Application of Machine Learning Based Top Quark and W Jet Tagging to Hadronic Four-Top Final States Induced by SM as well as BSM Processes

Jiří Kvita[a], Petr Baroň[a], Monika Machalová[b], Radek Přívara[a], Rostislav Vodák[b], Jan Tomeček[b]

[a] Joint Laboratory of Optics of Palacký University Olomouc and Institute of Physics of Czech Academy of Sciences, Czech Republic

[b] Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, Czech Republic

Faculty of Science
Palacký University Olomouc

Contact Information:
17. listopadu 1192/12
779 00 Olomouc, Czech Republic

Phone: +420 585 634 028
Email: monika.machalova@upol.cz

Introduction

We study the application of selected ML techniques to the recognition of a substructure of hadronic final states (jets) and their tagging based on their possible origin in current HEP experiments using simulated events and a parameterized detector simulation. The results are then compared with the cut-based method.

Simulations

Jets as hadronic final states are an inevitable consequence of the quantum chromodynamics (QCD) [1], the force between strongly interacting matter constituents of quarks and gluons. In hadron collisions, jets are important final states and signatures of objects of high transverse momentum. In cases of large jet transverse momenta, i.e. with a larger Lorentz boost in the plane perpendicular to the proton beam, decay products of hadronically decaying W bosons or top quarks are collimated so that they form one large boosted jet in the detector.

Preprocessing

1. Specific jets can be identified by focusing on those with masses in the ranges of [60, 100] GeV and [138, 208] GeV, jets outside these ranges are classified as light jets, as a result, we have four subsets: zp-sets and pp-sets for t jets, and zp-sets and pp-sets for W jets
2. Decomposition into the training and the test sets, training sets contain 80% and the test sets 20% of data from the original sets
3. Standardization of datasets by removing the mean and scaling to unit variance

Methods

Metrics

$$\text{Accuracy} \equiv \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

$$\text{Precision} \equiv \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall} \equiv \frac{\text{TP}}{\text{TP}+\text{FN}} \equiv \text{True positive rate}$$

$$\text{False positive rate} \equiv \frac{\text{FP}}{\text{FP}+\text{TN}}$$

Classifiers

- *Gradient boosting classifier* (GBC) - combining multiple simple predictors (here decision trees) to create a more powerful model

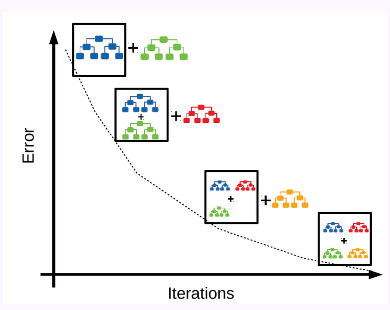

Figure 1: GBC

- *Multi-layer Perceptron classifier* (MLP) - based on neural networks.

Undersampling

- very distorted ratio between t-jets and light-jets (in the direction of t-jets)
- we settled for the undersampling applied to the training sets, which uses various techniques to remove data from the major class
- tested undersampling techniques: *Random undersampling*, *Cluster centroids*, *Near miss*, *Repeated edited nearest neighbor*

Cut-based algorithm

to identify jets coming from the hadronic decays of the W boson or a top quark by a simple cut-based algorithm

- W-jets if

$$0.10 < \tau_{21} < 0.60 \wedge 0.50 < \tau_{32} < 0.85 \wedge m_J \in [60, 100]\,\text{GeV}$$

- top-jets if

$$0.30 < \tau_{21} < 0.70 \wedge 0.30 < \tau_{32} < 0.80 \wedge m_J \in [138, 208]\,\text{GeV}$$

Data Structure

ID	File name	Number of jets
0	ascii_run_XY_pp_2tj_allhad_NLO_ptj1j2min200...	797 363
1	ascii_run_XY_pp_2tj_allhad_NLO_ptj1j2min60...	446 838
2	ascii_run_XY_pp_2tj_allhad_NLO_ptj1min200...	781 675
3	ascii_run_XY_zp_ttbarj_allhad_1000GeV...	449 606
4	ascii_run_XY_zp_ttbarj_allhad_1250GeV...	388 593

→

ID	File name	Number of jets
0	data_zp	838 199
1	data_pp	2 025 876

Table 1: Table of datasets

We have 5 different datasets, from which we subsequently created two new ones. The first one is the unification of the *zp-sets* (IDs 3 and 4) and the second one is the unification of the *pp-sets* (IDs 0–2).

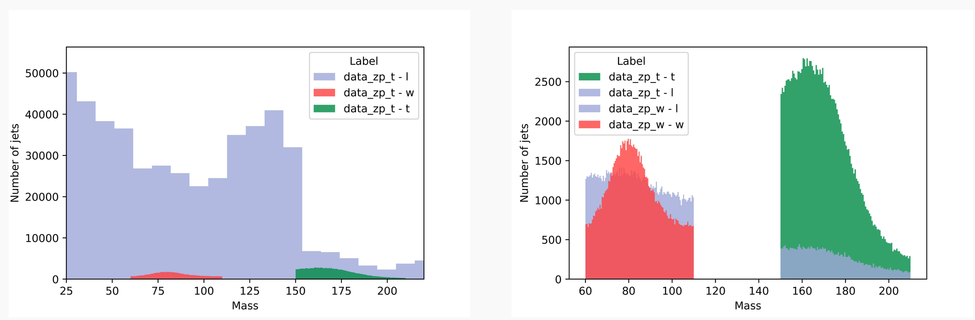

Figure 1: Structure of data_zp

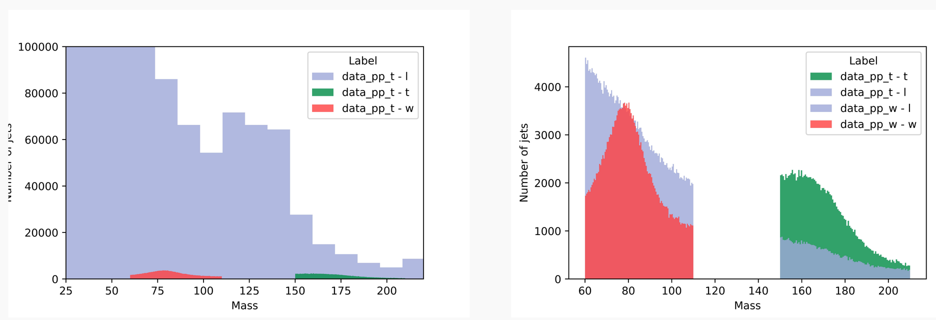

Figure 2: Structure of data_pp

The ratios between t-jets (W-jets) and light-jets are summarized in the following tables

Data set	t-jets	light-jets
data_zp_t	86%	14%
data_pp_t	72.5%	27.5%

Data set	W-jets	light-jets
data_zp_w	48%	52%
data_pp_w	42%	58%

Data set	t-jets	light-jets
data_zp_t	129 282	21 555
data_pp_t	127 029	48 000

Data set	W-jets	light-jets
data_zp_w	110 735	121 658
data_pp_w	180 169	251 422

Variables defined and used for each jet in the classification are as follows

event	$\Delta R(J,W)$	$\Delta R(J,t)$	p_T	η	ϕ	τ_{32}	τ_{21}	m	label
0	0.693589	0.280779	271.076000	-0.205725	1.034350	0.641589	0.304973	70.244600	l
0	1.152290	0.542026	161.364000	1.779510	-2.046550	0.678087	0.529191	67.632400	l
0	0.505954	0.876577	88.041000	0.431132	0.073586	0.468017	0.631805	7.432140	l
1	0.172936	0.046981	367.557000	-1.193480	-1.722920	0.840838	0.283345	75.302100	w
1	0.031584	0.143634	329.300000	-0.109191	1.337560	0.618819	0.205733	75.042200	w
2	0.143172	0.050171	501.473000	0.596318	-0.276567	0.605931	0.370552	171.372000	t

Table 2: Defined variables for each jet

Results

Performance of ML algorithms
For training and testing the respective algorithms, we used different sets. The algorithms for the prediction of t-jets were trained, after applications of undersampling methods, on a part of the data set data_zp_t and tested on the rest of data_zp_t and data_pp_t. The algorithms for the prediction of W-jets were trained on a part of the data set data_pp_w and tested on the rest of data_pp_w and data_zp_w.The performance of classifiers is shown via ROC curves derived based on test samples in Figure 3.

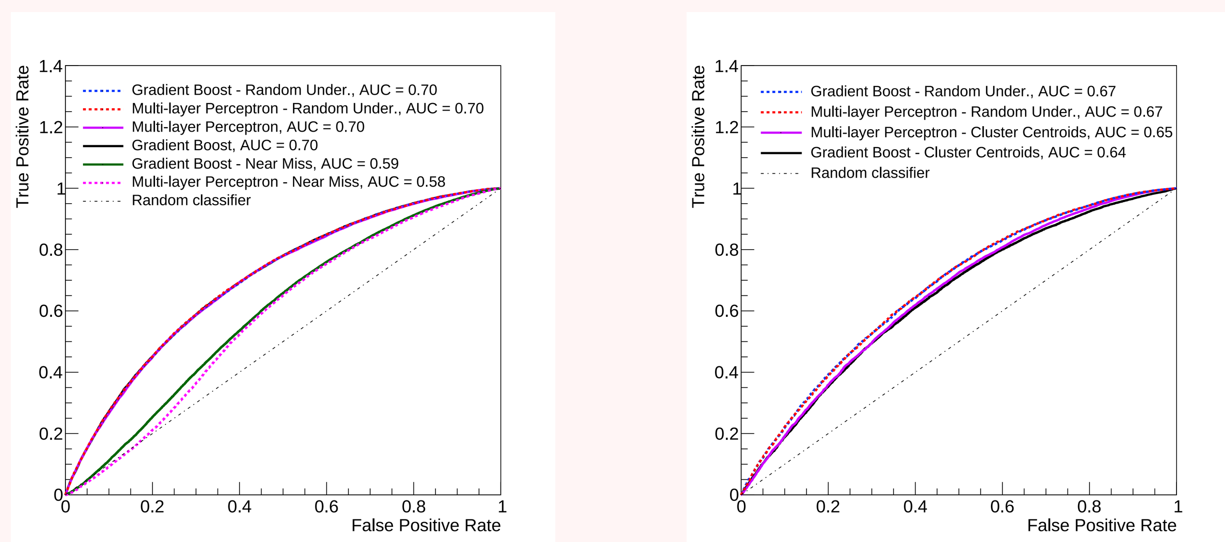

Figure 3: ROC curves summerising performance of W-tagging classifiers (left) and t-tagging classifiers (right)

Comparison of best ML method and cut-based algorithm In the Figure 4 we can see top tagging real efficiencies (red) and mistagging rates (blue) using cut-based (dashed lines) and ML-based (solid lines) of BSM $t\bar{t}y_0 \to t\bar{t}t\bar{t}$ as a function of jet mass (right). We can see that ML based algorithms give the same real efficiencies as cut-based, but significantly less fake efficiencies. Where real and fake efficiencies are defined as

$$\epsilon_{\text{real}} = \frac{\text{N(tagged \& matched)}}{\text{N(tagged \& matched)} + \text{N(not} - \text{tagged \& matched)}}$$

$$\epsilon_{\text{fake}} = \frac{\text{N(tagged \& not} - \text{matched)}}{\text{N(tagged \& not} - \text{matched)} + \text{N(not} - \text{tagged \& not} - \text{matched)}}$$

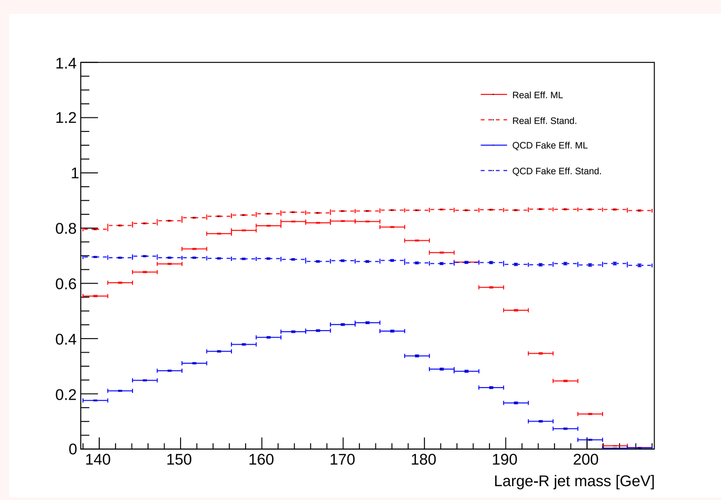

Figure 4: Efficiencies using cut-based and ML

References

[1] Franz Gross et al. 50 Years of Quantum Chromodynamics. December 2022.

[2] Inderjeet Mani and Jianping Zhang. Knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, 2003.

[3] Medium.com. Gradient boosting classifier, 2023. https://medium.com-/@hemashreekilari9/understanding-gradient-boosting-632939b98764.

[4] Andreas G. Müller and Sarah Guido. *Introduction to Machine Learning with Python*. O'Reilly, Beijing Boston Farnham Sebastopol Tokyo, 2016.

[5] Jane Yen and Yue-Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, April 2009.

Conclusions

The real efficiencies of cut-based method in both t-jets and W-jets tagging are high about 80%, mostly flat, but unfortunately also having high mistagging rates about 65-70%. While ML-based method has lower efficiencies, the mistagging rates are suppresed compared to cut-based method

Acknowledgements

The author would like to thank the grants of MSMT, Czech Republic, GAČR 23-07110S for the support.