found not to yield significant improvement. Different pruning strengths were tried out in order to determine a level at which overtraining is tolerable. The signal efficiency was found to be $7.3\pm1.3\%$ at the background rejection of $10^5$. Decorrelation and principal component analysis were tried out for the input variables, but they were found
to yield only minimal changes in the signal efficiency. The full training samples were used to obtain the results.

![A visualization of the yes/no chain of decisions of a boosted decision tree. Some of the nodes have been declared as signal or background.](image)

**Figure 3.** A visualization of the yes/no chain of decisions of a boosted decision tree. Some of the nodes have been declared as signal or background.

### 5.3. Support Vector Machine

The Support Vector Machine (SVM) learning algorithm is a recent addition to MVA methods. One of the first applications in HEP was the classification problem of signal/background discrimination in the $t\bar{t}$ dilepton channel [16].

The SVM maps the input vectors into the feature space through some non-linear mapping. In this space an optimal hyperplane is constructed and evaluated by a kernel function.

Potential advantages of the SVM method compared to the ANN method include the existence of only few user chosen parameters, ability to find global minimum, and correspondence to a linear method which makes the SVM theoretically easy to analyse.

For $\tau$ tagging the SVM was trained with a Gaussian kernel with $4 \times 10^3$ signal jets and $3.2 \times 10^4$ background jets. The SVM training time scales as $O(n^2)$, where $n$ is the size of the training sample. Therefore the training sample size was kept small for this method. The Gaussian kernel has two parameters ($\text{Sigma}$, $C$) and the optimisation was done in this parameter space with a grid scan. The training for individual points were run in parallel on a Linux cluster. The best signal efficiency was found to be $5.6 \pm 0.1\%$ at the background rejection of $10^5$ with the following parameters:

```
SVM_Gauss Sigma=0.5:C=17:Tol=0.001:MaxIter=20000:Kernel=Gauss
```

Figures 2 and 4 demonstrate the TMVA output of the SVM classifier with the Gaussian kernel.

### 5.4. Neural Networks

An interesting study related to our $\tau$ tagging is presented in [18], where an artificial neural network (ANN) was trained to choose the polarity of $\tau$ particles from the decay angles. It was shown that the $\tau$ helicity found by the ANN approximated well the optimal Bayesian classifier.

Of the three Multilayer Perceptrons (MLP) implementations, supported by TMVA, $\text{TMVA::Types::kMLP}$ was selected. Data for 6-15-1 MLP configuration with neurons of sigmoid type was trained for $10^3$ cycles (see Fig. 5.4) with $10^4$ signal jets and $4 \times 10^4$ background jets. A ROOT TMVA configuration for these settings can be written as follows:

```
17th International Conference on Computing in High Energy and Nuclear Physics (CHEP09) IOP Publishing
```
Figure 4. TMVA plot for parallel coordinates. In this kind of plot each line going through specific variable, as explained in Table 2, represent one event. Poorly classified background events with value 0.9–1.0 are selected from the vertical histogram on the left.

\[ \text{NSigTrain}=10000; \text{NBkgTrain}=40000; \text{SplitMode}=\text{Random}; \text{NormMode}=\text{NumEvents} \]

\[ \text{MLP}_v0 \text{ H:V:Normalise:NeuronType=sigmoid:N} \text{Cycles}=1000; \text{HiddenLayers}=N+9,N; \text{TestRate}=5; \text{VarTransform}=\text{PCA} \]

In addition to variable transformation \( \log(E_T) \), Principal Component Analysis, PCA, (see \text{VarTransform}=\text{PCA} above) was found to improve classification power. In our \( \tau \) tagging problem the PCA method simply performs a rotation in the 6-dimensional parameter space to a new coordinate system whose unit vectors are the eigenvectors of the system.

Convergence of the neural network training and for test data is shown in Fig. 5.4. The signal efficiency for the MLP discriminator was found to be 6.5\( \pm \)0.1\% at the background rejection of \( 10^{-5} \).

5.5. Summary of results

The performance of the various TMVA discriminators for the ideal \( \tau \) tagging problem is summarised in Table 3.

Overall discrimination performance of selected TMVA classifiers are demonstrated with Receiver Operating Characteristics (ROC) curves in Fig. 6. In order to take the preselection efficiencies into account, signal efficiency shown in Fig.6 should be multiplied by 0.17 and correspondingly background efficiency by 0.007. For example, the required \( 10^{-5} \) background efficiency corresponds to 0.9986 background rejection after preselections.

6. Conclusion

The usage of the TMVA package in ROOT for \( \tau \) identification in the framework of a charged Higgs boson study was discussed from the user’s point of view. It was observed, that the TMVA
Table 3. Summary of performance of various TMVA discriminators for the ideal $\tau$ tagging problem.

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>Signal efficiency (%) for background efficiency $10^{-5}$</th>
<th>$10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>$3.5 \pm 0.1$ (stat) $\pm 0.0$ (syst)</td>
<td>$1.6 \pm 0.1 \pm 0.1$</td>
</tr>
<tr>
<td>BDT</td>
<td>$7.3 \pm 1.3$ $\pm 0.1$</td>
<td>$2.6 \pm 0.8 \pm 0.1$</td>
</tr>
<tr>
<td>SVM</td>
<td>$5.6 \pm 0.1 \pm 0.1$</td>
<td>$1.7 \pm 0.1 \pm 0.1$</td>
</tr>
<tr>
<td>MLP</td>
<td>$6.5 \pm 0.1 \pm 0.2$</td>
<td>$2.2 \pm 0.1 \pm 0.2$</td>
</tr>
</tbody>
</table>

Figure 5. Evolution of training and validation errors of TMVA MLP classifier during 1000 training cycles.

Figure 6. Overall discrimination performance of selected TMVA classifiers. On the right a closeup image of background rejection vs. signal efficiency curves is shown.

The package has matured since CHEP’07 and it is now fully integrated to ROOT toolkit. It also provides an interface for adding new classifiers.

Some analysis code was prepared to evaluate the study case in the TMVA framework. The multivariate data-analysis techniques were found to be promising in $\tau$ identification. At $10^{-5}$ background efficiency, TMVA classifiers were found to yield signal efficiencies in the range 3.5–7.3 %. Several methods gave comparable results, which suggests that they are close to the
Bayesian limit of achievable ideal separation.

Areas where the study can be improved have been identified. One possibility would be to use more a fundamental set of variables, instead of those chosen in this study, such as the three-momenta components $p_x$, $p_y$, $p_z$ of the final state particles. This kind of jet analysis based on neural networks has been shown to simulate a sophisticated jet algorithm $k_\perp$ [19].

Acknowledgments

The work was funded partly by the Vilho, Yrjö and Kalle Väisälä fund of the Finnish Academy of Science and Letters.

References

[14] B. Roe et al., Boosted Decision Trees as an Alternative to Artificial Neural Networks for Particle Identification, arXiv:physics/0408124
[17] L. Garrido and V. Gaitan, Use of Neural Nets to Measure the Tau Polarization and its Bayesian Interpretation, UAB-LFAE-91-04, April, 1991 Universitat Atonoma de Barcelona

Appendix A. Analysis code for TMVA event evaluation

Analysis code for simulated $\tau$ tagging events is demonstrated in Listing 1 and corresponding example from TMVA run is shown in Listing 2.

Listing 1. Pseudocode for the event evaluation.

```cpp
TMVA::Reader reader
TTree tempTree
for tree in [signal_test_tree, background_test_tree]:
    output = {}
```
Listing 2. Example listing showing analysis code for TMVA.

MyEvaluate : Testing classifiers in order to obtain signal/background efficiencies for EVENTS
MyEvaluate : (not jets, which are the input to TMVA)

<table>
<thead>
<tr>
<th>MyEvaluate</th>
<th>Signal eff</th>
<th>Events</th>
<th>Bkg eff</th>
<th>Signal eff</th>
<th>Jets</th>
<th>Bkg eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generated</td>
<td>200000</td>
<td>89799840</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event preselection</td>
<td>47866 0.239 1470627 0.016</td>
<td>206079 4345182</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMVA preselection</td>
<td>34660 0.724 635923 0.432</td>
<td>34660 0.17 736086 0.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total preselection</td>
<td>0.173 0.007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MyEvaluate : Thus 1e−5 overall bkg event efficiency corresponds to 1.4e−03 bkg event efficiency by TMVA.
MyEvaluate : Signal and bkg event efficiencies have been scaled with the preselection.
MyEvaluate : Evaluation results ranked by best signal efficiency at 1e−5

MyEvaluate : MVA Signal event efficiency at bkg event efficiency (error):
MyEvaluate : Methods: @B=1e−6 @B=1e−5 @B=1e−4 @B=1e−3 @B=0.01
MyEvaluate : SVM_Gauss_0.5.17:0.0166(006) 0.0568(012) 0.1083(016) 0.1643(019) 0.1733(020)

MyEvaluate :