W vs. QCD Jet Tagging at the Large Hadron Collider

Bryan Anenberg: anenberg@stanford.edu; CS229

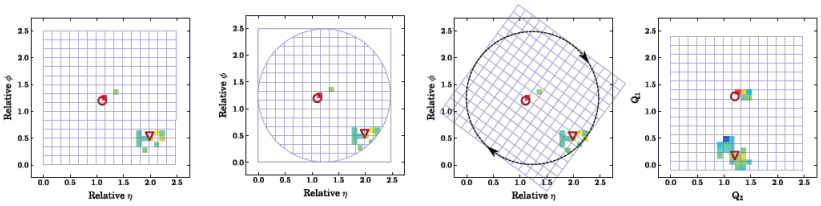
December 13, 2013

Problem Statement

High energy collisions of protons at the Large Hadron Collider (LHC) produce massive particles such as W, Z, Higgs bosons, and top quarks. A key task in the search for physics beyond the standard model is to study the kinematic configurations of these heavy particles. The massive particles are observed indirectly by the energy signature they generate. The heavy particles decay into quarks and gluons which deposit energy on the ATLAS calorimeter after hadronization. The collimated stream of particles produced by the hadronization of a quark or gluon is referred to as a jet. The goal of the project is to discriminate between jets that originate from boosted electroweak bosons such as W-boson and top quarks (referred to as W-jets) and those originating from light quarks or gluons (referred to as QCD jets). The motivation to improve jet tagging (classification) is due to the decrease in performance of standard techniques for reconstructing the decays of heavy particles with a large background of ordinary QCD jets. In this project compares the performance of a number of supervised learning algorithms such as SVM, Fisher Linear Discriminant, and random forest in their ability to distinguish between W-boson and QCD jets.

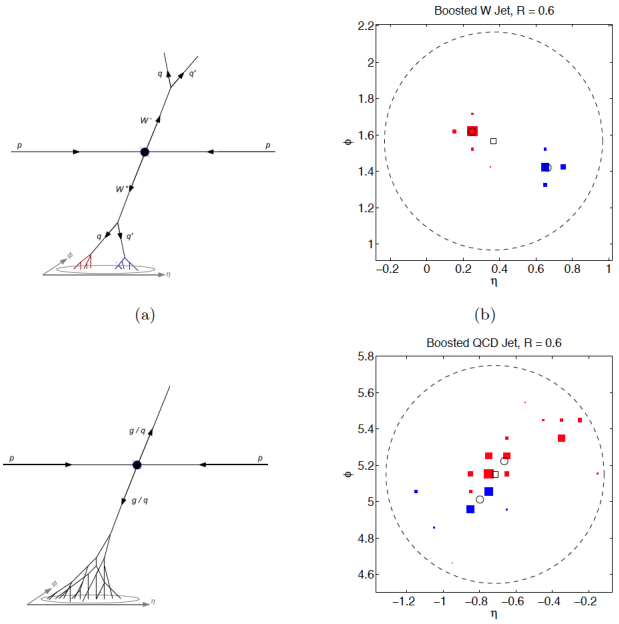
Jet Image

The data we consider is a Monte Carlo simulation of a top quark decay. The data is formatted as 25 by 25 pixel 2-D images where the pixel intensity corresponds to the transverse momentum (pT) of the particle as detected by the calorimeter. In unrotated image, the axes correspond to eta and phi. However, pre-processing rotates each image so that the clusters align vertically. As a result, the axes no longer exactly correspond to rapidity-azimuth plane (eta-phi), but to a spatial dimension unique to each image.



The first of the following images depicts a W boson

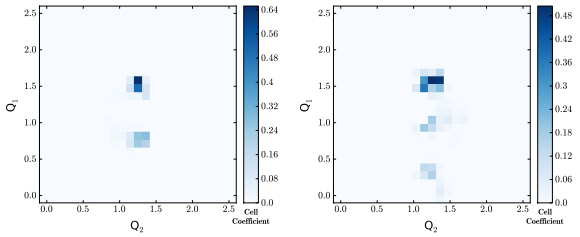
Figure 1: Hadronic decay sequence



(a) (b)

The diagrams on the left illustrate the typical decay sequence of either a W-boson or a QCD jet. a) W-boson decay. b) QCD event. Note that the (a) W jet is typically composed of two distinct p_T peaks, whereas the (b) QCD jet deposits its energy over a wide region of the calorimeter as a result of splitting observed in the event. Image curtosy of source 1.

decay into two quarks. W boson jets exhibit a two prong structure where each prong corresponds to a quark generated in the decay. The second image is an example of a QCD jet. QCD jets typically display an asymmetric intensity pattern. QCD jets typically have a single high p_T peak and a variable number of lower p_T peaks.



The data analyzed in this paper relies upon preprocessing algorithms written by Josh Cogan, graduate researcher with the ATLAS group at SLAC.

Classification performance on raw data

Support Vector Machine

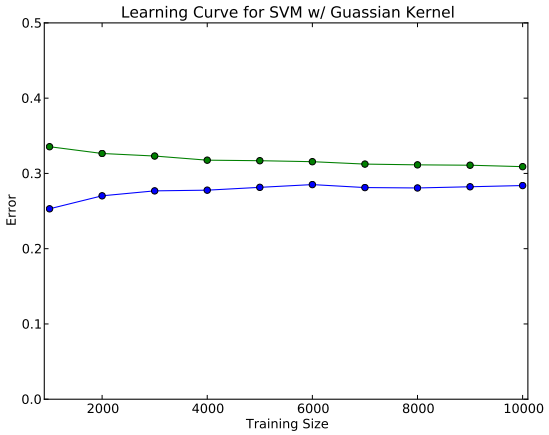
To establish a baseline for classification accuracy, we study the Support Vector Machine. Of the choices of kernel, the gaussian kernel SVM with $K(x,z) = exp\left(-\gamma||x-z||^2\right)$ performed the best. The optimize the choice of parameters, C for l_1 regularization and the γ coefficient of the kernel, we perform hold out cross validation. Hold-out cross validation on $C \in \{1, 10, 100, 1000, 10000\}$ and $\gamma \in \{0.001, 0.01, 0.1, 1, 10\}$ resulted in parameters $C = 100$ and $\gamma = 0.1$ yielding the lowest error when tested on the hold out cross validation set. The learning curve for the Gaussian kernel SVM is displayed below.

At best gaussian kernel SVM classified the jet image data with 31.69% error.

The learning curves generated for the Gaussian kernel SVM illustrate that the model possesses high bias. The test and training error are both high and at comparable levels. To improve the classification of the SVM, we generate additional features.

The goal of the classification algorithm is to maximize the ability to distinguish between the signal (W jets) and background (QCD jets). To evaluate the performance of a classification algorithm we introduce the measure known as "Significance" $S = {}^{TP}/\sqrt{FP}$ where TP is the true positive, the number of times the classifier correctly classified the signal and FP is the false positive, the number of times the classifier classifies the background as signal. We consider the $\sqrt{FP}$ because $\sqrt{FP}$ is the RMS of the Poisson distribution with mean FP. A classifier that achieves good discrimination between the signal and background should maintain a high Significance value. The significance value of the gaussian kernel SVM with $C = 100$ and $\gamma = 0.1$ is $S = 86.04$

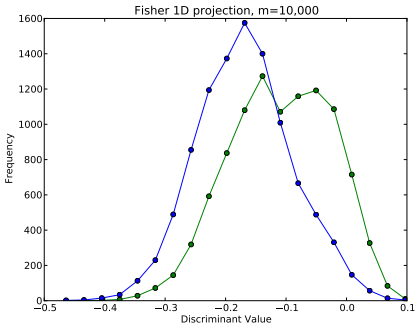
Figure 2: Learning Curve for Gaussian Kernel SVM

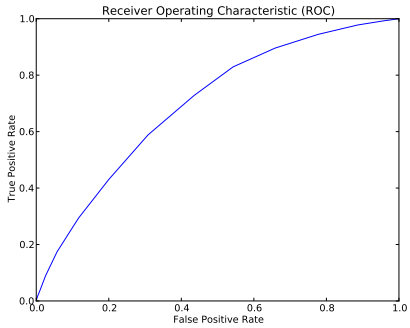


Blue: training data. Green: testing data. Training size of k indicates that the linear SVM was trained on a data set consisting of k W-jet samples and k QCD jet samples. The total training set size is $2k$.

Fisher Linear Discriminant

The classification of quark-initiated vs. gluon-initiated jets parallels the gender recognition problem. The similarity motivates the use of the Fisher Linear Discriminant (FLD) facial recognition algorithm on the classification of jets. The FLD objective is to perform dimensionality reduction by finding the direction by which the classes are most separate. The algorithm generates a classifier with a linear decision boundary by fitting the class conditional densities to the fisher criteria of maximizing between class scatter while minimizing the within-class scatter. For jet classification we only consider the two class FLD algorithm. In this case we have μ_1 and μ_2 the mean vectors of the two classes. M_1 and M_2 are the total number of samples, $x^{(i)}$for either class. The within class scatter matrix is given as $S_w = \sum_{i=1}^{2}\sum_{j=1}^{M_i}(x_j - \mu_i)(x_j - \mu_i)^T$ and between class scatter $S_b = \sum_{i=1}^{2}(\mu_i - \mu)(\mu_i - \mu)^T$ where $\mu = \frac{1}{2}(\mu_1 + \mu_2)$. The goal is to find an orthonormal projection matrix W_{opt} given by the optimization objective $W_{opt} = \underset{W}{argmax}\frac{|W^T S_b W|}{|W^T S_w W|}$. When trained on the raw jet pixel intensities, the FLD produced the following projection onto the one dimensional subspace. In this example we train the FLD with a 10,000 sample training set. The green curve corresponds to the quark-initiated jets. Blue curve is QCD jets.

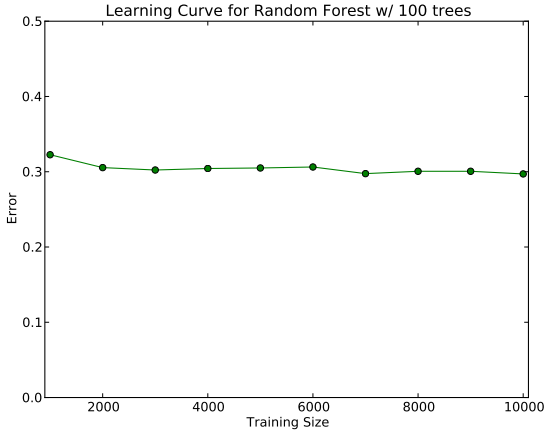




The Receiver Operating Characteristic (ROC) curve which illustrates the performance of FLD as the discrimination threshold is varied. When setting the decision boundary corresponding to signal efficiency of 50%, FLD algorithm performed with 35.3% error and $S = 75.355$ at best. This is notably worse performance than the SVM with gaussian kernel.

Random Forest

The third classification algorithm that performed well when classifying the jets was random forests. The random forest classifier builds a model by constructing $NTREES$ decision trees by repeatedly resampling the training data with replacement. The random forest classifies the test data by returning the consensus vote of $NTREES$. Every node of a decision tree corresponds to one of the input features. The edges between a node and its children give the possible values of that input feature. Each leaf of the tree corresponds to the binary classification of the total sample given the features represented by the path from the root to the leaf. When constructing the member trees of the random forest, each node shares an edge with the best random subset of the features. Initial tests revealed the ExtraTreesClassifier implemented in sklearn performs even better than the standard RandomForestClassifier. The ExtraTreesClassifier adds in an additional layer of randomness by choosing the best threshold among a set of randomly generating thresholds that are used to determine the best random subset of features to connect to each node.



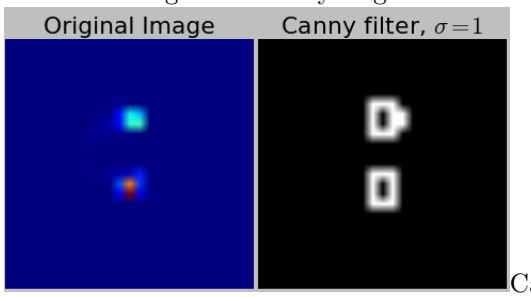
With $NTREES = 100$ the random forest classifier performed with 30.5% test error and significance of $S = 90.51$. Random forest classifier performed better than both the SVM and FLD algorithms.

Feature Expansion

In order address the bias of the model, we attempt to extract additional features from the data. Historically jet tagging techniques attempt to classify a jet by analyzing its substructure. This technique attempts to capitalize on the fundamentally different energy patterns of W jets and QCD jets. We employ a number of image processing techniques with the goal of extracting meaningful features from the jet images. For example, we would like to quantify the number of subjets present in each jet image where a subjet refers to a clusters of pixels with a p_T much larger than neighboring pixels and the relative p_T of the subjets.

Canny Edge Detection

Figure 3: Canny Edge Detection



Canny edge detection identifies the two subjets of a W jet.

The first technique we apply to expand the set of image features is canny edge detection. By applying edge detection to the images, we hope to more clearly distinguish the subjets from the background. W jets should exhibit edges around their two subjets where as QCD jets could

Figure 4: Peak Local Maximum Filter

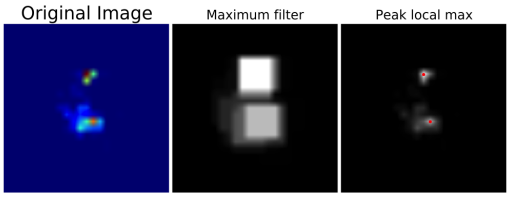


Image 1: W jet. The red points denote the coordinates of the local peaks of the image. The maximum filter merges regions within the rectangular region to identify the local maximum.

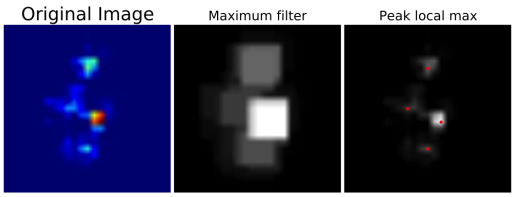


Image 2: QCD jet.

display a variable number of edges. The edge detection scheme assumes that there is an underlying continuous intensity function which is sampled at the image points. The edges are found by computing derivatives of this intensity function. σ varies the width of the Gaussian used to reduce the effect of noise present in the image. The choice of $\sigma = 1$ yielded the a filtered image with the most well defined edges. Training the Gaussian kernel SVM directly on the image filtered by the canny edge detector or with the additional features appended onto features of the original image did not improve in the classification of jet images. In fact, the testing error rate for the Gaussian kernel SVM ($C = 100$, $\gamma = 0.1$) when trained on 4000 samples increased from 31.78% to 35.75%. Since the raw data contains very little background noise, the edges distinguishing each subjet were already clear. Drawing edges with the edge detector could reshape or rescale the subjet edges inaccurately. Edge detection would generate more meaningful features when considering jet images that contain background p_T due to energy deposited by additional proton-proton collisions in the event.

Peak Local Maximum Filter

To ascertain information about the substructure of the jet image, we employ scikit-image's peak_local_max function to find the coordinates of local peaks (maxima) of the image. peak_local_max identifies the local maxima by first applying a maximum filter to identify the pixels with large values. Potential maxima that are located within a pre-selected radius are merged together. The coordinates of the merged maxima is returned as the coordinates of the local maxima of the original image.

We extract the following 5 features by using the coordinates of the local maxima:

1. Whether the jet image contains two subjets.

   Binary value: 1 if subjet contains two subjets (local maxima), 0 if it contains a different number of subjets.

2. p_T of the largest local maxima.

3. p_T of the second largest local maxima.

4. Difference in pixels between the two peaks with the greatest p_T values. The difference is given as $\Delta R = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ where x and y are the axes of the image. QCD jets are known to often have 1 high p_T subjet and one lower p_T subjet at wide angle. Thus we might expect that the ΔR for QCD jets is on average greater than the ΔR for W jets.

5. Ratio of the p_T of the largest peak j_1 and the second largest peak j_2, $\frac{j_1}{j_2}$. This feature quantifies the relative size of the two subjets. This measure is useful since the two subjets of a W jet should have similar p_T values. A QCD jet is more likely to have a single high p_T peak and a second smaller p_T peak.

The gaussian kernel SVM, FLD, and Random Forest algorithms all demonstrated a decrease in performance when trained with only the 5 feature extracted above. SVM performed with 38.2% testing error $S = 69.23$, Fisher with 40.13% testing error $S = 64.20$, and random forest with 38.8% testing error and $S = 69.24$. However, appending the additional 5 features onto the original image data yields marginal performance improvements for SVM and FLD. SVM performed with 31.5% testing error $S = 86.7$, Fisher with 35.2% testing error $S = 78.35$, and random forest with 30.89% testing error and $S = 89.64$. The results indicate that there is potential in extracting additional features from the image data, however further efforts are required for significant improvements in performance.

Feature Selection

We suspect that the intrinsic dimensionality of the data is much lower than 625 since all of the jet images from either class look similar. This insight motivates the discussion of how to reduce the dimensionality of the training set to leave only those features that are critical to the jet classification.

Principle Component Analysis (PCA)

The goal of PCA is to identify the subspace in which the data approximately lies. By projecting the data on the k principle components, this procedure can potentially extract the most characteristic features from the data. Performing hold-out cross validation on k, the number of principle components, we find that optimal choice is $k = 60$. However, even with this choice of k, after applying

PCA on the expanded data set the testing error of the gaussian kernel SVM is 38.53%, which is notably higher than the error measured without PCA.

Recursive Feature Elimination

Beyond PCA, we attempt to reduce the dimensions of the features by a number of techniques. Foremost, we apply recursive feature elimination by using a linear kernel SVM to assign weights to each feature. At each iteration, this feature elimination procedure eliminates the feature with the smallest weight from a trained SVM. The procedure is recursively repeated on the pruned feature set until the desired number of features to select is eventually reached. In addition to being computationally expensive for a feature size of over 600, recursive feature elimination does not achieve any noticeable improvement in classification. When trained on the reduced data set the SVM only manages to achieve 32.98% testing error with $S = 82.77$ and FLD performs with 35.32% testing error and $S = 77.67$. However, random forest observes an infinitesimal improvement to achieve 30.76% testing error with $S = 90.61$.

Tree-based Feature Selection

The random forest can also be used to determine the most relevant subset of features by using the average information gain achieved during the construction of the $NTREES$ voting decision trees. The Sklearn package is used to implement this procedure. When trained on the extended training samples, the tree-based procedure reduced the number of features to 187. This reduction in features allowed the classifiers to run more quickly and still demonstrate strong performance. In fact, the classifiers on performed better when trained on the reduced data set. The SVM performed with 31.58% testing error and $S = 86.62$, FLD performed with 35.24% testing error and $S = 78.29$, and random forest performed with 30.6% testing error and $S = 91.49$.

Conclusion

In this project the gaussian kernel SVM, Fisher Linear Discriminant, and the random forest classifiers were compared in their capabilities to discriminate between W and QCD jets. Results indicate that feature expansion techniques motivated by insight into the physical data can improve classification. However, identifying the most significant features still presents a problem. Feature selection techniques demonstrated a variety of results. Although most feature selection procedures did not yield significant improvement in classification, they still improve the speed of the classifiers. Overall the random forest and gaussian kernel SVM classifiers performed the best.

References

[1] Thaler, Jesse, and Ken Van Tilburg. Identifying Boosted Objects with N-subjettiness. Cambridge: Center for Theoretical Physics, MIT.

[2] Scikit Learn Python library

[3] Scikit-image Python image processing library

[4] Sakarkaya, Mutlu, Fahrettien Yanbol, and Zeyneb Kurt. Comparison of Several Classification Algorithms for Gender Recognition from Face Images. Istanbul: Yildeiz Technical University, n.d. Web. 14 Dec. 2013.

[5] Belhumeur, Peter N., Joao P. Hespanha, and David J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. N.p.: IEEE Transactions on Pattern Analysis and Machine Learning, n.d. Web. 14 Dec. 2013.