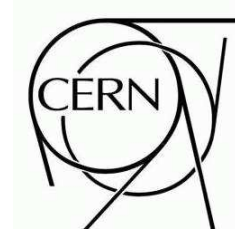


ATLAS NOTE

ATL-COM-PHYS-2007-016

December 5, 2007



Performance of boosted decision trees for combining ATLAS b-tagging methods

J. Bastos

LIP-Coimbra, University of Coimbra, P-3004-516 Coimbra

`bastos@lipc.fis.uc.pt`

Abstract

This note evaluates the performance of boosted decision trees for combining the information from different ATLAS b-tagging algorithms into a single jet classifier. The rejection of light quarks given by boosted decision trees is estimated using a Monte Carlo simulation of WH and $t\bar{t}$ events. It is shown that this approach yields significant gains in the rejection of light quarks with respect to a tagging algorithm based in 3D impact parameters and reconstructed secondary vertices.



1 Introduction

The correct identification of jets produced by b quarks is an essential requisite for precision measurements in the top quark sector and searches for the Higgs boson and physics beyond the Standard Model at the LHC. Several methods for tagging b-jets have been developed in the ATLAS experiment [1]. Additionally, the information from the most efficient taggers has been combined into joint likelihood functions in order to obtain better tagging performances than those given by individual taggers. The combination of tagging methods using a neural network approach was also contemplated [2].

Boosted decision trees is a learning technique first introduced in High Energy Physics by the Mini-BooNE experiment. It was reported [3] that particle identification with boosted decision trees has better performance than that with neural networks, in a Monte Carlo simulation of MiniBoone data. Subsequently, it has been applied to the search for radiative leptonic decays $B \rightarrow \gamma l \nu$ at BABAR [4], for detecting the first evidence of single top quark production by the D0 collaboration [5], and it also was proposed for supersymmetry searches at the LHC [6].

Boosted decision trees can be applied to any classification problem and have shown better performance than feed-forward neural networks for tagging b-jets in a toy Monte Carlo simulation of $WH \rightarrow l \nu q \bar{q}$ events [7]. This note evaluates the performance of boosted decision trees for combining the information from different ATLAS b-tagging algorithms into a single classifier. A comparison with the performance of a tagging algorithm based in 3D impact parameters and reconstructed secondary vertices is presented.

2 Boosted decision trees

The boosted decision tree algorithm consists of a set of decision trees. Initially, all jets are given the same weight $w^{(0)}$. After a tree is built, the weight of each jet is changed according to its classification given by the tree. The final classification of a jet is a weighted average of the individual classifications over all trees. A decision tree consists of a set of binary splits in the data. The algorithm begins with a parent node containing a training set of b-jet and u-jet patterns. Then, the algorithm loops over all possible binary splits in order to find the discriminating variable and corresponding separation value that optimizes a given figure of merit. This procedure is then repeated for the new daughter nodes until a stopping criterion is satisfied.

A node is called “signal node” if the sum of the weights of b-jets is greater than the sum of the weights of u-jets. Otherwise, it is called “background node”. A b-jet (u-jet) is correctly classified if it lands on a signal (background) node. If p designates the fraction of correctly classified jets in a node, its Gini index is defined to be $Q(p) = -2p(1-p)$. The optimal discriminating variable and separation value are the ones which maximize the figure of merit

$$Q_{split} = \frac{w_L Q(p_L) + w_R Q(p_R)}{w_L + w_R}, \quad (1)$$

where w_L and w_R are the sum of the jet weights in the left and right daughter nodes, respectively, and $Q(p_L)$ and $Q(p_R)$ are the Gini indices of the left and right daughter nodes. A node is not split if the optimal Q_{split} is smaller than its own $Q(p)$, or, alternatively, if it contains less events than a prespecified limit. Unsplit nodes are called “leaves”.

After the k th tree is built, the jet weights are updated. There are several methods to accomplish this. Here, we will consider the AdaBoost algorithm [8]. First, the total misclassification error ϵ_k of the tree, defined as the sum of the weights of misclassified jets over the sum of the weights of all jets, is calculated:

$$\epsilon_k = \sum_{i_{miss}} w_i^{(k)} / \sum_i w_i^{(k)}, \quad (2)$$

where i loops over all jets in the training sample. Then, the weights of misclassified jets are increased (*boosted*)

$$w_i^{(k+1)} = \frac{w_i^{(k)}}{2\varepsilon_k}, \quad (3)$$

while the weights of correctly classified jets are decreased

$$w_i^{(k+1)} = \frac{w_i^{(k)}}{2(1 - \varepsilon_k)}. \quad (4)$$

Finally, the tree $k + 1$ is constructed using the new weights $w^{(k+1)}$.

After K trees are trained their performance can be evaluated with a test sample of jets. The final classification of jet i is a weighted sum of the classifications over the individual trees

$$F_i = \sum_{k=1}^K \log\left(\frac{1 - \varepsilon_k}{\varepsilon_k}\right) f_i^{(k)}, \quad (5)$$

where $f_i^{(k)} = 1(-1)$ if the k th tree makes the jet land on a signal (background) leaf. Therefore, signal (b-jet) patterns will have large positive scores, while background (u-jet) patterns will have large negative scores. Also, trees with lower misclassification errors ε_k are given more weight when the jet classification is calculated. Further details of the boosted decision tree algorithm employed here can be found in [9].

3 Monte Carlo simulation

The analysis described in this note was performed with Monte Carlo data samples produced in the framework of the CSC production using a full simulation of the ATLAS detector with Athena 12.0.X. The detector geometry tag was ATLAS-CSC-01-02. The following physics channels were considered:

- $WH \rightarrow l\nu_l b\bar{b}$ with $m_H = 120 \text{ GeV}/c^2$ (5850 CSC sample).
- $WH \rightarrow l\nu_l u\bar{u}$ with $m_H = 120 \text{ GeV}/c^2$ (5851 CSC sample).
- $t\bar{t} \rightarrow b\bar{b}l\nu_l jj$ with $m_t = 175 \text{ GeV}/c^2$ (5205 CSC sample).

The WH samples were produced with the PYTHIA generator [12], while the $t\bar{t}$ samples were generated with AcerMC [13]. The parton distribution functions of the protons were described by CTEQ6L and both initial and final state radiation were included in the simulation. For the reconstruction of jets, the jet cone algorithm with $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} = 0.4$ was considered. Only jets with transverse momentum $p_T > 15 \text{ GeV}/c$ and pseudorapidity $|\eta| < 2.5$ were selected. The association between reconstructed jets and partons is based on the ΔR distance between the jet and the parton. A jet is matched to a parton if $\Delta R < 0.3$. Events without a reconstructed primary vertex were rejected.

4 ATLAS tagging algorithms

Tagging techniques exploit specific properties of B-hadrons to differentiate them from the large background of jets produced by light quarks and gluons. The long lifetime of B-hadrons results in displaced vertices formed by tracks from their decays. Also, tracks from B- and D-hadron decays typically have large impact parameters. In a different approach, soft-lepton tagging searches for low transverse momentum leptons inside jets, originating from semileptonic decays of B-hadrons.

The most developed approaches for tagging b-jets at ATLAS are based on the impact parameter of tracks and reconstructed secondary vertices [1]. The impact parameter based tagger, assigns to each jet a weight defined as the sum over all tracks in the jet of the log ratio of the likelihood functions for a track to belong to a b-jet and to a light jet. These likelihood functions are obtained from impact parameter distributions. B-jets will typically have large positive weights when compared with light jets. The secondary vertex based tagger [14] performs an inclusive vertex search in order to find displaced secondary vertices from B- and D- hadron decays inside the jet. After secondary vertices are reconstructed, several discriminant variables are built from their properties. These discriminant variables are taken from the SV1/SV2 vertex reconstruction algorithm. The information from taggers based in soft muons [10] and soft electrons [11] was also considered. Additionally, other variables show some discriminating power such as the largest transverse and longitudinal impact parameter significances of tracks in the jet, the largest transverse momentum of tracks in the jet and the transverse momentum of the jet. In order to obtain these discriminant variables only tracks with $p_T > 1$ GeV/c and ΔR distance relative to the jet axis smaller than 0.4 are considered. Note that, because in the node splitting procedure only the best discriminating variable is selected, boosted decision trees are insensitive to the inclusion of variable with weak discriminating power, while the training time only scales linearly with the number of input variables. The boosted decision trees were fed with the following list of input variables:

- IP2D: jet weight from transverse impact parameters.
- IP3D: jet weight from 3D impact parameters.
- SV1: jet weight from secondary vertices.
- SV2: jet weight from secondary vertices.
- SVBU: jet weight from secondary vertices.
- lhsig: combination of jet weights IP1D, IP2D and SVBU.
- weight: combination of jet weights IP3D and SV1.
- softm: soft muon based tagger.
- softe: soft electron based tagger.
- mass: mass of the particles which participate in the vertex fit.
- efrc: ratio between the total energy of particles in the vertex and the total energy of all charged particles in the jet.
- n2t: number of two-track vertices.
- ntrk: number of track in vertex.
- largest transverse impact parameter significance of tracks in the jet.
- largest longitudinal impact parameter significance of tracks in the jet.
- largest transverse momentum of tracks in the jet.
- jet transverse momentum.
- multiplicity of tracks in the jet.

5 Results

Boosted decision trees (BDT) were implemented using the StatPatternRecognition package [9]. The reader is referred to the StatPatternRecognition manual for a comprehensive description of the algorithm employed here. The boosting of the jet weights was accomplished using the AdaBoost algorithm [8]. The trees were fed with the 18 discriminant variables mentioned in the previous section.

5.1 WH events

The decision trees were trained with a training sample consisting of 15000 b-jet patterns and 50000 u-jet patterns, obtained from the $WH \rightarrow lv_1q\bar{q}$ data, with $q\bar{q} = b\bar{b}, u\bar{u}$. The best results were obtained with a minimum number of jets per leaf of about 7000 and no significant improvement in the tagging performance was observed after one hundred tree iterations. An unbiased evaluation of the boosted decision trees performance is obtained using a distinct sample of b-jet and u-jet patterns (test sample).

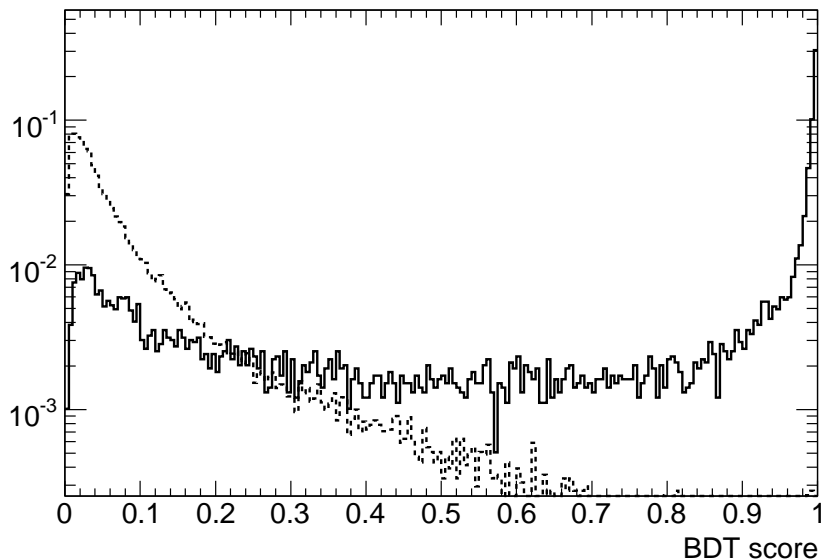


Figure 1: Normalized distributions of jet scores given by the boosted decision tree algorithm. The solid (dashed) line corresponds to b-jets (u-jets).

Figure 1 shows the distributions of jet scores, normalized to be within the interval $[0, 1]$, given by the boosted decision tree algorithm for a test sample of b-jets (solid line) and u-jets (dashed line). The distribution of b-jet scores shows a sharp peak at the value 1. On the other hand, the distribution of u-jet scores exhibits a smoother shape. Jets with a score above some specified threshold value are tagged as b-jets. The threshold value is contingent on the desired efficiency for tagging b-jets $\varepsilon_b = N_b^{tag}/N_b$, where N_b is the number of b-jets in the data and N_b^{tag} is the number of tagged b-jets, or, alternatively, on the tolerated level of contamination by light jets.

The black dots in Figure 2 show the light jet rejection $R_u = \varepsilon_u^{-1}$ as a function of the b-jet selection efficiency, given by boosted decision trees. For comparison, the rejections given by a b-tagging algorithm based in 3D impact parameters and reconstructed secondary vertices (IP3D+SV1) are depicted as gray squares. One can observe that the performance of both approaches are similar for high b-tagging efficiencies. However, for smaller efficiencies BDT outperforms IP3D+SV1. For a b-tagging efficiency of 60% the light jet rejection given by BDT is about 50% higher than that given by IP3D+SV1.

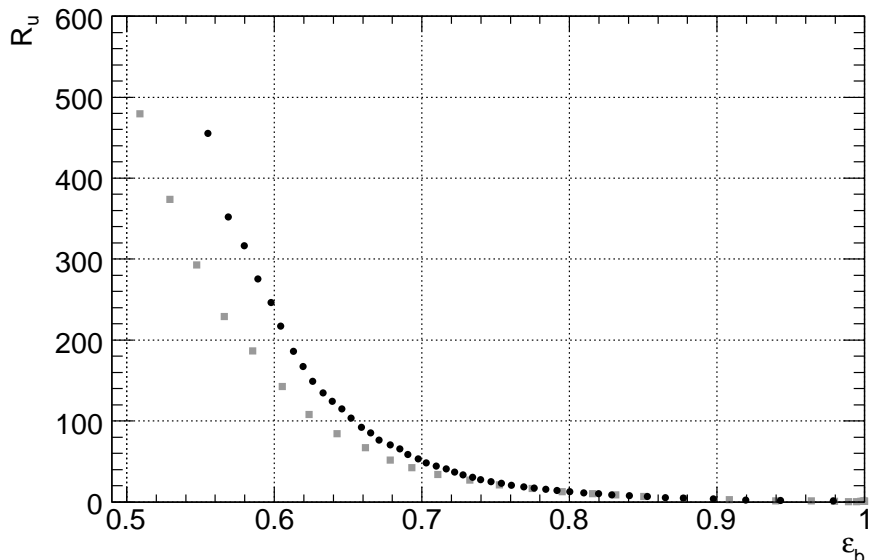


Figure 2: Light jet rejection as a function of b-jet efficiency for $WH \rightarrow lvq\bar{q}$ events, given by boosted decision trees (black dots) and IP3D+SV1 tagger (gray squares).

5.2 $t\bar{t}$ events

The trees were also trained with 50000 b-jet patterns and 50000 u-jet patterns, from $t\bar{t} \rightarrow b\bar{b}lv_ljj$ events. Again, the best performance of BDT was obtained with a minimum number of jets per leaf of about 7000 and 100 trees. Figure 3 shows the results for a test sample of jets from $t\bar{t} \rightarrow b\bar{b}lv_ljj$ events. For a b-tagging efficiency of 60% the light jet rejection given by BDT is about 45% higher than that given by IP3D+SV1.

Also of note is that the performance of both BDT and IP3D+SV1 is lower in $t\bar{t}$ events with respect to WH events. For a b-tagging efficiency of 60%, the light jet rejection of BDT (IP3D+SV1) in $t\bar{t}$ events is about 15% (7%) smaller than in WH events. On the other hand, the performance of both algorithms is expected to be better in $t\bar{t}$ events because in this channel the jets are more central than in WH events. The reduced performance of the algorithms in $t\bar{t}$ events may be attributed to the contamination of light jets by heavy flavour, which is more frequent in the high jet multiplicity environment of $t\bar{t}$ events [1]. In order to corroborate this hypothesis a “purification” cut was imposed on u-jets in $t\bar{t}$ events: only u-jets without a heavy quark within a $\Delta R < 0.8$ distance are selected. Figure 4 shows the light jet rejections given by IP3D+SV1 and BDT when purified u-jets are considered. It can be observed that, after purification and for a b-tagging efficiency of 60%, the light jet rejection in $t\bar{t}$ events given by IP3D+SV1 is increased by about 40% and the light jet rejection given by BDT is increased by about 65%. Furthermore, the light jet rejection given by BDT is about 70% higher than that given by IP3D+SV1.

Finally, the performance of BDT for $t\bar{t}$ events when the algorithm is trained with WH events was also studied. For a b-tagging efficiency of 60% one finds that, when the training is performed on WH events, the light jet rejection in $t\bar{t}$ events is decreased by about 20%. For purified light jets, the light jet rejection is decreased by about 25%. The reduction in the performance of BDT is expected since the algorithm was trained to learn a different pattern. Still, the performance of BDT in $t\bar{t}$ events trained with WH events is about 20% higher than that of IP3D+SV1. For purified light jets the performance of BDT is about 25% higher than that of IP3D+SV1.

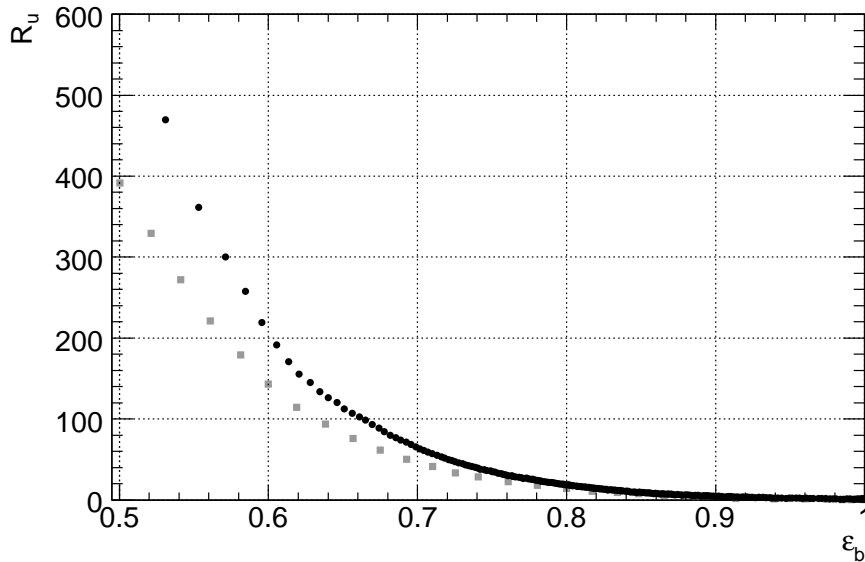


Figure 3: Light jet rejection as a function of b-jet efficiency for $t\bar{t}$ events, given by boosted decision trees (black dots) and IP3D+SV1 tagger (gray squares).

6 Conclusions

The performance of boosted decision trees for tagging b-jets at ATLAS was evaluated using a sample of $WH \rightarrow l\nu_l q\bar{q}$ and $t\bar{t} \rightarrow b\bar{b}l\nu_l j j$ Monte Carlo events. The results show that boosted decision trees outperform the IP3D+SV1 tagger. For the WH channel and for a b-tagging efficiency of 60%, the light jet rejection given by BDT is about 50% higher than that given by IP3D+SV1. For the $t\bar{t}$ channel the light jet rejection given by BDT is about 45% higher than that given by IP3D+SV1. When purified light jets are selected, the light jet rejection given by BDT in $t\bar{t}$ events is about 70% higher than that given by IP3D+SV1.

Acknowledgments

I would like to thank Laurent Vacavant and Karl Jakobs for reading the manuscript and for the valuable comments. This work was supported by grant SFRH/BPD/20616/2004 of Fundação para a Ciência e Tecnologia.

References

- [1] S. Corr eard, V. Kostioukhine, J. Lev eque, A. Rozanov and J.B. de Vivie, *b-tagging with DC1 data*, ATL-PHYS-2004-006 (2004).
- [2] M. Wolter and A. Kaczmarska, *Combining b-tagging methods using a neural network approach*, ATL-INDET-2000-023 (2003).
- [3] B.P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu and G. McGregor, *Boosted decision trees, an alternative to artificial neural networks*, Nucl. Instr. and Meth. A543 (2005) 577.
- [4] I. Narsky, *Optimization of signal significance by bagging decision trees*, physics/0507157.

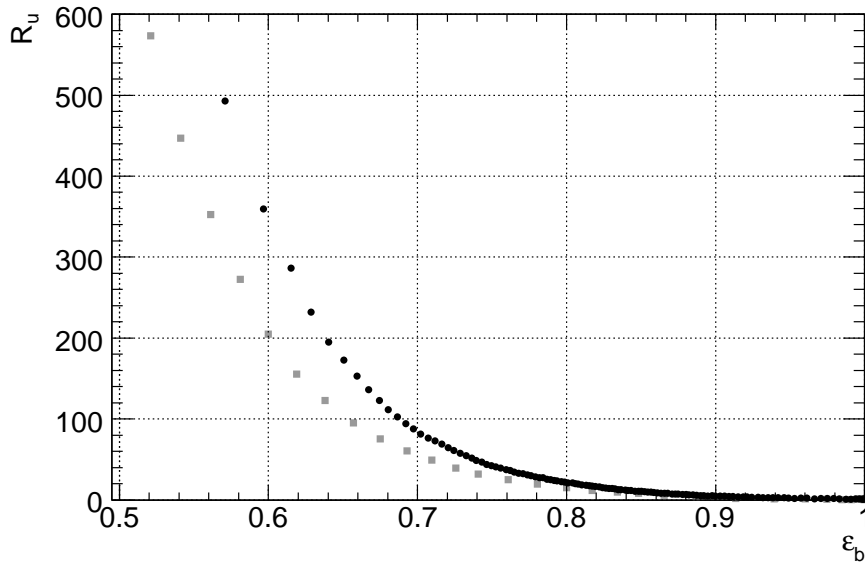


Figure 4: Light jet rejection as a function of b-jet efficiency for $t\bar{t}$ events after purification of light jets, given by boosted decision trees (black dots) and IP3D+SV1 tagger (gray squares).

- [5] V. M. Abazov *et al.* (D0 Collaboration), *Evidence for production of single top quarks and first direct measurement of $|V_{tb}|$* , hep-ex/0612052.
- [6] J. Conrad and F. Tegenfeldt, *Applying rule ensembles to the search for supersymmetry at the Large Hadron Collider*, JHEP 0607, 040 (2006).
- [7] J. Bastos, *Tagging heavy flavours with boosted decision trees*, physics/0702041 (2007).
- [8] Y. Freund and R.E. Schapire, *Experiments with a new boosting algorithm*, in: Proceedings COLT, ACM Press, New York, (1996) 209.
- [9] I. Narsky, *StatPatternRecognition: A C++ package for statistical analysis of High Energy Physics data*, physics/0507143 (2006).
- [10] H. Bachacou *et al*, *b-tagging performance with soft muons*, ATLAS CSC-BT3 note, in preparation.
- [11] F. Derue *et al*, *b-tagging performance with soft electrons*, ATLAS CSC-BT4 note, in preparation.
- [12] T. Sjostrand, P. Eden, C. Friberg, L. Lonnblad, G. Miu, S. Mrenna and E. Norrbin, *Computer Physics Commun.* 135 (2001) 238.
- [13] B. P. Kersevan and E. Richter-Was, *The Monte Carlo event generator AcerMC version 1.0 with interfaces to PYTHIA 6.2 and HERWIG 6.3*, *Comput. Phys. Commun.* 149 (2003), 142.
- [14] V. Kostioukhine, *Secondary vertex based b-tagging*, ATL-PHYS-2003-033 (2003).