CERN summer student programme project report

Studying b- and c-tagging for a potential $H \to c\bar{c}$ measurement

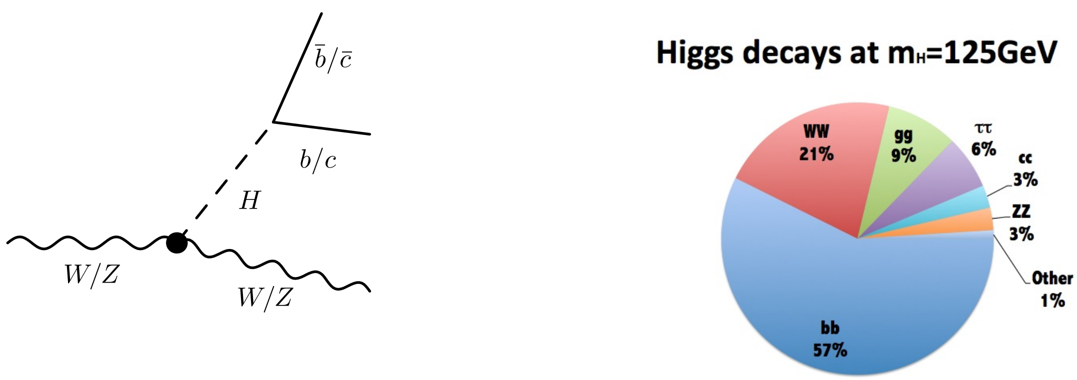
Vytautas Dūdėnas

August 8, 2014

1 Introduction

After the discovery of the Higgs boson [1, 2], we have turned to the study of its properties. In particular, we want to measure the couplings to quarks. Therefore we study the Higgs produced in association with a W or Z boson and decaying into two quarks. The Feynman diagram of this process is shown in figure 1(a). Because of the large background, this is a challenging analysis[3].

57% of the time it decays into bottom - antibottom pair and 3% of the time - $c\bar{c}$ as shown in Figure1(b). That makes measuring the decay $H \to c\bar{c}$ even more difficult. The background can be reduced on the $H \to b\bar{b}$ final state by making use of the latest algorithms for the identification of b quarks (b-tagging) [4]. Moreover, c-tagging is harder than b-tagging. Hence for this process to be identified, there is a need for improvement in c-tagging techniques.



(a) W/Z channel for Higgs production

(b) Branches of Higgs decays

Figure 1: Higgs production and decays

1.1 Jet identification

In experiment we do not observe quarks due to the confinement, but hadrons, which are clustered into jet. We, nevertheless, try to determine the flavour of the quark from the hard process inside the jet. Basic properties of hadrons that let us identify the flavour of the jet are shown in table 1. The crucial properties are the lifetime and the mass of the hadrons. B hadrons have a lifetime of 1.6 ps, so the distance they travel from their production until decaying into jets is significant enough to be reconstructed from the tracks of the decay products. The place of this decay is called secondary vertex and the parameters of it can be reconstructed. On the other hand, light-flavoured hadrons typically have a very small lifetime, hence the place where the jet produced by it's decay approximately coincides with the primary vertex (where it has been produced). Therefore the identification is based on the secondary vertex reconstruction of the b flavoured hadron. The typical jets coming from these hadrons are shown in figure 2.

Flavour	B	C	UDS
mass	$5.3\,GeV$	$1.7-1.8GeV$	$O(MeV)$
hadron lifetime ($c\tau$)	0.5mm	0.1-0.3mm	typically small

Table 1: Basic properties of B, C and UDS flavoured hadrons.

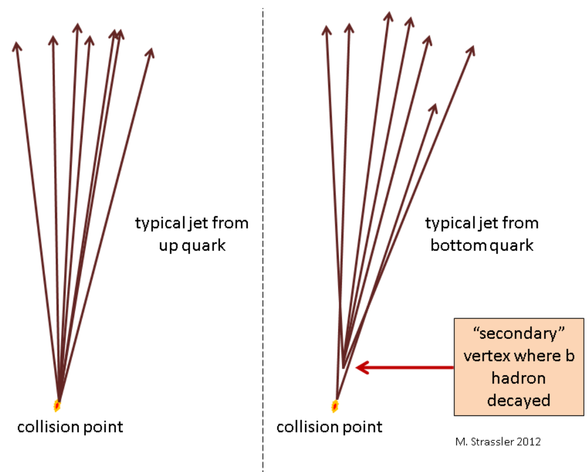


Figure 2: Typical light flavoured and b flavoured jets

c flavoured hadrons have smaller lifetime than b hadrons, but their secondary vertex can also be reconstructed. Because the difference of masses and lifetimes of b and c flavoured hadrons are not as significant as B vs light, it is harder to distinguish them. In many physics analyses, most of the background is from light-flavoured jets, so for an overall identification, the rejection of light flavoured backgrounds is more important.

For an identification of c jets this whole picture gets even more complicated. Distinguishing between b and c is as challenging as before, but now, since it has smaller lifetime, it is harder to reconstruct the secondary vertex and the discrimination between c vs light is much worse.

1.2 b-tagging procedure

The toolkit for multivariate data analysis (TMVA for short) is used to classify the jets as signal (containing b) or background (no b)[5]. It uses machine learning techniques to classify signal versus background. The training of the classifier is performed as follows. First, one takes jets from QCD or ttbar samples. Then one determines b, c, and light flavoured jets from the generator information. The sample is split into two parts for testing and training. Variables with high discriminating power are given to the TMVA training, to make classification, in particular the signal and background efficiencies for a certain cut on the classifier output value. After training is finished, the evaluation is made on the other part of the sample to check the performance, i.e. signal and background efficiencies.

In the standard CMS CSV b-tagger, they categorize the sample into three separate categories. The categorization is based on whether or not a secondary vertex can be reconstructed. Entries of the sample that have well reconstructed secondary vertex are put into the so-called RecoVertex category. The entries that have at least two tracks with impact parameter significance greater than 2 are put into PseudoVertex category. And when there is no secondary vertex reconstruction possible at all, corresponding entries go into the NoVertex category. These categories have a different number of variables, i.e. PseudoVertex has more variables than NoVertex, because it has some information about secondary vertex, and RecoVertex has the largest number of variables.

	b vs c	b vs udsg	c vs udsg
1	2D distance btw 1st and 2nd V signif.	2nd vertex E/Etot	2nd vertex E/Etot
2	2nd vertex mass	2D distance btw 1st and 2nd V signif.	3D IP significance (highest)
3	2nd vertex E/Etot	2nd vertex mass	jet Eta
4	3D IP significance (highest)	Distance btw jet axis and 2nd V direction	jet Pt
5	3D IP significance (2nd highest)	Eta relative to jet axis (2nd V)	track Eta relative to jet axis (1st V)
6	2D IP of track> m_c	jet Pt	track Eta relative to jet axis (2nd V)
7	track Eta relative to jet axis (1st V)	Eta relative to jet axis (1st V)	2D IP of track> m_c
8	Distance btw jet axis and 2nd V direction	3D IP significance (highest)	2nd vertex mass
9	N of tracks in 2nd V	jet Eta	N of tracks in jet
10	track Eta relative to jet axis (2nd V)	N of tracks in 2nd V	track parallel P along jet axis

Table 2: Important variables in $90 < p_t < 150$ and $abs(\eta) < 1.2$ bin. Variables from the pseudo- and reco-vertex categories are given in blue, and varaibles from the reco-vertex category only are in red.

Method	Settings
BDTG	NTrees=1000::BoostType=Grad:Shrinkage=0.03:UseBaggedBoost:GradBaggingFraction=0.9:nCuts=500:MaxDepth=5:MinNodeSize=0.001"
Neural network	"H:!V:NeuronType=tanh:VarTransform=N:NCycles=600:HiddenLayers=N+5:TestRate=5"
Likelihood	"H:!V:TransformOutput:PDFInterpol=Spline2:NSmoothSig[0]=20:NSmoothBkg[0]=20:NSmoothBkg[1]=10:NSmooth=1:NAvEvtPerBin=50"

Table 3: Settings for methods used in the TMVA training.

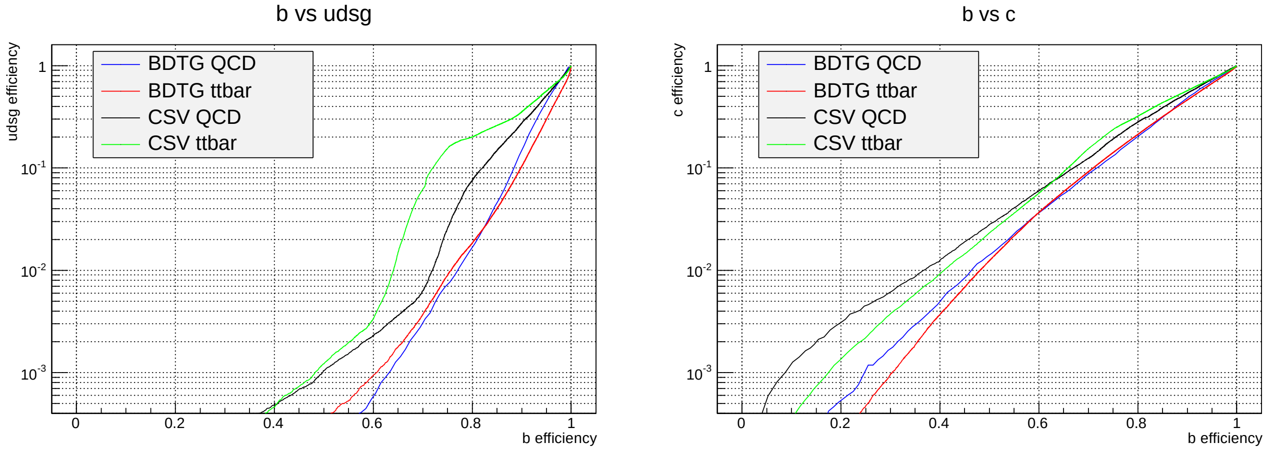
After categorizing in this way, the training is performed separately for those categories. Then, knowing ratios of these categories are combined them and evaluation is made it on all categories. This algorithm is called "combined secondary vertex" (CSV).

2 Results

Instead of training in separate categories, we apply the same category weights as used for CSV before the training and then train all categories together. We find this to be a well performing method for b-tagger. Then, we apply the same procedure for c-tagging. First we try BDTG method for training. All settings used for training as they are booked from TMVA can be found in table 3. We concentrate on $90 < p_t < 150$ GeV and $abs(\eta) < 1.2$ bin. We checked the most important variables from the training as given by TMVA (table 2). The most important variables for b vs light and b vs c are roughly the same. Most of them are from PseudoVertex category with two variables from RecoVertex. The highest discrimination is given by the distance between secondary and primary vertices, mass (summed from tracks) and energy (ratio between energy and total energy) of secondary vertex. For discrimination of c vs light, the variables from RecoVertex become less important, but energy of the secondary vertex still has the highest discrimination power.

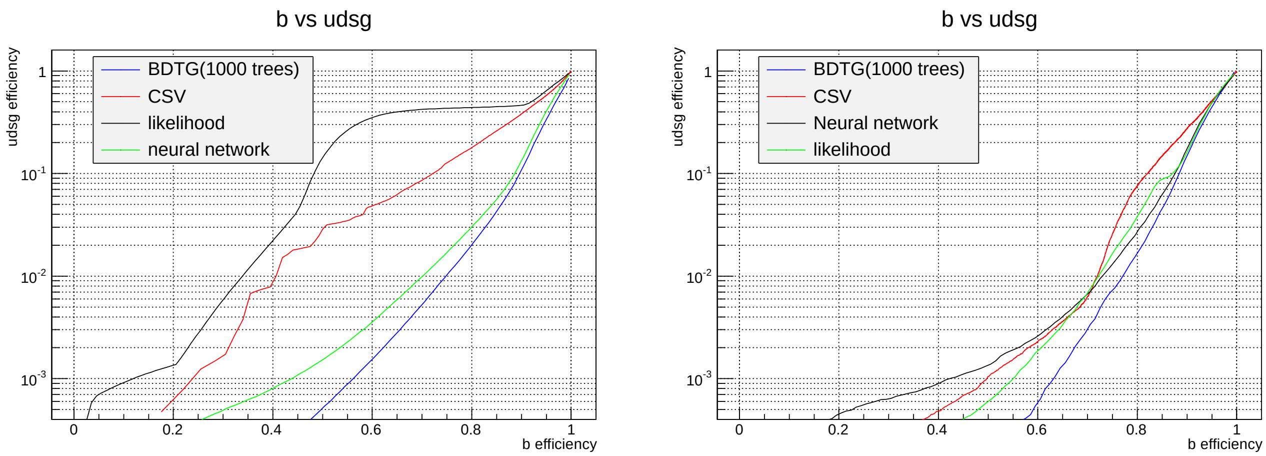
After training with BDTG for b-tagging, we evaluate on ttbar sample and find that the performance is very similar. Hence it suggests that our results are valid. We also compare it with the official CSV output. The efficiencies of signal and background (ROC curves) of evaluation on ttbar and QCD samples are shown in figure 3. Furthermore, we compare the BDTG with other MVA methods, and find best performance for BDTG as well (figure4). So we apply this on c-tagging only in QCD sample to get ROC curves (figure5). The BDTG(1000) has the same settings as before and BDTG(100) has looser settings for training. We compare our performance with the performance obtained by ATLAS [6] in table 4. It shows that we get efficiencies of the same order. Signal efficiencies near different background efficiencies of b- and c- taggers are summarized in table 5.

To estimate the signal significance for $H \to c\bar{c}$ analysis, we use same strategy as $VH \to b\bar{b}$ analysis in ref. [3] for an integrated luminosity of 18.9 fb^{-1}. Taken the same background, we rescaled all cross sections by a factor of 1.785 which corresponds to going from center of mass energy 8 TeV to 14 TeV. Also, knowing branching ratio of $\frac{H\to c\bar{c}}{H\to b\bar{b}} = 0.05$, we obtained $H \to c\bar{c}$ cross section. This rough estimation gives us the best significance ($\frac{S}{\sqrt{S+B}}$, where S and B stands for signal and background respectively) near ~50% of signal efficiency, namely 0.005 (table 6). We can estimate what significance we will have for luminosity of 5000 fb^{-1} by $0.005 * \sqrt{5000/20} \approx 0.08$. Hence there is a little significance for this process and it will be very challenging to measure the $H \to c\bar{c}$ decays process if it has a cross section as in the Standard Model.



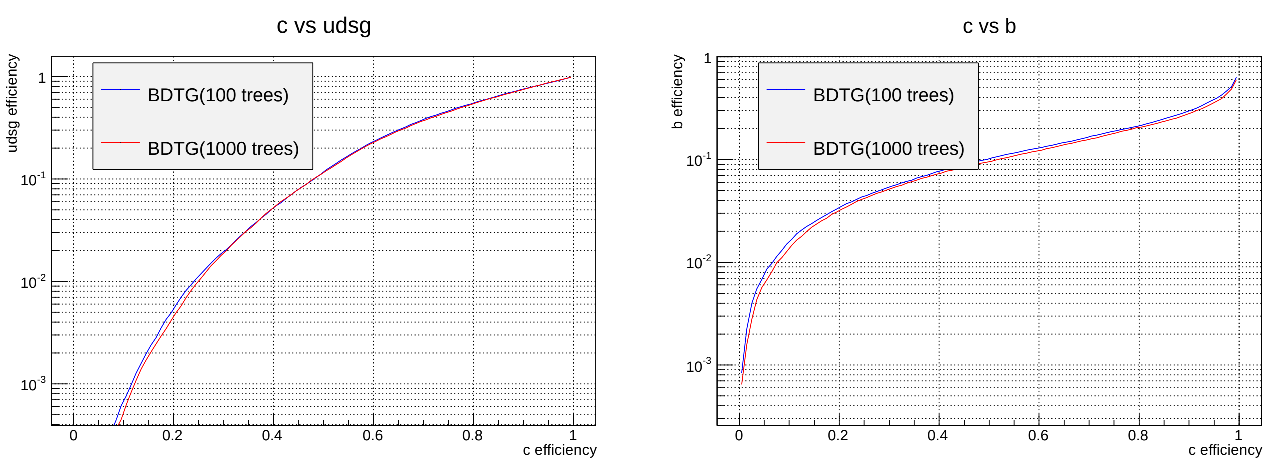
(a) B vs UDSG (b) B vs C

Figure 3: Signal and background efficiencies for $90 < p_t < 150\,\mathrm{GeV}$, $abs(\eta) < 1.2$.



(a) In the whole p_t and η range (b) In the range of $90 < p_t < 150$ and $abs(\eta) < 1.2$

Figure 4: Comparisons of different methods in the whole $p_t$and η range and in a bin of $90 < p_t < 150$ and $abs(\eta) < 1.2$ for b as a signal and udsg as a background.



(a) C vs UDSG (b) C vs B

Figure 5: C - tagging efficiencies in $90 < p_t < 150$ and $abs(\eta) < 1.2$ bin.

	BDTG	Atlas	BDTG	Atlas
c	20%	20%	95%	95%
b	3%	4%	38%	48.5%
udsg	0.45%	0.14%	90%	-

Table 4: Comparisons of efficiencies obtained with the BDTG approach and efficiencies presented in ATLAS.

Background efficiencies(%)	BDTG(%)	CSV(%)
B vs UDSG		
10	89	82
1	77	72
0.1	63	50
B vs C		
10	72	68
1	47	37
0.1	25	8

Background efficiencies(%)	BDTG(%)
C vs UDSG	
10	48
1	25
0.1	12
C vs B	
10	52
1	8
0.1	1

Table 5: Efficiencies for b and c taggers in $90 < p_t < 150$ and $abs(\eta) < 1.2$ bin.

c	48%	25%	12%
b	9%	4%	1.8%
udsg	10%	1%	0.1%
$\frac{S}{\sqrt{S+B}}$	0.005	0.0041	0.0027

Table 6: Signal significance for various efficiencies in $90 < p_t < 150$ and $abs(\eta) < 1.2$ bin.

3 Conclusion

We studied b-tagging by training a multivariate classifier with BDTs. Curves of efficiencies of signal and background show promising results. The validity of the results are justified by checking with another independent sample. Applying the same training to c-tagging we obtain comparable performance as ATLAS[6]. With the efficiencies from the c-tagging evaluation and the analysis strategy of the $VH \to b\bar{b}$ analysis from ref. [3] , we find that the expected significance for a $H \to c\bar{c}$ signal is small.

References

[1] G. Aad *et al.* [ATLAS Collaboration], Phys. Lett. B **716** (2012)

[2] S. Chatrchyan *et al.* [CMS Collaboration], Phys. Lett. B **716** (2012) 30 [arXiv:1207.7235 [hep-ex]].

[3] http://arxiv.org/pdf/1310.3687.pdf

[4] http://cds.cern.ch/record/1581306?ln=en

[5] http://arxiv.org/abs/physics/0703039

[6] http://cds.cern.ch/record/1562880/files/ATLAS-CONF-2013-068.pdf