CS229. Fall 2016.
Final Project Report.

Quark-gluon tagging in the forward region of ATLAS at the LHC.

Rob Mina (robmina), Randy White (whiteran)

December 15, 2016

1 Introduction

In modern hadron colliders, such as the Large Hadron Collider (LHC) at CERN, sufficiently high energies are reached that the properties of the elementary constituents of matter can be probed. Particle collisions in the center of the ATLAS detector cause the creation of observable high-energy particles, of which quarks and gluons comprise one of the largest physically "interesting" components. They deposit large signals, called "jets", within the surrounding detector subsystems.

On a collision-by-collision basis, the classification of each jet by the type of particle that produced it is called "tagging". The relative locations and sizes of hits within a jet are correlated with its type in complex ways that are only partially understood physically.

In the "forward" region of the detector, particles traveling with small angles relative to the beam axis are observed. Due to budgetary and mechanical constraints, the physical layout of the detector is different in this region compared to the "central" region; in particular, no ultra-fine granularity tracking information is available to resolve hits very near the interaction site, and the angular granularity of the detector is coarser. The combination of these and other detector effects causes jet classification algorithms that perform well in the central region to fail in the forward region.

We compare methods for jet classification for simulated high-energy collisions in the central and forward regions at the upgraded Run III ATLAS detector. Construction will begin on the detector upgrades in 2018, so it is an ideal time to develop and evaluate improved algorithms using simulation. In particular, we explore geometric and observed variables in order to optimize the performance of algorithms such as logistic regression, support vector machines, and neural networks.

2 Related Work

Jet tagging is the subject of much current research. Efforts range from defining new features, such as the EEC discussed later in this note [1], to applying traditional multivariate learning techniques to simulated interactions [2], and finally to applying such techniques to actual data observed by ATLAS and other experiments [3] [4]. These efforts have proven effective in the central region, where the detectors have the best spatial and energy resolution, but they have not been applied to the less well-understood forward region.

More cutting-edge efforts include applying convolutional neural networks and other deep learning techniques to simulated jets in order to explore performance at hypothetical next-generation collider experiments [5].

3 Dataset description

A sample of ATLAS Run III Upgrade simulated proton-proton collisions was selected to provide the training and testing dataset for this task. The simulated event energy and luminosity conditions reflect those expected for the "HL-LHC" running period post-2018. The sample consists of about 25,000 "dijet" events with two jets of approximately equal momentum oriented opposite one another. Such events are advantageous because the signal jets tend to have much higher energy than the background "pile-up" interactions. In order to avoid contamination by this uninteresting background, only the highest energy jet in each event is considered. This yielded a sample of 17,023 gluon jets and 7,835 quark jets.

A full simulated event record is about 7 MB in size, and contains a great deal of extraneous information unrelated to the jets of interest. Therefore, a simple data structure called a ROOT tree was created with only the necessary jet variables [6], which was only about 16 MB for the entire sample.

The variable $|\eta|$ characterizes the angle of a jet with respect to the beam. Since the detector sensitivity is not constant with varying angle, the data sample has been binned according to $|\eta|$, and tagging discriminants will be developed indepently for each bin. The choice of bins, along with the fraction of the total quark and gluon samples in each bin, are shown in Table 1. The $|\eta|$ distributions for both quark and gluon jets in this sample are shown in Fig. 1.

Region	$\|\eta\|$ bounds	q-jet fraction	g-jet fraction
R1	$\|\eta\| < 1$	0.404	0.567
R2	$1 \leq \|\eta\| < 2$	0.296	0.302
R3	$2 \leq \|\eta\| < 2.8$	0.192	0.105
R4	$2.8 \leq \|\eta\| < 3.2$	0.0688	0.0177
R5	$3.2 \leq \|\eta\| < 4$	0.0391	0.0078

Table 1: The five bins in $|\eta|$, as well as the fraction of total quark and gluon jet samples present in each bin. Note that gluon jets are more central than quark jets, in the sense that they tend to have larger angle with respect to the beam axis. One difficulty of building powerful jet classifiers for the forward region is that there are fewer samples to train on compared to the central region.

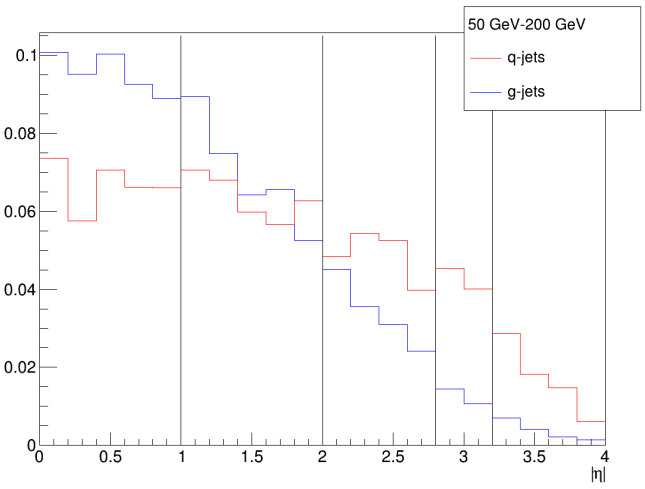

Figure 1: $|\eta|$ distributions for quark and gluon jets in the sample, normalized to 1. The vertical black lines delineate boundaries between different $|\eta|$ bins.

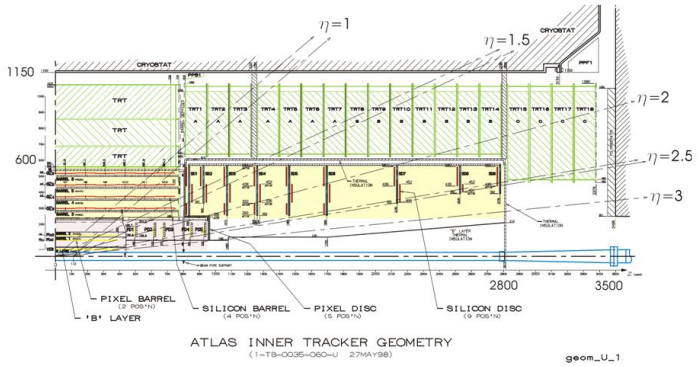

Figure 2: A schematic of the ATLAS inner detector, showing representative values of $|\eta|$.

4 Features

Most discriminating variables effective for quark-gluon tagging relate to the multiplicity of particles within the jet, which can be estimated using either the number of charged-particle tracks or the number of topological calorimeter clusters observed within the jet radius [3]. Variables related to the correlations between jet constituent directions and energies are also possible to exploit [1].

Discriminating variables extracted from the described dataset are listed in Table 2, along with two quantities characterizing their discriminating power. Denote the effiency for selecting quark jets as eff_q, and that for gluon jets as eff_g. This is simply the ratio of the number of jets selected by a discriminant over the total number in the sample. For a discriminant that uses only one of the variables, there is a simple functional relationship between the two efficiencies, so the gluon efficiency can be written as a function of quark efficiency: $\text{eff}_g = \text{eff}_g(\text{eff}_q)$. The curve obtained by this relationship is called a ROC curve, as for example in Fig. 3. The first quantity is $\text{eff}_g(0.5)$, the efficiency for selecting gluons when the quark selection efficiency is 50%. The second quantity is the separation achieved by the variable over all possible values, which is obtained by integrating the quantity $\text{eff}_q - \text{eff}_g(\text{eff}_q)$ from 0 to 1.

Some of the extracted variables use truth information from the simulation that is not available for actual data, while others (reco) use only information obtained from the event reconstruction. Truth variables perform better than reco variables in all regions, and the best performing variable is the EEC defined in [3]. The best-performing reco variable differs between the central region, where tracking is most precise, and forwards regions, where the calorimeter is more precise.

5 Methods

Because of the small number of examples in the forward region, we chose to implement learning algorithms that could be trained effectively using such limited statistics. We also limited the input number of features by using only two or three of the discriminating variables discussed in the previous section. This is possible because two features that both contain information about multiplicity (N), are highly correlated and do not provide as much information as one multiplicity and one width feature (W). Given a sufficient number of examples, it would in principle be possible to achieve better results by combining multiplicity or width variables from the tracker and the calorimeter, which are independent, but this was not pursued in this project.

Variable	R1: g-jet eff at 50%	R1: integrated separation	R4: g-jet eff at 50%	R4: int. sep.
$N_{trk,reco1000}$	0.161	0.256	0.476	0.024
$N_{trk,reco500}$	**0.157**	**0.259**	0.476	0.023
$N_{trk,truth}$	0.126	0.283	0.134	0.292
$N_{90\%constit,reco}$	0.191	0.220	0.352	0.090
$N_{constit,truth}$	0.102	**0.299**	0.123	**0.322**
$W_{trk,truth}$	0.152	0.213	0.129	0.244
$W_{calo,reco}$	0.204	0.207	**0.335**	**0.130**
$W_{calo,truth}$	0.116	0.236	0.118	0.252
EEC_{reco}	0.191	0.215	0.346	0.116
EEC_{truth}	**0.077**	0.294	**0.075**	0.306

Table 2: Discriminating variables used in this study, along with two quantities characterizing their effectiveness described above. For the g-jet efficiency at 50% working point, smaller is better. For the integrated separation, larger is better. In each column, the best-performing variable based on truth information is EEC (bolded), though the best-performing reconstruction variable is different in the forward region compared to the central region (also bolded). These variables are described above.

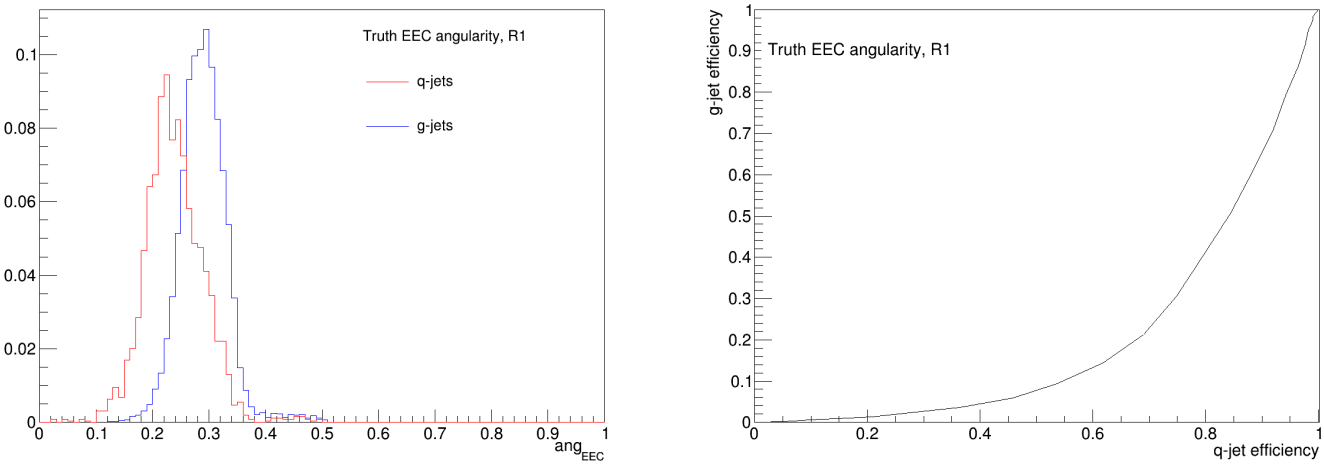

Figure 3: The track-based energy-energy-correlation angularity ([3]) distributions for the R1 region, along with the ROC curve obtained by a thresholded decision stump on this variable.

Each feature was scaled (independently for each $|\eta|$ region) to have mean zero and standard deviation 1.

The algorithms chosen were logistic regression (LR) using the "L2" norm as penalty with an intercept term, a support vector machine (SVM) with various kernel functions described below, and a sequential dense neural network (NN) with one or two hidden layers using the rectified linear activation function. For the SVM and NN, the prediction error on the training sample was used as the objective function.

In LR, a weight vector θ of length $n+1$, where n is the number of features, is optimized so as to maximize the logistic likelihood function $h_\theta(x)^y(1-h_\theta(x))^{1-y}$, where $h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)}$, x is an example feature vector, and y is the correct label. A regularization term is added to the likelihood function to penalize weight vectors with very large values. In particular, the Newton-Raphson method was used to perform this optimization via the scikit-learn package [7].

An SVM is capable of optimizing a linear decision boundary with dimensionality equal to the number of examples (a "separating hyperplane"). A nonlinear decision boundary in the feature space can be obtained by using a kernel function. The SVM maximizes the gap between examples of the different classes, where the shape of the gap is determined by the choice of kernel function. Predictions for new examples are then made by observing on which side of the separating plane they fall. The scikit-learn "SVC" class was used to perform the optimization.

Finally, NN's are capable of optimizing an arbitrary non-linear decision boundary in any dimension. The individual cells in the network contain weight and bias terms that are optimized with respect to some objective function evaluated on the output of the network, most often using an efficient algorithm called backpropagation in conjunction with gradient descent. Backpropagation is an application of the chain rule that allows computation of the gradient at each layer in the network, iterating "backwards" from the outward layer towards the input layer. The scikit-learn multi-layer perceptron classification class was used to perform the learning.

6 Results

Since LR is the most efficient of the three algorithms to train, we evaluated the best combination of variables to use in each $|\eta|$ region by using all possible combinations of $N+W$, $N+EEC$, and $W+EEC$ features as input to the LR classifier. We also assessed combinations of three features: one multiplicity, one width, and one correlation function. We did not evaluate performance for combinations of reconstruction variables with truth information, since such combinations do not give reproducible results. The hyperparameters for the LR algorithm are the convergence tolerance and the regularization strength, and we determined that the default values of 0.0001 and 1.0, respectively, gave the best performance across the regions and feauture combinations.

Our three metrics for evaluating classification performance were the prediction accuracy on the test set (larger is better), the gluon efficiency at 50% quark efficiency (smaller is better), and the integrated separation described in the features section above (larger). Computing the latter two metrics requires computing a one-dimensional discriminant from the classifier, which was done using the class probabilities evaluated on the test dataset. Usually, the feature combination that performs best on one of these metrics also optimizes the others.

After determining the best two- and three-feature combinations for each region, an SVM was trained on each of these. We examined the performance of linear and polynomial kernel functions of degree 3-5, and determined that these were not superior to the default radial basis function. We also required the model to train class probabilities in addition to optimizing the decision boundary, so as to compute the same performance metrics.

We determined that the optimal choice of number of layers for the NN so as to optimize accuracy without overfitting was two, with 10 nodes in the first layer and 5 in the second. The scikit-learn implementation of the NN classifier uses the cross-entropy loss function, and we chose the 'lbfgs' solver which uses a form of backpropagation to optimize node weights. We determined that the rectified linear unit function $f(x) = \max(0, x)$, was the optimal choice for activation function. Since the NN is non-linear in nature, we added the transverse momentum of the jet, p_T, and the jet charge to the inputs. These variables are not discriminative in nature, but do contain correlations with the discriminative variables that could be exploited by a nonlinear classifier.

The performance of each of these classifiers in all $|\eta|$ regions can be seen in Table 3. Only the performance for reconstruction variables is shown for brevity. In all cases, the multivariate classifiers outperform the one-feature classifiers studied above.

7 Conclusion

The LR and NN algorithms achieve similar results, though the SVM did not perform as well. More surprisingly, the performance in the far-forward region R5 was not significantly degraded compared to the central region. This implies that fairly pure and efficient quark

Region	Variables	Logistic Regression	SVM classifier	NN classifier
R1	$N_{trk,reco500} + W_{calo,reco}$	**0.81**, **0.102**, **0.29**	0.80, 0.122, 0.22	0.80, 0.118, 0.28
R1	$N_{trk,reco500} + W_{calo,reco} + EEC_{reco}$	0.81, **0.092**, 0.29	0.80, 0.122, 0.22	**0.81**, 0.093, **0.31**
R2	$N_{trk,reco1000} + W_{calo,reco}$	0.77, 0.121, 0.27	0.77, 0.137, 0.21	**0.79**, **0.101**, **0.29**
R2	$N_{trk,reco1000} + W_{calo,reco} + EEC_{reco}$	0.77, 0.122, 0.28	0.77, 0.130, 0.21	**0.79**, **0.084**, **0.29**
R3	$W_{calo,reco} + EEC_{reco}$	**0.71**, **0.155**, **0.23**	0.70, 0.173, 0.22	0.66, 0.253, 0.20
R3	$N_{90\%constit,reco} + W_{calo,reco} + EEC_{reco}$	**0.71**, **0.155**, **0.23**	0.70, 0.175, 0.22	0.66, 0.253, 0.20
R4	$N_{90\%constit,reco} + W_{calo,reco}$	0.61, **0.355**, **0.11**	0.58, 0.402, 0.04	**0.64**, 0.388, 0.09
R4	$N_{trk,reco500} + W_{calo,reco} + EEC_{reco}$	**0.63**, **0.266**, **0.16**	0.58, 0.402, 0.05	0.60, 0.372, 0.08
R5	$N_{trk,reco500} + W_{calo,reco}$	0.62, 0.185, 0.12	0.71, 0.332, 0.15	**0.74**, **0.148**, **0.24**
R5	$N_{trk,reco500} + W_{calo,reco} + EEC_{reco}$	0.64, 0.216, 0.12	0.71, 0.246, 0.22	**0.74**, **0.148**, **0.24**

Table 3: Performance for different classifier algorithms in each $|\eta|$ region. In each column, the first number is the prediction accuracy, the second is the gluon efficiency at 50% quark efficiency, and the third is the integrated separation. The optimal value for each |eta| region and number of features is bolded.

jet samples will be available all the way to $|\eta| = 4.0$, enabling new physics searches at the upgraded ATLAS detector. Future efforts should focus on acquiring much larger datasets in order to exploit classifiers with more input features and more sophisticated structures. Increasing the number of examples by an order of magnitude would probably be sufficient to more carefully explore the differences in performance between these algorithms and combinations of variables.

References

[1] Andrew J. Larkoski, Gavin P. Salam, and Jesse Thaler. Energy Correlation Functions for Jet Substructure. *JHEP*, 06:108, 2013.

[2] Jason Gallicchio and Matthew D. Schwartz. Quark and Gluon Tagging at the LHC. *Phys. Rev. Lett.*, 107:172001, 2011.

[3] Georges Aad et al. Light-quark and gluon jet discrimination in *pp* collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Eur. Phys. J.*, C74(8):3023, 2014.

[4] Tom Cornelis. Quark-gluon Jet Discrimination At CMS. In *Proceedings, 2nd Conference on Large Hadron Collider Physics Conference (LHCP 2014): New York, USA, June 2-7, 2014*, 2014.

[5] Patrick T. Komiske, Eric M. Metodiev, and Matthew D. Schwartz. Deep learning in color: towards automated quark/gluon jet discrimination. 2016.

[6] Rene Brun and Fons Rademakers. Root – an object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(1):81 – 86, 1997.

[7] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. *CoRR*, abs/1309.0238, 2013.