PROCEEDINGS OF SCIENCE

An Overview of the b-Tagging Algorithms in the CMS Offline Software

Christophe Saout[*]
CERN, Geneva, Switzerland
E-mail: christophe.saout@cern.ch

The CMS Offline software contains a widespread set of algorithms to identify jets originating from the weak decay of b quarks. Different physical properties of B hadron decays like lifetime information, secondary vertices and soft leptons are exploited. The variety of selection algorithms range from simple and robust ones, suitable for early data-taking and online environments as the trigger system, to highly discriminating ones, exploiting all the information available. For the latter, a generic discriminator computing framework has been developed that allows to exploit the full power of multi-variate analysis techniques in a flexible way.

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

[*] on behalf of the CMS collaboration

1. Introduction

The ability to efficiently identify jets from the hadronisation of b quarks plays a crucial role in many of the physics analyses at the LHC. This includes Standard Model signatures like $b\bar{b}$ or $t\bar{t}$ production, Higgs production in association or decaying into $b\bar{b}$ pairs as well as e.g. SUSY channels with $\tilde{t}$ decays.

The b quark exhibits particular properties that make it possible to tag jets that contain B hadrons, namely the lifetime (≈ 1.5 ps), mass (≈ 4.2 GeV/c^2) and the high decay multiplicity (≈ 5 charged tracks on average). Furthermore, the b quark decays weakly, giving rise to $e^{\pm}$ and $\mu^{\pm}$ in 36% of the cases and possibly also causing subsequent c quark or τ decays.

The constructable signatures exploited by the lifetime-based b-tagging algorithms are tracks with significant impact parameters and secondary vertices. The CMS all-silicon tracking system [1] with impact parameter resolutions of up to 10 μm in the $r-\phi$ and 20 μm in the $r-z$ plane is particularly suited for this task.

2. Basic b-Tagging Algorithms

The simplest tagging algorithm directly utilises the measured impact parameters of the tracks. In order to get the best possible handle on the compatibility of the track with the primary vertex, the signed 3-D impact parameter significance is used, which is defined as $sign \cdot \frac{value_{IP}}{error_{IP}}$. The sign is chosen depending on the hemisphere of the point of closest approach of the track to the primary vertex with respect to the jet direction, so that tracks from decays with lifetime populate the positive side. Quality cuts on the tracks stringently reject possible fake tracks (which tend to exhibit high impact parameters) by requiring very well reconstructed tracks and a minimum distance of 0.7 mm to the jet axis and an $IP < 2$ mm.

The “Track Counting” algorithms [2] choose the signed I.P. significance of exactly one track per jet as algorithm discriminator, namely the one with the n^{th}-highest value in the jet, where n is 2 (the “High Efficiency” variant) or 3 (“High Purity”). Tracks are considered to lie inside the jet, if the track momentum fulfills $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} < 0.5$ with respect to the jet axis. Figure 1 shows the distribution of the discriminator, where the large positive tails for b-jets clearly stand out. This simple algorithm is also run in the time-critical CMS online high-level trigger.

The “Jet Probability” algorithm [2] goes a step further and takes all signed track I.P. significances in a jet into account. For each individual track, a probability is computed for the track being compatible with the primary vertex by looking up a *pdf*, defined for different track quality categories. These individual probabilities are then combined into a probability for the whole jet. This algorithm is known to be among the most powerful ones. The downside, however, is that that the *pdf*’s have to be carefully calibrated. The performance of the algorithm is depicted in figure 2. Note that in the light flavour jets gluons are also included and their splitting into heavy flavour quarks is excluded.

The “Simple Secondary Vertex“ [3] algorithm uses the significance of the 3-D flight distance distance between a reconstructed secondary and the primary vertex. The primary and secondary vertices are reconstructed using an adaptive reconstruction algorithm using simulated annealing techniques for robustness [4]. The performance of the algorithm is comparable to the “Track

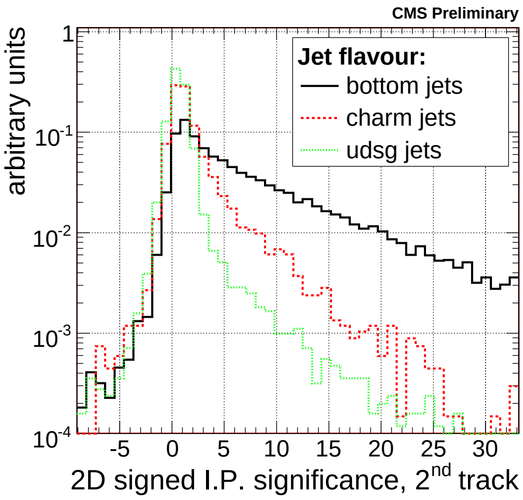

Figure 1: The 2-D signed impact parameter significance of the second-most significant track

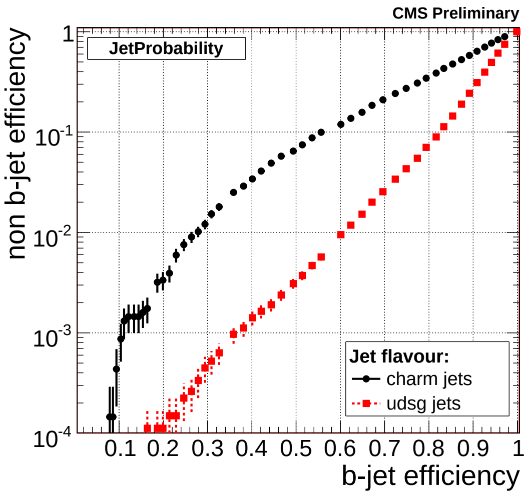

Figure 2: The performance of the "Jet Probability" algorithm on an inclusive $pp \rightarrow t\bar{t}$ sample

Counting" algorithm, but an outstanding property of this algorithm is its robustness with respect to non-optimal alignment of the tracking system. This makes the algorithm particularly interesting for early data-taking scenarios, as can be seen in figure 3.

The "Soft Lepton" algorithms [5] exist in different flavours and combine several variables from a reconstructed lepton inside the jet from a semi-leptonic b- or c-decay using a simple feed-forward neural network. The most discriminating variables are the lepton's I.P. and transverse momentum with respect to the jet axis. Figure 4 shows the $e^{\pm}$ and $\mu^{\pm}$ taggers, the latter with an additional variant that does not include the lepton I.P. for the discriminator computation.

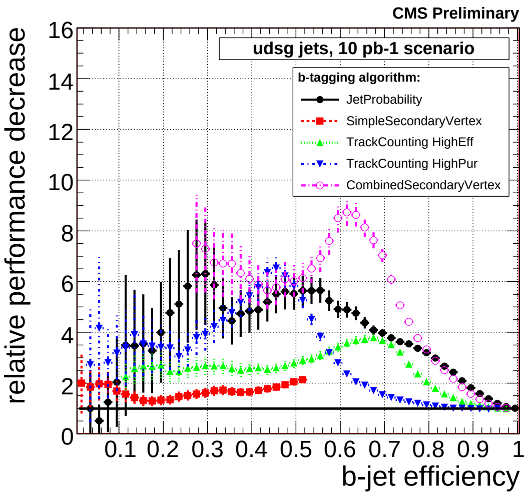

Figure 3: The relative increase in mistag rate at a given b-tagging efficiency with respect to ideal detector alignment at an expect tracker alignment achievable with 10 pb^{-1}

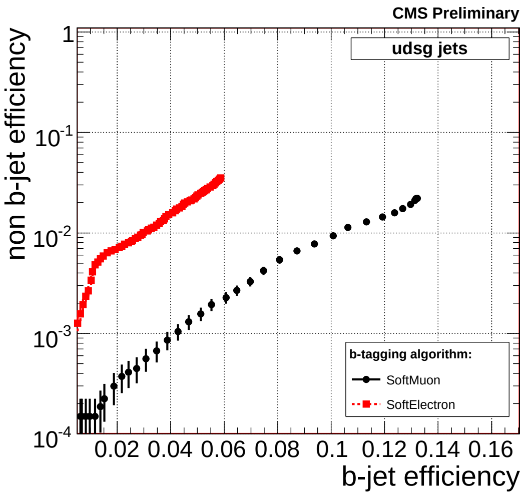

Figure 4: The performance for three "Soft Lepton" b-tagging algorithms

3. Multivariate Analysis Techniques for Event Reconstruction

Within the CMS software environment (CMSSW), a dedicated abstraction layer allows the transparent use of Multivariate Analysis Techniques for purposes of event reconstruction. Its modular implementation allows almost arbitrary combination and layering of variable transformations in order to obtain a final discriminator. Very simple and commonly used techniques like linear discriminants and Likelihood Ratios are available as built-in modules, as well as a set of standard preprocessing techniques. More sophisticated algorithms, like Artificial Neural Networks or Boosted Decision Trees, are available through third-party interface plugins, particularly to the Toolkit for MultiVariate Analysis (`TMVA` [6]), a comprehensive MVA package that is bundled with `ROOT` [7]. The framework allows a seamless integration of these packages into the software environment, including full access to to the CMS Conditions Database for storage and live retrieval of training data. This allows to run natively inside the framework without requiring a separate intermediate data format e.g. for training. The layout of the evaluation network is freely configurable via an XML trainer description language, and an example is shown in figure 5.

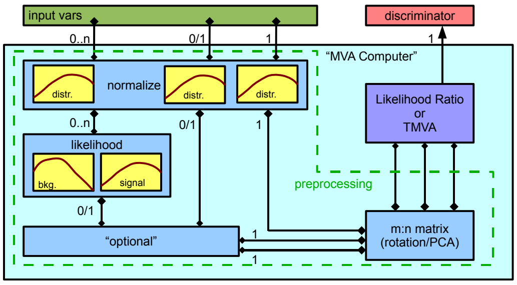

Figure 5: An example of how the MVA network layout is structured for the Artificial Neural Network variant of the "Combined Secondary Vertex" b-tagging algorithm

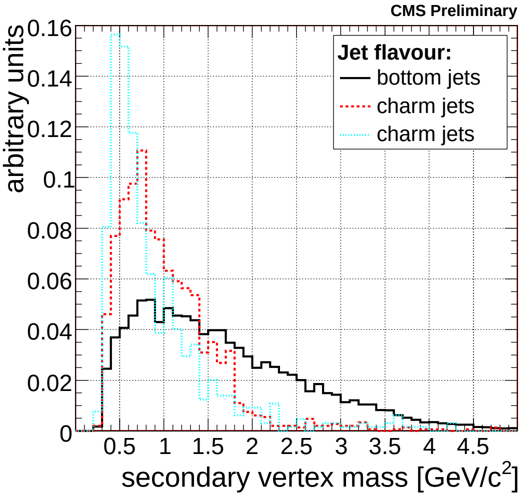

Figure 6: The secondary vertex mass input variable for the "Combined Secondary Vertex" algorithm

By feeding all variables computed by the low-level b-tagging algorithms listed in the previous chapter into the MVA architecture, an algorithm exploiting all the possible physics content can easily be constructed to provide a best-performing algorithm. One such example is the "Combined Secondary Vertex" algorithm [8], which essentially combines most information available in the event. This includes track variables like signed I.P. significances, but mainly, if available, secondary vertex variables like flight distance significance, invariant mass (as shown in figure 6) and energy fraction of the secondary vertex. The default version of the algorithm combines all variables using a Likelihood Ratio with *pdf*'s in different categories. A variant using an Artificial Neural Network is also available. If well trained, this algorithm yields a significant improvement over the already well-performing "Jet Probability" tagger.

A very different approach at combining discriminative information in order to obtain a more efficient tagging algorithm is to combine the discriminators of existing algorithms based on complementary physical input. For instance the output of the lifetime tags can be combined with the output of the soft lepton tags. By using a neural network to combine the output of the "Combined Secondary Vertex" tagger and the two $e^{\pm}$ and $\mu^{\pm}$ "Soft Lepton" tags, the b-tagging efficiency can be increased from 60% to about 65% for the same light flavour mistag rate.

4. Conclusions

A comprehensive set of *b*-tagging algorithms are ready for use in CMS, based on impact parameters, secondary vertices and soft leptons. The simple algorithms will play an important role for the first data-taking, whereas later the discriminative power of Multivariate Analysis Techniques will allow for the best possible b-tagging performance for discovery purposes.

References

[1] **CERN/LHCC 2006-001**, CMS Collaboration: "The CMS Physics Technical Design Report, Vol 1"

[2] **CMS NOTE-2006/019**, A.Rizzi, F.Palla, G.Segneri: "Track impact parameter based b-tagging with CMS"

[3] **CMS PAS BTV-07-003**, CMS Collaboration: "Effect of misalignment on b-tagging"

[4] **CMS NOTE-2007/008**, R.Fruehwirth, W.Waltenberger, P.Vanlaer: "Adaptive Vertex Fitting"

[5] **CMS NOTE-2006/043**, A.Bocci, P.Demin, R.Ranieri, S.de Visscher: "Tagging b jets with electrons and muons at CMS"

[6] J. Stelzer: "TMVA - the toolkit for multivariate data analysis", proceedings of the ACAT2008 conference, also see the URL http://tmva.sourceforge.net/

[7] "ROOT - An Object-Oriented Data Analysis Framework" see the URL http://root.cern.ch/

[8] **CMS NOTE-2006/014**, C.Weiser: "A Combined Secondary Vertex Based B-Tagging Algorithm"