Applying Machine Learning Techniques for Jet Flavor Identification

Eric Culbertson
Unviversity of Virginia
emc5ud@virignia.edu

Abstract

The field of experimental particle physics aims to uncover the fundamental properties of the universe through measuring the interactions of the most basic particles that make up matter. Of these fundamental particles, my research involves the study of quarks, which are the building blocks of matter. Quarks cannot exist in isolation, and any quark that we produce in our experiment will produce an associated shower of particles known as a jet in the same direction of the quark. The focus of this study is to create a model that can identify a quark by the properties of the measured shower of particles that it produces. In order to allow for supervised learning, the model was trained on simulated events meant to emulate data collected at the Large Hadron Collider.

1 Introduction

1.1 The CMS Experiment

My physics research is conducted as a part of the CMS experiment at the European Organization for Nuclear Research (CERN). The CMS experiment is one of four major experiments conducted at CERN's large hadron collider (LHC). The LHC is one of the largest machines created by man, which can accelerate protons to speeds exceptionally close to the speed of light. The collider is shaped like a large (27km circumference) doughnut, and it directs and accelerates protons through its circular vacuum chamber using a combination of superconducting magnets and radiofrequency chambers ("Accelerators", 2017). During run-time, millions of protons are accelerated in both direction around the ring, and are carefully guided so that they only collide at four points at regular time intervals.

The compact muon solenoid (CMS) is a detector located at one of these collision points. Because these collisions are at high energy, the collisions produce millions of fundamental particles going in every direction. The properties of these high energy fundamental particles can lead physicists to important discoveries, such as the detection of the Higgs Boson in 2012. In order to analyze the massive amount of data produced at the CMS experiment, there are over 4,000 collaborating scientists and students across the globe. ("About CMS", 2017)

Because of the CMS detector's ability to measure the properties of microscopic particles, it can be considered as one of the most advanced microscopes in the world. The detector is able to see the particles produced in an interaction by measuring the energy that the particles deposit in each of the CMS detector's thousand scintillating crystals.

Scintillation is simply a process in which a detecting material gives off light which then can be converted into an electric signal. Because the particles flow outward at such a rapid pace, determining what types of particles produce each energy deposit is an extremely complicated task requiring many stages of physics analysis. My goal for this project is to add another tool that physicists can use to help discriminate fundamental particles ("CMS").

1.2 Physics Background

The culmination of our progress in many avenues of physics is summarized in something called the standard model. This model aims to describe the properties of every fundamental particle and how those particles interact with each other. According to the standard model, there are 17 fundamental particles that make up all of the known matter in the universe and that mediate the forces between them. For the scope of this paper, only talk about the particles that are relevant to jet production will

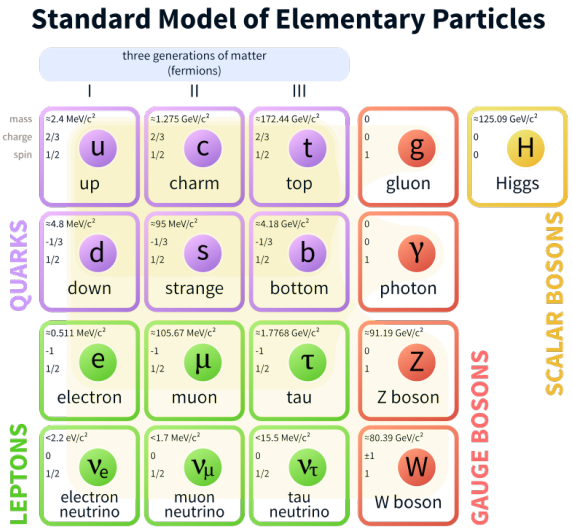

Figure 1: The fundamental particles

be discussed.

The particles that compose matter can be broken up into two basic groups called quarks and leptons. Each group is composed of exactly 6 particles, which are related in groups of two known as generations. The biggest difference between leptons and quarks is that leptons can exist on their own while quarks cannot exist separate from other quarks. For example, up and down quarks belong to the same generation and they are most commonly grouped together to create most every-day matter found in the universe, such as protons and neutrons.

The reason that leptons such as electrons can exist independently while quarks must exist in groupings is because of the strong force. The strong force, mediated by a particle called the gluon, is an exceptionally powerful attractive force between particles that have what is known as a color charge. Confusingly, color in particle physics has nothing to with the color we can see. Instead, it is analogous to electric charge or mass. The more particles have, the stronger the force. As the name implies, the strong force distinguishes itself from other forces such as gravity and electromagnetism due to its magnitude. Also, the strong force is unique because its magnitude increases with distance, much like how pulling apart a metal spring gets harder and harder the more spread out it becomes (”Standard Model”).

The immense strength of the strong force combined with its unique scaling with distance leads to an interesting phenomenon known as confinement.

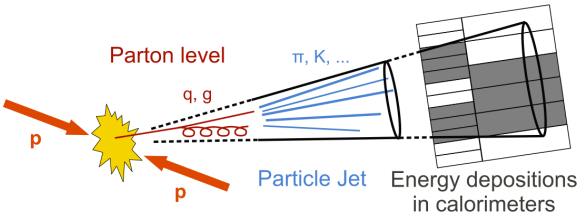

Figure 2: When high energy partons (quarks and gluons) are produced, a cone of additional particles is produced in their wake.

Essentially, as two quarks get pulled apart, more and more energy needs to be put into the system to separate them. Eventually, once the input energy has reached a certain threshold, this energy is converted into two new quarks which each combine with quarks that have been isolated. In physics, this process is an example of a system favoring a lower energy state. Once the energy put into the system becomes too high, it is more energetically favorable for some of that energy to be converted into two new particles.

Particle collisions at the LHC are an example of a process that can provide enough energy to rip quarks apart. Due to confinement, however, the energetic quarks will create a cluster of new particles in the direction of its path. These new particles create a cone of outgoing particles known as a jet, which is shown in figure 2. Because the gluon also interacts via the strong force, it too has the property of confinement.

Many important physics processes, such as the process that produces the Higgs Boson, have quarks as an end product. In order to learn more about these fundamental physics interactions, it is imperative that physicists know about the quarks that are produced. However, identifying the specific type of quark based on the properties of the measured jet is a difficult problem. In fact, currently the best methods of jet identification merely label it as one of two categories of quarks. The task of this paper was to improve jet classification in ideal cases by constructing a model which can categorize a jet into one of three classes of jet producing particles. The classifications, also known as flavors include light quarks (up, down, strange, charm), heavy quarks (bottom and top), and gluons.

2 Supervised Sample Generation

In order to have a sample that allows for supervised learning, the jets analyzed by my model were created using Monte-Carlo physics simulations. First the underlying processes that produce quarks needed to be simulated using a framework called MadGraph5_amc@NLO. This software can simulate the physics behind a proton-proton collision at the specified run time conditions and simulate the properties of what type of particles can come out. The output of this software is the kinematics of the outward going particles, some of which are quarks and gluons.

Once the quarks and gluons are made, they are fed into another piece of software that simulates the parton shower. The parton shower is the result of the containment described in the previous section. Because quarks and gluons cannot exist on their own, each particle creates hundreds of more in their wake. The output of the parton shower simulator PYTHIA8 is a file containing hundreds of particles per event. An event is simply a timestep in which one underlying particle process occurs. After the parton shower finishes, the hundreds of output particles in the event must be clustered into outward going cones. The anti-k_t algorithm was chosen to do this clustering, and the C++ library FastJet handled the implementation. The clustering algorithm's parameters were chosen to be in line with current best practice in physics analysis. The clustering algorithm was not tampered with because the model must learn to the associate quarks with the same type of jets used in real physics analysis.

After the clustering algorithm runs, there is now a list of particle jets that need to be associated back to the original partons (quarks and gluons) that could have produced them. The software that clusters the jets is agnostic to the types of particles that make up the cluster, however. That means in order to train our model using a supervised sample, the jets need to manually be paired up with the partons produced by Madgraph5 beforehand. This was done by comparing at the kinematics of the jets to the kinematics of the partons in each event, and matching the jets to the quark by looking for pairs that are close in both direction and energy. The required closeness in direction and energy was systematically adjusted in order to minimize the amount of mislabeled entries in the sample.

2.1 Preprocessing

In order to aid in the training of our model, there were a few steps taken to ensure that training was more effective. First, the attributes had to be normalized so that the machine learning algorithm implemented did not give that attribute significantly more importance. For features with clearly defined upper and lower bounds, the values were linearly scaled to be between zero and one. Most features however had long tails in their distributions. For these I chose to normalize the features so that the values had a mean of zero and a standard deviation of one.

The matching procedure described in the previous section was not perfect, and some work had to be done to measure the error rate of the correct labeling. Obviously if the label provided a supervised model is often incorrect, it will prevent the model from learning the true classifications. To get an estimate of the error in our matches, I determined the matching rate of each fundamental particle produced in our sample. This plot can be seen in figure 3. As expected, the matching rate of the gluon is lower than the rate of the other fundamental articles. This is because gluons often split into multiple jets, which the current matching procedure doesnt account for. The fact that the gluon matching rate is lower than the other particles is promising because it shows that the matching procedure doesnt purely produce false positives. By instead looking at particles which are known not to produce jets, we can estimate how often that a particle erroneously gets matched with a random unassociated jet. This was determined to be around 30 percent. The error rate was reduced to 20 percent by requiring the matches to be closer in energy and direction, but we were cautious about not making the criteria too strict. This is because that stricter matching criteria biases our sample to only produce higher energy particles, as lower energy particles are more likely to have a spread out jet. We finally settled on an angular distance ΔR^2 of .3 and a fractional energy difference of .1.

Another step done before the training involved the creation of new features to train with and the removal of others. An example of one such feature are the the momentum in the x-y direction, with the z-axis being the direction of the incident proton beam. This is because the direction in the x-y plane for all particles is essentially random. I didnt have such prior knowledge on all the fea-

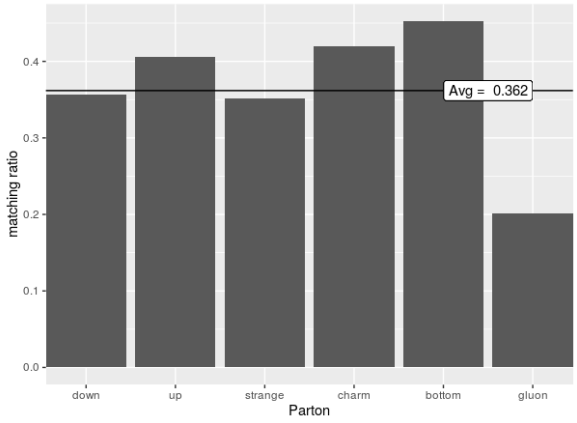

Figure 3: Of the particles produced, the fraction that match up with a jet is plotted. The gluons produced show a significantly lower matching rate.

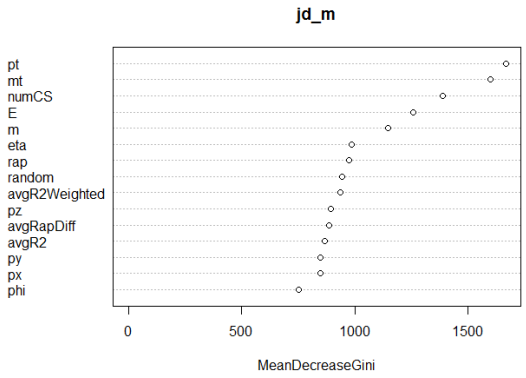

Figure 4: The estimated variable importance is plotted for each jet attribute. The attributes with a lower performance than the random feature likely are not worth keeping.

tures, however, so I ran a random forest algorithm to estimate variable importance. As a baseline for variable importance, I included an extra truly random feature. All features that performed worse than this baseline were cut. Finally, I added a few features that I thought could be good predictors. First, I added a feature that kept track of the magnitude of the momentum in the x-y plane, which turned out to be one of the top predictors. Also, I included a metric for how spaced out the particles within a jet were, because I noticed that light jets on average had a thinner cone than gluon jets.

3 Results and Conclusions

My first attempts at training a model were done using a vraiety of linear classifiers. A confusion matrix of the results can be seen in figure 5. Con-

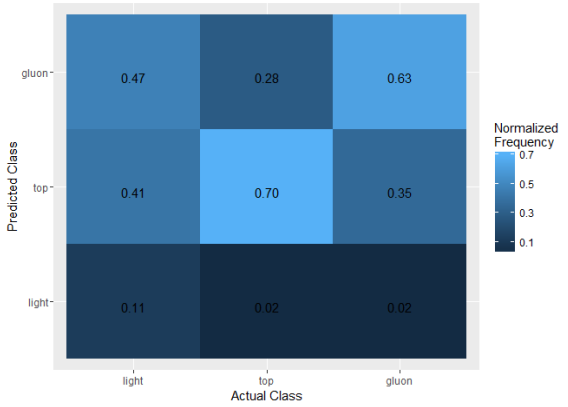

Figure 5: Confusion matrix showing the performance a basic decision tree.

sidering the 20 percent mismatch in the supervised sample, the results from this training is actually better than expected. Higher purity is not often achieved in physics events without significantly cutting out sample size. The surprising performace is likely because the jet properties trained on in this experiment are much simpler than in real data. For this study, I did not include interactions of the jet with the detector. This is clearly a major step in actual physics analysis, and would make our sample much harder to accurately train on.

Although the top and gluon jets could be discriminated with decent accuracy, this was not the case for light jets. There are a few reasons this might be the case. The most simple explanation is that light jets do not appear as often as the other flavors. In the sample studied, only about 10 percent of the sample consisted of light jets. This means that if the model had trouble deciding a category, predicting a light quark would always be a bad decision. So, it had a bias against learning to classify light quarks. This general accuracy was true when studied with many different classifiers such as knn, decision trees, and support vector machines.

Class imbalance is a well studied issue in classification problems and the correct approach can vary from case to case. One methodology to help a model predict rare events is to change the models loss function to have a greater penalty for misclassifying the rare event. This can work well when the rare event is more important to predict than the other categories. However in this study, the light flavor of quarks is not inherently more important. Also, because the sample is simulated, this lends itself to a sampling based approach. It is easy to

scale up the number of jets created when making the sample, and then randomly cutting out a certain amount of the majority classes. In this way the dataset can be more balanced while not compromising sample size. Additionally, this avoids common pitfall in oversampling where the same small target class is fed in many times to a model (Fawcett).

The larger sample size of balanced data was then fed into a neural network with one hidden layer. This neural networks was run on 50,000 samples each with 10 features to train on. These features included angle away from beam, momenta, calculated mass, jet topologies, and jet energy. The results from this neural network can be seen below. As expected, having an equal proportion of classes led to more balanced results. The heavy quarks (top and bottom) were once again the easiest jets to discriminate. This was expected by simply examining some of the attribute distributions, as bottom quarks had a larger amount of high energy instances. Overall, the performance on gluon and heavy flavor remained largely the same while the light flavor classification greatly improved. Although future study is needed to see how these results could be improved, this likely would involve improving the purity of the sample trained on. Perfect accuracy cant be expected if the model is trained to learn associations that dont actually exist. One method to remedy this problem would be to create an even larger sample, and cut out the instances where there is less uncertainty about the true matching. As stated before though, this could bias the model to only be able to learn the jet classification of high energy jets.

The logical next step in this process would to take the predicted particles from each jet and try to guess the underlying process that produced the event. Also, the simulated data can be run through software designed to emulate the CMS detector in order to have a dataset similar to what is actually measured in a particle physics experiment.

Acknowledgments

The guidance of Professor Chris Neu was important for the decisions about how to construct my sample. Additionally, I would like to thank the HPC division at UVa for their guidance on how to effectively use the Rivanna computer cluster to create my sample and train my models.

References

"CMS" *CERN Accelerating Science.* 8 May 2017.

Fawcett, Tom. "Learning From Imbalanced Classes." Silicon Valley Data Science. 1 February 2017

"Understanding Color." Giacomo, A. Fisica University, Pisa. Nov. 2001.

"How an Accelerator Works." *About CERN.* 8 May 2017.

"SM of Particle Physics." `https://en.wikipedia.org/wiki/Standard_Model#/media/File:Standard_Model_of_Elementary_Particles.sv`

"The Standard Model." *About CERN.* 8 May 2017.