Jet Identification in QCD Samples at DØ

Zachary D. Hodge
Austin Peay State University
Columbia University
Nevis Laboratories

July 31, 2008

Abstract
The jets of b-hadrons and gluons in QCD Monte Carlo samples can be identified and separated from light-quark jets other background jets with the use of the ROOT toolkit TMVA and the Neural Network and Boosted Decision Tree algorithms contained within it.

1 Introduction

1.1 FermiLab Tevatron Collider and DØ Experiment

The FermiLab Tevatron Collider is located outside Batavia, Illinois. It is a proton/antiproton collider and is capable of 1.96 TeV center of mass energy collisions. In the ring, two individually enclosed beams circle in opposing directions, one beam of protons, the other with antiprotons. The DØ experiment is located at one of the points on the main ring where the two beams cross. The DØ detector consists of multiple tracking and energy deposition layers. Each layer has a specific purpose in capturing event data. These all aid in the reconstruction of events, which is the basis of this study.

Figure 1: Detector Assembly [9]
1.1.1 Silicon Detector and Fiber Tracker

The central tracking detectors provides the capability of measuring charged particles originating from the point where the initial interaction of the proton and antiproton collide (primary vertex). The silicon microstrip tracker (SMT) and central fiber trackers (CFT) surround the beam pipe, the inner most part of the detector. The SMT helps determine the position of charged particles near the location of the primary vertex and the CFT determines a charged particle's momentum and charge. Directly outside of the fiber tracker is a solenoid capable of producing a 2T magnetic field. A charged particle's path will bend while in the presence of a magnetic field, this allows the trackers determine charge and momentum.

![Figure 2: Inner Tracking Chamber](image)

1.1.2 Liquid Argon Calorimeter

Outside of these trackers is the liquid Argon (LArg) and depleted Uranium electromagnetic and hadroninic calorimeter. The calorimeter provides the ability to measure the energy of electromagnetically and strongly interacting particles. The calorimeter is divided into 3 regions and each are fully enclosed to help maintain the liquid Argon at a constant 90K. The electromagnetic calorimeter uses sheets of depleted uranium as the absorber material, where as the hadronic calorimeter uses copper. The absorber material is very dense and causes a particle to radiate as it passes through it. This subsequently causes a shower of particles to be produced which ionize the liquid argon and allow for a signal to be collected. This provides the ability to determine a particle’s energy as it is completely deposited in the detector.

1.1.3 Muon Tracking System

Surrounding all of these detectors is the muon tracking system. The muon tracking system uses a magnetic field and drift chambers to track muons. A 2T torodial magnetic field is used to bend muon paths and determine momentum. Drift chambers provide the ability to track a muons position.
1.2 Bottom quarks, Hardronization, and Jets

At the FermiLab Tevatron Collider protons and antiprotons are accelerated in opposite directions and then crossed to produce collisions with center of mass energies of 1.96 TeV. During one of these collisions, many different particles are produced. Many of these particles can be tracked and measured through the use of the detector described previously. Some processes cannot be measured directly, such as the top quark and Higgs boson production (Figure 5). These processes are very interesting, and can be evaluated through the existence of bottom quarks in their decays.

1.2.1 Bottom quarks

Bottom quarks are a third generation fermion and have a significantly higher mass (4.2 GeV) than all the other quarks, excluding the top. The third generation is named this because it was the last grouping to be discovered. Fermions are particles with 1/2 integer spins (obey Pauli Exclusion Principle) and contained within this family are quarks and leptons. Bottom quarks carry a fractional electric charge of -1/3e. Bottom quarks are produced in many different ways. Some processes which can create bottom quarks are a top quark which will decay into at least two and possibly six bottom quarks. Also the decay of the theoretical Higgs boson is expected to decay in a similar way, producing bottom quarks.

1.2.2 Hadronization and Jets

Color charged particles (parton) do not exist freely, and will immediately form a bound state with another parton that is created from the vacuum. This is due to color confinement, which is where a parton is stretched in its bound state, and finds a more energetically appealing situation if another parton is created to bind to. This process is known as hadronization. Quarks will hadronize into either a meson or baryon. Mesons are a two quark configurations and Baryons are three quark configurations. B-hadrons have a relatively longer lifetime than other light-quark hadrons, of approximately $10^{-12}$ s, and therefore travel a few millimeters before decaying. The created hadron will then proceed to decay and produce even more particles, forming a collimated spray of particles known as
a jet. This jet can be detected in the electromagnetic and hadronic calorimeter. The data collected on the jet can be reconstructed to infer the existence of the parton and further study its characteristics.

2 Variables, selection and KS Test

2.1 Algorithms and Variables

There are currently four separate b-tagging algorithms used at DØ. Three of which rely on charged particle tracks to differentiate between b-jets and light-jets, while the fourth uses the presence of a muon. Particles in the detector that leave a straight line are known as tracks. The DØ Neural Network (NN)
uses a list of predefined variables from the tagging algorithms to determine whether a jet presents the characteristics of a b-jet or that of other light-quark jets. The neural network connects many different, seemingly uncorrelated data into a single variable.

2.1.1 Jet Tagging Algorithms

The Counting Signed Impact Parameters (CSIP) algorithm counts the tracks contained within a jet that have a large impact parameter significance to the primary vertex. The Jet Lifetime Probability Tagger (JLIP) uses the impact parameters from all tracks within a jet and combines it to one variable, the Jet Lifetime Probability (JLIP Prob). The Secondary Vertex Tagging (SVT) uses the tracks from a jet that are displaced from the primary vertex, then reconstructs a displaced vertex based on those tracks. The Soft Lepton Tagging (SLT) uses the presence of muon in the jet to tag it.

2.2 Selection of Variables based on a KS Test

Each of the algorithms produce sets of variables that can help in further discrimination of b-jets from light jets. The Kolmogorov-Smirnov goodness of fit test is used to determine which of the variables provide the greatest separation power between b-jets and light jets.

2.2.1 Kolmogorov-Smirnov Test

A Kolmogorov-Smirnov (KS) Test is a goodness of fit test, used to determine the difference between two one-dimensional distributions.

\[ D_{n,n'} = \sup |F_n(x) - F_{n'}(x)| \]  

The KS test compares the locations of two empirical cumulative distribution functions to determine the greatest distance between the two. Variables with a higher KS value have distributions that differ significantly and thus provide greater separation power between the two distributions. The KS test is calculated over all variables to determine which have the greatest separation power.

2.2.2 KS Test results and Variable Selection

The final variables are selected by choosing those with the greatest separation between b-jet and light jet samples. The variables are ranked from greatest separation to least, and the top 24 variables were chosen to be used. See Table 1 for the list of selected variables and corresponding KS values. The table is a list of variables created by its corresponding jet tagging algorithms. For example svt_dlsig is the decay length significance (dlsig) as calculated by the Secondary Vertex Tagging (svt) algorithm.

3 MLP and BDT

3.1 ROOT and TMVA

This entire study was done using ROOT and the TMVA program contained within it. The ROOT framework provides the platform to do data set on the
<table>
<thead>
<tr>
<th>Variable</th>
<th>KS Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>svt_dllsig</td>
<td>0.5570</td>
</tr>
<tr>
<td>mu_all_put</td>
<td>0.5376</td>
</tr>
<tr>
<td>svt_dll</td>
<td>0.5178</td>
</tr>
<tr>
<td>svt_dllsig_z</td>
<td>0.5048</td>
</tr>
<tr>
<td>svt_dll_3d</td>
<td>0.4717</td>
</tr>
<tr>
<td>svt_dllsig_3d</td>
<td>0.4601</td>
</tr>
<tr>
<td>svt_dll_z</td>
<td>0.3927</td>
</tr>
<tr>
<td>jlip_prob</td>
<td>0.3076</td>
</tr>
<tr>
<td>mu_all_csi_comb</td>
<td>0.2912</td>
</tr>
<tr>
<td>svt_lifetime</td>
<td>0.2501</td>
</tr>
<tr>
<td>svt_svtrkpt</td>
<td>0.2127</td>
</tr>
<tr>
<td>svt_svpt</td>
<td>0.2126</td>
</tr>
<tr>
<td>svt_mass</td>
<td>0.1999</td>
</tr>
<tr>
<td>svt_ptcomass</td>
<td>0.1997</td>
</tr>
<tr>
<td>svt_ptrel</td>
<td>0.1985</td>
</tr>
<tr>
<td>svt_ntracks</td>
<td>0.1985</td>
</tr>
<tr>
<td>svt_ntracks_pv</td>
<td>0.1980</td>
</tr>
<tr>
<td>svt_trkpt</td>
<td>0.1902</td>
</tr>
<tr>
<td>svt_ntracks_pv</td>
<td>0.1828</td>
</tr>
<tr>
<td>svt_chisq</td>
<td>0.1627</td>
</tr>
<tr>
<td>jlip_ntracks</td>
<td>0.1365</td>
</tr>
<tr>
<td>svt_chisqndof</td>
<td>0.1318</td>
</tr>
<tr>
<td>jet_numtracks</td>
<td>0.1127</td>
</tr>
<tr>
<td>jet_ntracks_cpf0</td>
<td>0.1034</td>
</tr>
</tbody>
</table>

Table 1: Komogorov-Smirnov Test Results
input layer, then sent to the first hidden layer where the data is analyzed and determined whether it presents background or signal like qualities. The MLP contains hidden layers that each do analysis on the data to further separate signal from background. Each hidden layer calculates some linear combination of the input variables, compares its results to known signal and background events, calculates the error, and then reweights and repeats the process. MLPs are known as a feedforward architecture, meaning it sends information learned on one hidden layer to the next hidden layer. The output of the MLP ANNs require training to be effective at differentiating between a signal event and a background event and are therefore trained on 3000 known signal and background events.

3.3 Boosted Decision Tree

Boosted decision trees (BDT) are a form of binary decision tree that operate on a yes/no decision as to what the variable should be classified as. BDTs operate over one variable at a time, performing individual decisions that culminate in either classifying that variable as signal or background. Figure 5 shows a schematic of a decision tree. Once passing through each layer of the tree, the variable will reach a terminal node (leaf) that classifies it as either signal or background as a result of the path it took to reach that leaf. Statistical fluctuation in a sample can lead to a misclassification on a variable. Misclassification is where a known signal event lands on a background terminal node or vice versa. To account for these misclassifications, the decision tree undergoes a process known as boosting. The act of boosting is training a variable several times using a boosted (reweighted) sample. Each time a variable is misclassified, it is reweighted to make it signal like, then is repeated in the BDT. This helps to decrease statistical fluctuation in the response as a result of the training sample.

![Decision Tree Diagram](image)
3.4 Comparison of MLP and BDT

The background rejection versus signal efficiency curves of the MLP and BDT classifiers were compared to determine which provided better feedback at separating b-jets from light-quark jets. Both the BDT and decorrelated BDT (BDTD) showed a greater performance over the MLP of approximately 3%. (See Figure 6) Because of the performance boost in the BDT classifier, it was chosen to carry out the remainder of the study. Further investigation was put into determining what the best setting were for the BDT. Figure 7 shows multiple tests on the different BDT settings, with BDT5 exhibiting the greatest performance curve (another 3% performance gain). The settings to achieve this (BDT5) are as follows:

- NTrees=1000
- BoostType=AdaBoost
- SeparationType=GiniIndex
- NCuts=20
- PruneMethod=ExpectedError
- PruneStrength=5

![Background rejection versus Signal efficiency](image)

Figure 7: Background Rejection vs. Signal Efficiency of MLP, BDT, and BDTD

4 Boosted Decision Tree with Variables

4.1 BDT on Z → bb Monte Carlo Sample

The BDT was trained and tested using Z → bb Monte Carlo samples to determine the ability of the BDT to separate b-jets from light-quark jets. Predefined
$Z \rightarrow bb$ trees were input as signal events and $Z \rightarrow qq$ were input as background events.

### 4.1.1 Results of BDT with b-jets and light-quark jets

It is fairly obvious that the BDT was capable of separating b-jets from light-quark jets (See Figure 8). The background and signal distributions show a considerable difference. Background events are shifted toward a -1 value, whereas the signal events are shifted to a +1 value. There is a little crossing between the two distribution around 0 but this is small compared to where the bulk is located.

### 4.2 BDT on quark/gluon/other jets Monte Carlo Sample

The BDT was trained and tested using quark jets, gluon jets, and background other jets Monte Carlo Samples to determine the ability of the BDT to separate quark jets from gluon jets and other jets in the event. An event sample was preprocessed with cuts to artificially separate the quark jets, gluon jets and other jets from the MC sample.

### 4.2.1 Results of BDT with quark jets and other jets

The first BDT test was operated over all quark jets as signal and other jets as the background. Other jets are are jets produce from some interaction that is not directly related to a quark or gluon event. The BDT shows fairly good capability of separating the two jet types (See Figure 9). The Background rejection vs signal efficiency curve shows a decrease efficiency, but this can be
attributed to the way cuts were applied in preprocessing to separate the jets. All jets were required to have at least two tracks, which consequently makes the other jets have characteristics like that of quark jets. With this considered the BDT still shows good ability in separating the two.

4.2.2 Results of BDT with quark jets and gluon jets

The second test using the BDT classifier was on quark jets set as signal and gluon jets set as background. This test showed a significantly lower performance than all previous tests. The BDT had difficulty separating quark and gluon jets. The output distributions are nearly overlapping (See Figure 12) and has a very low efficiency curve (See Figure 13). These results are not unexpected though, as many of the variables for quark and gluon jets have similar distributions and values.

5 Conclusion

This study showed that through the use of TMVA and the Boosted Decision Tree (BDT) classifier algorithm, assorted high energy jets can be separated in Monte Carlo samples. The BDT classifier showed a ~3% efficiency gains over the Multi-layer Perceptron (MLP) in separating b-jets. The BDT classifier also showed ability in separating all quark jets from other background jets. The BDT’s separation power was suppressed in quark and gluon jets but this can be attributed to the similar variable characteristics in the two jets. The Boosted Decision Tree has proven to be a capable classifier algorithm in jet separation.
Figure 10: Background Rejection vs Signal Efficiency b-jets and light-quark jets

Figure 11: BDT Classifier Response: Signal = quark jets Background = other jets
Figure 12: Background Rejection vs Signal Efficiency quark jets and other jets

Figure 13: BDT Classifier Response: Signal = quark jets Background = gluon jets
Figure 14: Background Rejection vs Signal Efficiency BDT quark jets and gluon jets
References


