Characterization of Jet Charge at the LHC

Thomas Dylan Rueter, Krishna Soni

Abstract

The Large Hadron Collider (LHC) produces a staggering amount of data - about 30 petabytes annually. One of the largest difficulties in analyzing this data is the analysis of hadronic ”jets” of particles produced by quantum chromodynamic processes. Due to the thresholds that particles must exceed in order to have their data stored, it is not possible to simply add the charges of every particle in a jet to calculate its charge. In this paper, we train a support vector machine (SVM) to classify jet charges for up and anti-up quark jets at a better efficiency and rejection rate than can be achieved with the traditional jet charge calculation and cut method. Through the selection of good, physically motivated features and careful regularization, our SVM does slightly outperform the traditional method, as evidenced by a comparison of ROC curves for both methods.

I. Introduction

The Large Hadron Collider (LHC) is the largest and most powerful particle accelerator in the world, built with the goals of understanding and testing the Standard Model of particle physics, as well as searching for physics beyond the Standard Model. Having already discovered the Higgs boson in Run I, the LHC has been upgraded to a higher center of mass energy and has begun a second round of data collection, dubbed Run II. One of the largest difficulties in analyzing the data produced by the LHC is the analysis of hadronic ”jets” of particles produced by quantum chromodynamic (QCD) processes. Being able to characterize these jets to determine their sources is crucial to both testing the Standard Model as well as understanding any new physics the LHC might find.

This problem is made difficult by the incomplete nature of the data; only particles above certain momentum thresholds are recorded by the ATLAS detector. If every particle were recorded, calculating the jet charge would be a simple matter of adding the charges of all ofthe particles in a given jet. Furthermore, the physics of jets is not well understood analytically due to both the wide variety in jet structure (e.g. the number of particles in a jet can vary wildly) and the difficult nature of QCD calculations (it is strongly coupled, and including many final state particles exponentially complicates the calculation). As a result, physicists have tried to use heuristic measures to obtain information about jet charge, and the goal of this paper is to apply machine learning algorithms and techniques to the problem based on physical intuition to improve on these measures.

II. Related Work

Determining the charge of these jets has been done in the past via a summation of the particle tracks within the jet weighted by their transverse momentum in the detector (p_T) [3]:

$$Q_j = \frac{1}{(p_{T_j})^\kappa} \sum_{i \in Tr} q_i \times (p_{T_i})^\kappa \qquad (1)$$

Here Tr represents the set of particle tracks recorded for the jet, and κ is a number which defines the weighting, $0 \leq \kappa \leq 1$. Jets are classified as positively or negatively charged based on whether their calculated charge is above or below a given threshold [3] [2]. The efficacy of this method certainly hints that perhaps momentum weighted charge averages are a very relevant feature for jet charge classification, but there is a large overlap between the positive and negative jet charge distributions, as seen in Figure 1. We believe a larger set of features could provide better separation of the two classes.

Other recent work in jet classification has focused on computer vision, using image analysis tools on pictures of jets constructed from detector information [1]. This inclusion of position information, specifically how far a given particle is from the jet's center, seems like an important feature for charge classification. Studies have also used neural networks and multivariate discriminant analyses for jet tagging with considerable success [4]. Support vector machines (SVMs) have been applied to similar jet classification problems and been successful [5]. We believe that the two-class nature of this problem lends itself particularly to an analysis based on a SVM approach, as the SVM algorithm is capable of taking in large feature sets and finding separating hyperplanes which may inform better feature selection in future calculations.

III. Dataset and Features

Our initial training set consists of a list of 9576 simulated up (+2/3 charge) and anti-up (-2/3 charge) quark jet events with at least five particles each, which were generated by the ATLAS collaboration using a particle simulation software called **PYTHIA**. For every jet in an event there is aggregate information about the jet's p_T, η, ϕ, calculated momentum-weighted jet charge for κ of 0.3, 0.5, and 0.7, and the true value of the jet charge (negative or positive). η is the pseudorapidity of the jet, which represents the angle of the jet's center relative to the beam axis, and ϕ is the azimuthal angle of the jet's center in the detector. For each jet there is a list of the information stored for each particle within jet, containing the particle's charge, p_T, η, ϕ, and dR to jet axis. Here dR is defined by:

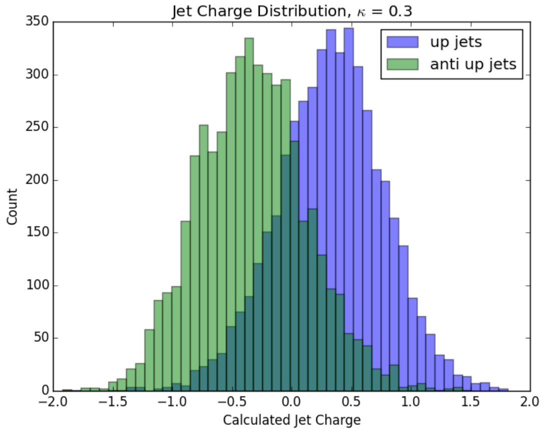

Fig. 1. Calculated jet charges for positive (up) and negative (anti-up) jets in our training data, using κ = 0.3.

$$dR = \sqrt{(\eta_{particle} - \eta_{jet})^2 + (\phi_{particle} - \phi_{jet})^2} \tag{2}$$

Based on the idea that particles which carry a larger fraction of the jet p_T carry more information about the jet charge, we defined our initial features as the fraction of the jet p_T, the charge, and the dR values of the five particles with the largest momenta. This raw feature selection seemed to over-fit our data, as we had a training error of 18.2% compared with a test error of 48%. We believe this is due to the SVM algorithm not being scale invariant, as the p_T values of our leading particles take on values from 1 GeV to 100 GeV. In order to normalize these values to the interval [0,1], we considered instead the fraction of the jet momentum z_i accounted for by the i^{th} track:

$$z_i = \frac{p_{T_i}}{p_{T_{jet}}} \tag{3}$$

We also thought that perhaps another reason for our poor performance was the incomplete nature of the information contained within the first five particles. The traditional jet charge calculations sum over every particle in a jet, which provides much more information than just the particles with leading p_T. In an attempt to use information from the whole jet, we defined features which involved summing over all of the particles. We defined our features as:

$$Q_{1,\kappa} = \sum_{i \in Tr} q_i z_i^{\kappa} \tag{4}$$

$$Q_{2,\kappa} = \sum_{i \in Tr, z_i > 0.1} q_i z_i^{\kappa} \tag{5}$$

$$Q_{3,\kappa} = \frac{\sum\limits_{i \in Tr} q_i \left|\Delta\eta_i\right|^{\kappa}}{\sum\limits_{i \in Tr} \left|\Delta\eta_i\right|^{\kappa}} \tag{6}$$

$$Q_{4,\kappa} = \frac{\sum\limits_{i \in Tr, z_i > 0.1} q_i \left|\Delta\eta_i\right|^{\kappa}}{\sum\limits_{i \in Tr} \left|\Delta\eta_i\right|^{\kappa}} \tag{7}$$

Each of these was calculated for κ = 0, 0.3, 0.5, 0.7, and 1, giving us 20 total features. The $\Delta\eta$ weighted averages were used in addition to momentum weighted averages to get a measure of particles' rapidity relative to the jet's rapidity (note rapidity can be approximated by the pseudorapidity η for these light quark masses) linked with the charge. The sums over tracks with $z_i > 0.1$ are designed to pick out the particles with the highest momentum, which there is good reason to believe carry more information about the jet charge.

IV. Methods

Given that our data was not linearly separable, we were drawn to using an RBF Kernel to parametrize a C-Support Vector Classification function (**sklearn.svm.SVC()**) to perform our classification.

A. Support Vector Machines

Support Vector Machines (SVMs) are commonly used in machine learning as binary classifiers. We aim to calculate a maximum margin hyperplane between classes of elements from our training set. In this case, we chose to use an SVM to classify a sufficiently interpret if a jet's charge was negative or positive. We may formalize our problem by constraining:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \begin{cases} y_i = +1 \\ y_i = -1 \end{cases} \tag{8}$$

Where y_i corresponds to either positive or negative jet charge. We define the dual optimization problem:

$$max_{\alpha_i \geq 0} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \left\langle x^{(i)} \cdot x^{(j)} \right\rangle \tag{9}$$

subject to the constraint $\sum\limits_{i=1}^{m} \alpha_i y^{(i)} = 0$. We may define a solution where $w = \sum_i \alpha_i y_i x_i$ which is a linear combination of our training vectors (This is found by applying the stationarity KKT condition on our dual optimization problem). We also define a penalty function (where C is a penalty parameter for misclassifying a sample) and consider

$$|w|^2 + C \sum_i \zeta_i \tag{10}$$

which is subject to:

$$y_i(x_i \cdot w + b) \geq 1 - \zeta_i \text{ where } \zeta_i \geq 0 \tag{11}$$

Here, our goal in the C-SVM, is to minimize $|w|^2$ (where $|w|$ is the norm vector to the separating hyperplane) in our penalty function. In this problem we arbitrarily set $C = 1$ since we did not have much information about how our feature data correlated to our training labels.

B. RBF Kernel

Now, it is useful to define some function which maps elements on some higher dimensional space in order to make them linearly separable (and thus usable for our SVM). Here, we define some function (say $\phi(x)$) which maps our features onto some higher dimensional space. From this, we define a Kernel Function which is the composition of an inner product of our feature mapping functions of $\phi(x)$:

$$\mathbf{K}(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{12}$$

In our example we used an RBF Kernel, or a (Radial Basis function defined as:

$$\mathbf{K}(x_i, x_j) = e^{-\gamma |x_i - x_j|^2} \tag{13}$$

The above equation can easily be seen as form of Gaussian distribution. Here, we leverage the properties of this distribution (which by definition is linearly separable) to improve the separability of our feature mapping. In addition, γ is a weighting parameter which defines the relative influence of a given training features. For this problem, we used a $\gamma = \frac{1}{N}$ where N is our number of training features. We made this assumption as we wanted our γ to be agnostic of the relative importance of our training features.

V. Results and Discussion

Feature selection and regularization was the most difficult part of this project, due to the lack of analytic solutions to jets. When our initial feature set of the p_T, dR, and charge of the 5 leading p_T particles was overfitting, we attempted to solve the problem by normalizing the p_T values as well as performing a dimensional reduction on the feature set via Primary Component Analysis. The SVM trained on these reduced features performed better, but was still worse than the jet charge calculation. We also briefly tried a boosted decision tree, using the AdaBoost algorithm and weak decision tree classifiers, but the test error was unaffected as the training error decreased. It was only once we implemented the feature set in section 3 (without PCA) that we achieved good classification on the up/anti-up quarks, as evidenced by the ROC curve calculated on the up/anti-up training set using cross-validation. Since we had a fairly large amount of data we used hold out cross-validation, training on 80

Figure 2 shows ROC curves for the jet charge calculation in Equation 1 for differing values of κ, as calculated on our training set. The ROC curve traces out the positive jet tagging efficiency, defined as the number of positive jets above the threshold divided by the total number of positive jets, and negative jet rejection, defined as the number of negatively charged jets divided by the number of false positives, as we move the charge threshold [2].

A perfect jet classification scheme would have a flat ROC curve at the maximum negative rejection. Our SVM ROC curve was calculated by scanning through thresholds on the SVM scores (calculated using the **sklearn.svm.SVC.decision_function** tool) of our test set. Our SVM slightly outperformed the traditional jet charge calculation for positive tagging efficiencies above .3, which is where ATLAS searches use these curves most often, thus, making a useful tool for tagging jet charge.

We also received a list of 9686 down (-1/3 charge) and anti-down (+1/3 charge) quark

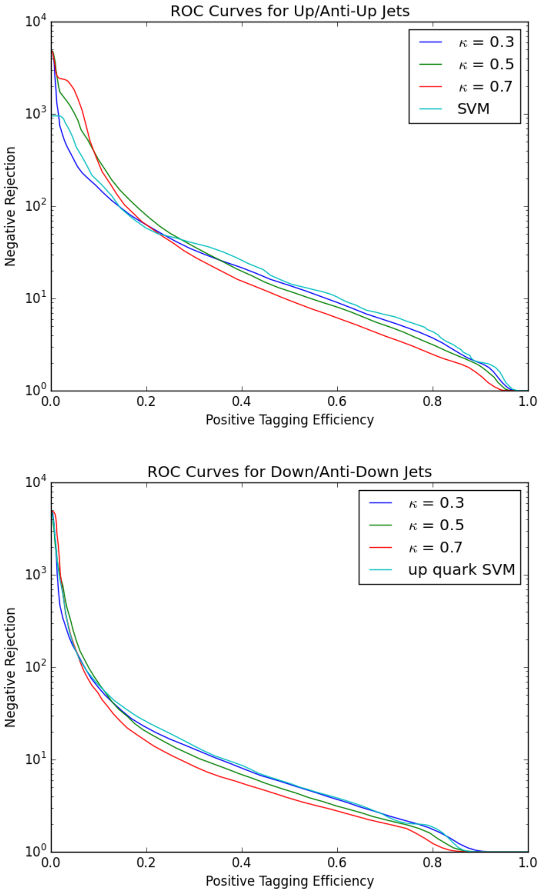

Fig. 2. **Top:** ROC curves for traditional jet charge calculation on up/anti-up jets (κ values) and our SVM, tested using hold out cross-validation. **Bottom:** ROC curves for jet charge calculation for down/anti-down jets (κ values) and our SVM, trained on the full up/anti-up jet data and tested on the down/anti-down jet data.

jet events, which we used to test our SVM after training it on the entire up/anti-up jet data set. Figure 2 shows that our SVM performed essentially the same as the best jet charge calculation for all positive tagging efficiencies. The down/anti-down data is much more difficult to classify, due to the total jet charge being a factor of 2 smaller than the up/anti-up jets. This causes a larger overlap of positive and negative jets around 0 calculated jet charge,as seen in Figure 3.

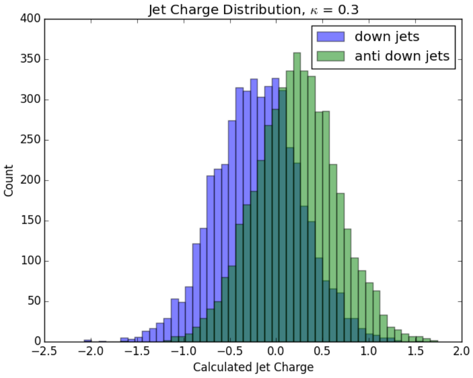

Fig. 3. Calculated jet charge with $\kappa = 0.3$ for the down/anti-down jet data.

VI. Conclusions and Future Work

We successfully classified both up/anti-up and down/anti-down jets using a SVM-based analysis and good feature selection. It became apparent to us that this classification problem is not only highly non-linear, but also cannot be adequately solved based on sampling only a few of the jet's particles. Had we received the down/anti-down data earlier, it would have been interesting to try to use a support vector regression algorithm to accurately determine whether a jet was up, anti-up, down, or anti-down based on a regression on the charge. Using the position information of the particles in the jet to analyze the jet as an image with a convolutional neural network would be another interesting method of trying to classify jets as positive or negative.

References

[1] Josh Cogan, Michael Kagan, Emanuel Strauss, and Ariel Schwarztman. Jet-images: computer vision inspired techniques for jet tagging. *Journal of High Energy Physics*, 2015(2):1–16, 2015.

[2] ATLAS collaboration et al. Jet charge studies with the atlas detector using $\sqrt{s}$= 8 tev proton-proton collision data. ATLAS-CONF-2013-086, 2013.

[3] David Krohn, Matthew D Schwartz, Tongyan Lin, and Wouter J Waalewijn. Jet charge at the lhc. *Physical review letters*, 110(21):212001, 2013.

[4] Mostafa Mjahed. Identification of the quark jet charge in $e^+e^- \rightarrow w^+w^-$ using neural networks and discriminant analysis methods. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 449(3):602–608, 2000.

[5] Federico Sforza, Vittorio Lippi, Giorgio Chiarelli, and Sandra Leone. Rejection of multi-jet background in a hadron collider environment through a svm classifier. In *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2011 IEEE*, pages 1404–1408. IEEE, 2011.