Boosting b-tagging performance at high transverse momentum

Daria Koshelenko (National Research Nuclear University MEPhI)
Main supervisor: Dr. Andrea Coccaro (University of Geneva)
2nd supervisor: Dr. Richard John Hawkings (CERN)

August 18, 2017

Abstract

B-tagging algorithms play a vital role in many physics analysis. Increasing of the collision energy in the Run 2 data led to a substantial increase of the flight length of b-hadrons, which makes the b-tagging process more complicated. In this paper, the studying of improving the b-tagging algorithm by using the number of clusters in Insertable b-layer and a Pixel Detector of the ATLAS detector is presented.

Introduction

The identification of b-jets which is usually called as b-tagging is crucial for the ATLAS experiment in many measurements like studying properties of the Higgs boson, and for searching new physics beyond the Standard Model.

Generally, b-tagging is based on the reconstruction of charged particle tracks in the Inner Detector of the ATLAS experiment. During the long shutdown between Run1 and Run2 of the LHC machine, the Insertable B-Layer (IBL) [1], was installed in the ATLAS detector between a new beam pipe with a smaller radius and the Pixel Detector.

It is expected that a significant improvement in the resolution because of the IBL will lead to a significant improvement of b-tagging. The enhancements of the b-tagging algorithms will be described in this note.

This note is organized as follows. The LHC machine and the ATLAS detector are briefly described in sections 2 and 3 respectively. The event samples and the jet selection selections are described in Section 4. Section 5 presents the description of b-tagging, it's main algorithms and the final b-tagging algorithms based on them. It's improvement process described in detail in Section 6. Finally, conclusions are presented in Section 7.

1. The LHC Machine

The Large Hadron Collider (LHC) is a circular hadron accelerator based at CERN near Geneva. This machine uses the 27 km underground tunnel and collides protons with the centre of mass energy equal to 13 TeV. The LHC hosts four experiments: ATLAS, CMS, ALICE and LHCb located at Points 1,5, 2 and 8 respectively.

2. The ATLAS Detector

The ATLAS Detector [2] (A Toroidal LHC ApparatuS) is a multipurpose 4π-particle detector with a cylindrical geometry which has 25 meters in height, 44 meters in length and weighs about 7000 tons. ATLAS includes an Inner Detector (ID), a calorimeter system, a Muon Spectrometer (MS), a magnet system and a trigger system.

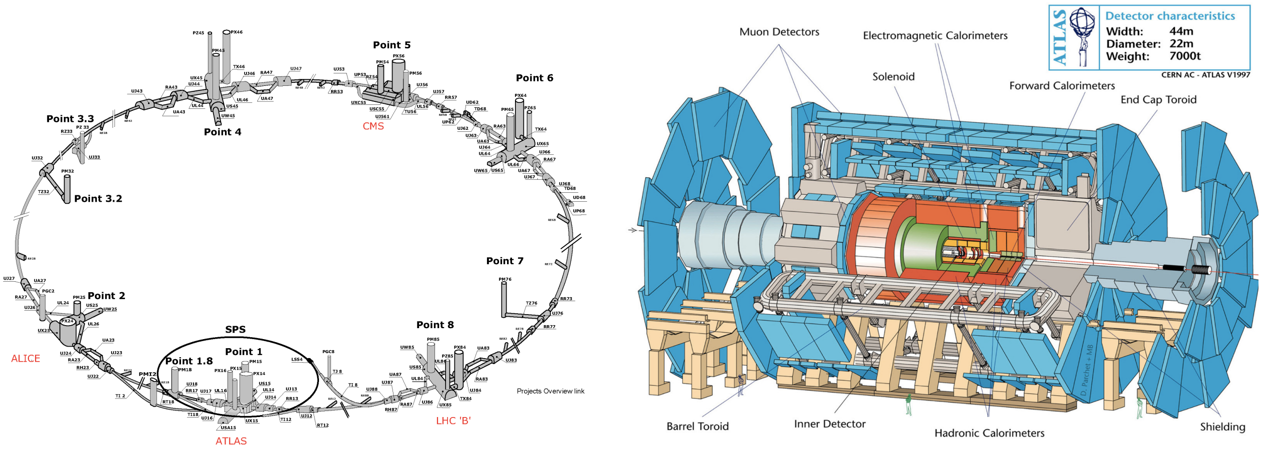

Figure 1: The LHC machine and the ATLAS Detector.

The inner detector (ID) ensures precision track measurements of charged particles in the pseudorapidity range of $|\eta| < 2.5$. It consists of the Insertable B-layer (IBL) which is the innermost pixel layer, three silicon pixel layers of the Pixel Detector and eight layers of silicon microstrip detectors of the Semiconductor Tracker surrounded by straw tubes of the Transition Radiation Tracker, which enhances electron identification in the region $|\eta| < 2.0$. All these subsystems combined provide a measurement of 50 points per track. The size and mutual arrangement are presented on figure 2.

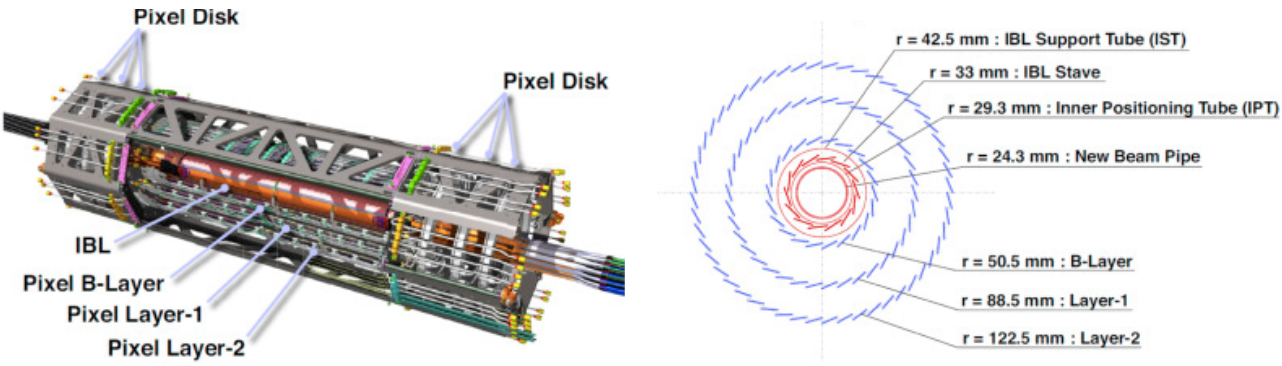

Figure 2: IBL and Pixel Detector of the ATLAS Detector.

The ATLAS calorimetry system includes the inner Electromagnetic (EM) surrounded by a Hadronic calorimeter (HC). The EM calorimeter utilizes liquid argon (LAr) as the active material and lead as the sampling medium covering the pseudorapidity range $|\eta| < 3.2$. A steel/scintillating-tile HC provides coverage for $|\eta| < 1.7$. The endcap and forward regions span $1.5 < |\eta| < 4.9$.

The MS consists of high-precision tracking chambers and a trigger system that provide triggering and tracking capabilities in the ranges $|\eta| < 2.7$ and $|\eta| < 2.4$ respectively.

The ATLAS magnet system includes a central solenoid surrounding the ID and producing a 2 T axial magnetic field, and a toriod system consisting of a barrel plus two endcaps which produces a maximum of 4 T in the MS.

The ATLAS trigger system consists of a harware-based Level-1 trigger followed by a software-based High Level Trigger (HLT) and Event Filter based on software frameworks. It allows reducing the accepted rate from 40 MHz to 1 KHz, and these are passed on to a data storage system for offline analysis.

3. B-tagging process

Due to the strong interactions, quarks can't be free so once they are produced in proton-proton collisions, they combine into hadrons and then form a hadronic jet. For understanding the nature of every event happened in the collision it's critically important to know the type of jet. In simulation it's defined as follows: in a cone with $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} < 0.3$ the tracks associated with this jet as shown on figure 3 are explored (so-called jet labelling). If it contained a b-hadron, this jet is labeled as a b-jet. In case of failure, this operation repeats for searching for c-hadrons and tau-leptons. If nothing was found than it's a light-flavour jet.

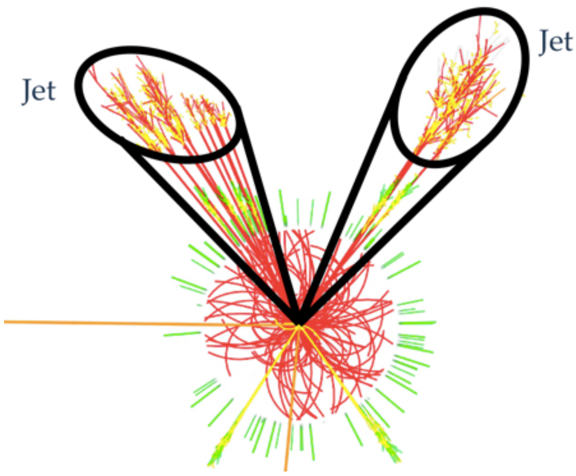

Figure 3: The b-tagging process.

4. Samples and jet selection

All the b-tagging steps were studied and enhanced with $t\bar{t}$ events corresponding to 13 TeV proton-proton collisions simulated with Powheg and Pythia6 Monte Carlo generators and CT10 parton distribution functions. Event generator samples are passed through GEANT4-based full simulation.

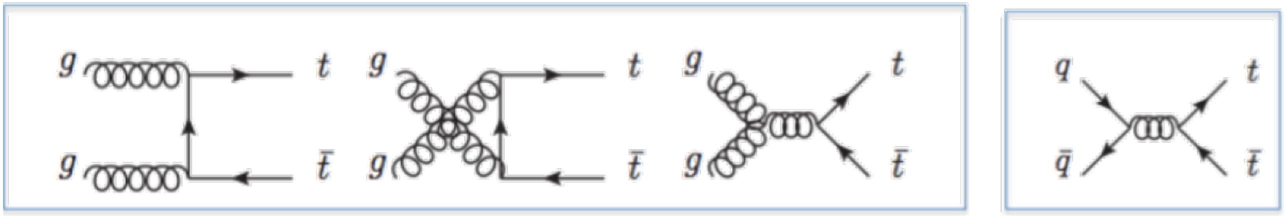

Figure 4: Feynman diagrams of $t\bar{t}$ events production.

Jet selection includes following selections:

- Transverse moment of the jet should be bigger than 20 GeV due to the too high background below this threshold: $p_T > 20$ GeV;
- Pseudorapidity selection: $|\eta| < 2.5$ because of the tracking detector acceptance;
- JVT>0.59 for jets with $p_T < 60$ GeV and $|\eta| < 2.4$. This selection is responsible for pileup jets removal. Pile-up is a low additional transverse momentum formed in the collisions of proton bunches in the processes of supplemental interactions in additional to signal. While in-time pileup arises from additional pp interactions in the current bunch-crossing, out-of-time pileup refers from previous ones. They can both be removed by a selection on the jet vertex tagger (JVT) [3]. The JVT algorithm based on KNN neural network and uses a part of jet's total momentum associated with the primary vertex as an input variables.

5. B-tagging algorithms

There are three basic b-tagging algorithms used in ATLAS physics analyses [4]:

- Impact parameter based algorithm;
- Inclusive secondary vertex reconstruction algorithm;
- Decay chain multi-vertex reconstruction algorithm.

Impact parameter based algorithm use the significance of the impact parameter of the tracks matched to the jet. The probability density functions (PDFs) for this value is used to define ratios of the b- and light-flavour jet hypotheses.

Secondary Vertex Finding Algorithm (SV) based on the reconstruction of the secondary vertex inside the jet. The first step consists of reconstructing two-track vertices using the candidate tracks. Events go through the preselection: tracks are discarded if a secondary vertex they form can be identified as a vertex of a long-lived particle's decay, photon conversions or hadronic interactions. Second step is a reconstruction of a single vertex using the tracks that pass the preselection process. Outlier tracks are removed iteratively.

Decay Chain Multi-Vertex Algorithm (JetFitter) uses the topological structure of weak b- and c-hadron decays inside the jet and tries to reconstruct the full PV → b → c-hadron decay chain.To find a common line on which the primary vertex and the bottomand charm vertices lie, the Kalman filter is used.

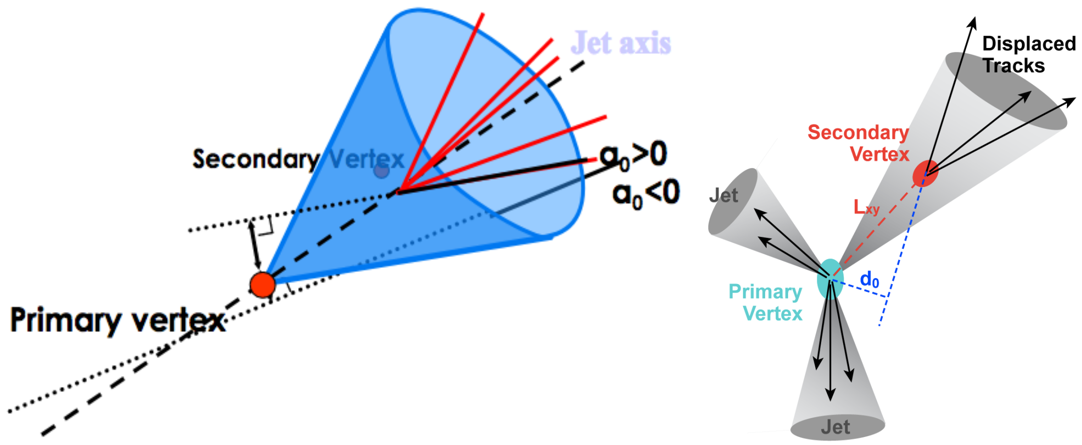

Figure 5: Event analysis in b-tagging algorithms.

For combining the benefits of all these algorithms, the Multi-variative algorithm (MV2) is used. It's defined as an output of BDT algorithm that uses 24 variables which are output from 3 basic ones as input parameters and also jet's transverse momentum and pseudo-rapidity. All these variables are summarized in Table 1. In this note the MV2c10 which is a training set consisting of b-jets as a signal and a mix of 93% light-flavour jets and 7% c-jets as a background is used.

For the MV2c10 distribution analysis the values of efficiency and rejection should be explained. Signal efficiency is the ratio of the number of events with b-jets that passed the selection to the total number of b-jets events, background rejection of c- and light-flavour jets events is defined as an inverse efficiency:

$$\varepsilon = \frac{N_{passed}}{N_{all}}, \quad r = \frac{1}{\varepsilon}.$$

Working points are defined as an MV2 cut value corresponding to a fixed b-jet efficiency on a $t\bar{t}$ sample. Table 2 shows the four working points calculated for the

MV2c10 b-tagging algorithm which include a value of MV2c10 cut, signal b-jets efficiency and background rejection, where background is light-jets and c-jets events together.

Input	Variable	Description
Kinematics	$p_T(jet)$	Jet transverse momentum
	$\eta(jet)$	Jet pseudo-rapidity
IP2D, IP3D	$\log(P_b/P_{\text{light}})$	Likelihood ratio between the b- and light jet hypotheses
	$\log(P_b/P_c)$	Likelihood ratio between the b- and c-jet hypotheses
	$\log(P_c/P_{\text{light}})$	Likelihood ratio between the c- and light jet hypotheses
SV	$m(\text{SV})$	Invariant mass of tracks at the secondary vertex assuming pion masses
	$f_{\text{E}}(\text{SV})$	Fraction of the charged jet energy in the secondary vertex
	$N_{\text{TrkAtVtx}}(\text{SV})$	Number of tracks used in the secondary vertex
	$N_{\text{2TrkVtx}}(\text{SV})$	Number of two track vertex candidates
	$L_{xy}(\text{SV})$	Transverse distance between the primary and secondary vertices
	$L_{xyz}(\text{SV})$	Distance between the primary and secondary vertices
	$S_{xyz}(\text{SV})$	Distance between the primary and secondary vertices divided by its uncertainty
	$\Delta R(\text{jet}, \text{SV})$	ΔR between the jet axis and the direction of the secondary vertex relative to the primary vertex
Jet Fitter	$N_{\text{2TrkVtx}}(\text{JF})$	Number of 2-track vertex candidates (prior to decay chain fit)
	$m(\text{JF})$	Invariant mass of tracks from displaced vertices assuming pion masses
	$S_{\text{xyz}}(\text{JF})$	Significance of the average distance between the primary and displaced vertices
	$f_{\text{E}}(\text{JF})$	Fraction of the charged jet energy in the secondary vertices
	$N_{\text{1-trk vertices}}(\text{JF})$	Number of displaced vertices with one track
	$N_{\geq\text{2-trk vertices}}(\text{JF})$	Number of displaced vertices with more than one track
	$N_{\text{TrkAtVtx}}(\text{JF})$	Number of tracks from displaced vertices with at least two tracks
	$\Delta R(\vec{p}_{\text{jet}}, \vec{p}_{\text{vtx}})$	ΔR between the jet axis and the vectorial sum of the momenta of all tracks attached to displaced vertices

Table 1: The 24 input variables used for the MV2 b-tagging algorithm.

MV2c10 cut value	b-jets efficiency	Background rejection
0.6022	60	19.2
0.7011	70	10.1
0.7715	77	6.4
0.8491	85	3.7

Table 2: Working points for the MV2c10 b-tagging algorithm.

6. MV2c10 Algorithm improvement.

It is of interest to study the possibility of improvement the MV2c10 algorithm by using number of hits in the IBL and all the layers in the Pixel Detector. A hit is defined as a pixel segment in which at least 0.05 MeV had been deposited [5].

New MV2c10 based on boosting decision tree that uses not only the previous 24 variables which have already been discussed but also the number of clusters in these 4 layers. The distributions of them are presented in figure 6.

For characterising information about the efficiency at various purity levels, and the integral of the efficiency-rejection (the "ROC") curve is presented on the figure 7.

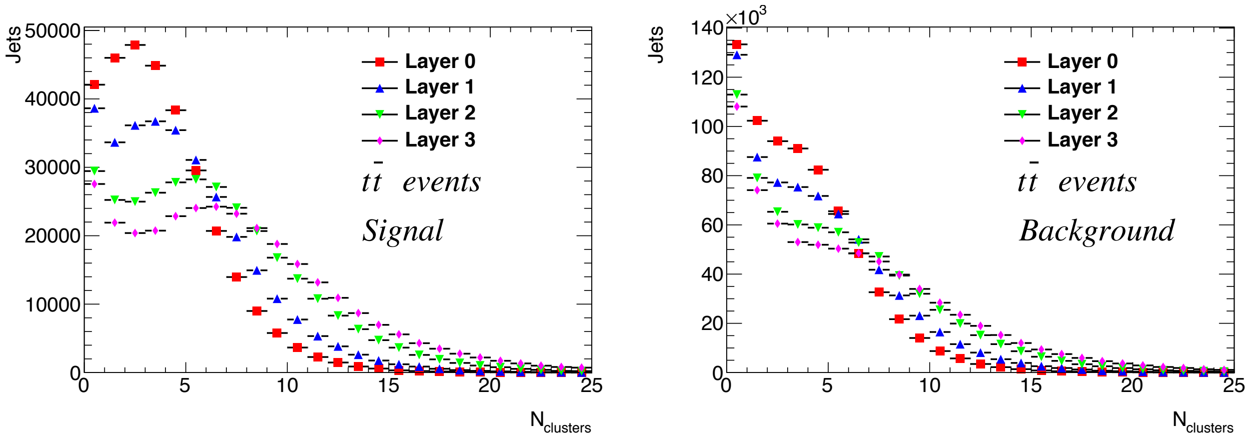

Figure 6: new variables distributions.

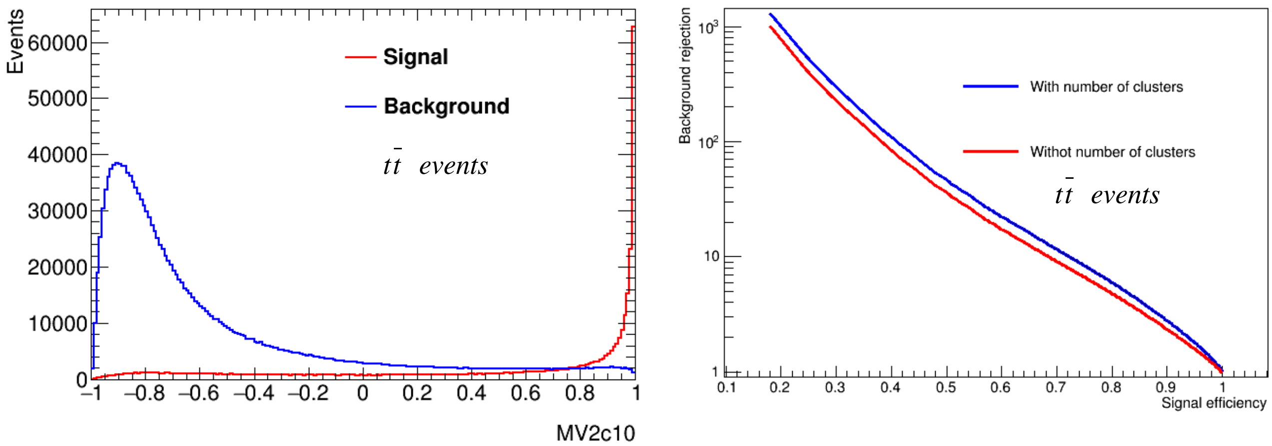

Figure 7: MV2c10 distribution and ROC-curve for MV2c10 training.

7. Results and conclusion.

To assess the appropriateness of using the proposed changes of MV2c10 algorithm, the values of rejection for working points were calculated and presented in the table 3. Comparison of the presented results demonstrates the growth of background rejection, which allows us to conclude that the use of the number of hits is advisable. The update of the b-tagging MV2c10 algorithm improved the background rejection by 10 percent which can be considered a significant achievement. The algorithm should also be tested on higher statistics sample. In the future it seems expedient to test the working capacity of a new version of the MV2c10 algorithm using Run2 data.

b-jets efficiency [%]	MV2c10 cut value	Background rejection without number of clusters, r_0	Background rejection with number of clusters, $r_{clusters}$	$k = \frac{r_{clusters}}{r_0}$
60	0.6022	19.2	21.9	1.1
70	0.7011	10.1	11.5	
77	0.7715	6.4	7.3	
85	0.8491	3.7	4.2	

Table 3: Working points for the MV2c10 b-tagging algorithm updates.

8. References

[1] ATLAS Collaboration, ATLAS Insertable B-Layer Technical Design Report, CERN-LHCC-2010-013. ATLAS-TDR-19. https://cds.cern.ch/record/1291633

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

[3] ATLAS Collaboration, Tagging and suppression of pileup jets with the ATLAS detector, ATLAS-CONF-2014-018.

[4] ATLAS Collaboration, Expected performance of the ATLAS b-tagging algorithms in Run-2, ATL-PHYS-PUB-2015-022.

[5] B. Todd Huffman, Thomas Russell, and Jeff Tseng, Tagging b quarks without tracks using an Artificial Neural Network algorithm, arXiv:1701.06832v3 [hep-ex] .