Thesis

Wei Tang

Contents

1 Abstract 3

2 Introduction 3

3 Experiment Background 4
  3.1 The Large Hadron Collider 4
  3.2 The ATLAS Detector 4

4 Search for New Resonances 4
  4.1 Event Selections 5
  4.2 Signal Generation 5
  4.3 Background Simulation 6
  4.4 Decay Polarizations 7

5 Standard Model Signal and Background 11
  5.1 Baseline and D2 Cuts 11
  5.2 Background and Signal Features 12
  5.3 Features Preprocessing 14

6 Traditional Box Cuts 15

7 Support Vector Machine 15

8 Random Forest 15

9 Neural Net 15
  9.1 Structure and Terminology 15
  9.2 Methodology 17
  9.3 Test Set 17
  9.4 Results 20
     9.4.1 All Features 20
     9.4.2 High Level Features 21
1 Abstract

The Standard Model is by far the most encompassing physics theory. With the recent discovery of the Higgs Boson, the Standard Model has performed extremely well against experimental data. However, the theory is intrinsically not complete. Many extensions of the Standard Model predict new resonances decaying to a W boson and a photon. This thesis presents a search for such resonances produced in proton-proton collisions at $\sqrt{s} = 13$ TeV using a dataset with an integrated luminosity of 36.1 $fb^{-1}$ collected by the ATLAS detector at the Large Hadron Collider. The W bosons are identified through their hadronic decays channel. The data are found to be consistent with the expected background in the entire mass range investigated. Upper limits are set on the production cross section times decay branching ratio to $W + \gamma$ of new resonances with mass between 1.0 and 6.8 TeV.

The second part of this thesis looks at the extraction of the Standard Model $W + \gamma$ production. The main background noise for the extraction comes from the Standard Model production of $\gamma$ and jets. In order to streamline and improve the accuracy of the event filtering process, this thesis developed a deep neural net classifier to identify signal and background decay products and improved the signal to noise ratio.

2 Introduction

This thesis first introduces background information on CERN, LHC and the ATLAS experiment in section 3. The later part is divided into two parts. The first part consists of a talk presented at the 2017 American Physical Society Division of Particles and Fields conference. At CERN, we analyze proton-proton collision experimental data to search for evidence of new particles. The first part examines possible decay channels of an unknown massive boson $X$ decaying into $V$ plus a gamma, where $V$ can be either a $W$ or a $Z$ boson. The production mechanism can be either gluon-gluon fusion or $q \bar{q}$ annihilation. Figure 1 illustrates the possible production and decay channels.
Section 4 describes preliminary results from a generic search for new massive X bosons. The ultimate goal of the search is to explore the X mass range from 250 GeV to highest energy attained in 13TeV p-p collision. Section 4.4 explains the underlying reason for observed detector inefficiencies for spin 1 decays.

The second part of the thesis attempts at superseding the traditional box cuts method employed in filtering signal events using machine learning classification techniques. Section 5 describes the background and signal events we attempt to classify. Section 6 presents the classification results as obtained using the traditional box cuts methods. Sections 7, 8 and 9 present the classification performances using support vector machine, random forest and neural net. Section 10 analyzes the effects of using different neural net parameters. Section 12 looks at combining different classification structures into an ensemble model.

Last but not least, this thesis outlooks the future directions and discusses the strengths and weaknesses of the research.

3 Experiment Background

3.1 The Large Hadron Collider

3.2 The ATLAS Detector

4 Search for New Resonances

The boson decay products from the unknown massive boson X can further undergo either leptonic or hadronic decay. In the leptonic decay, the Z boson further decays into a pair of either electron and positron or $\mu^+\mu^-$. 

Figure 1: $X \rightarrow V\gamma$
4.1 Event Selections

The leptonic decay events were selected using single electron and muon triggers with a nominal transverse momentum larger than 26 GeV, supplemented by di-lepton triggers with lower thresholds. The Z bosons which decay to a pair of leptons are selected using well measured, isolated electron and muon pairs with invariant mass within ±15 GeV of the invariant mass of Z bosons. The photons are required to be isolated with pseudo rapidity $|\eta| < 1.37$ or $1.52 < |\eta| < 2.37$ and transverse momentum $P_T > 10$ GeV. These requirements ensure that we are using the detector region with highest granularity. The final $Z\gamma$ event selection requires the photon to have transverse momentum greater than 30% of the combined $Z\gamma$ invariant mass.

4.2 Signal Generation

In the signal simulation, unknown massive boson X decays with a narrow width to a Z boson and $\gamma$. The narrow width chosen is 4MeV. Figure 2 on page 5 shows the $Z\gamma$ combined invariant mass distribution with all selection cuts applied.

![ATLAS Simulation Preliminary](image)

Figure 2: $Z\gamma$ Invariant Mass Simulation

Figure 3 on page 6 shows the total signal efficiency as a function of $M_X$. The signal efficiency of all the selection cuts combined range from 30-50% for increasing $M_X$. 

5
Event samples are generated using POWHEG-BOX interfaced with Pythia, with $M_X$ ranges from 200 GeV to 2.4 TeV.

4.3 Background Simulation

There are two sources of background. The dominant background is the Standard Model production of a Z boson plus a photon, with smaller contributions from Z + jets with the hadronic jet misidentified as a photon.

The $Z\gamma$ invariant mass data points are plotted with the background-only fit, as shown in
Figure 4a on page 6. The solid blue line is the background prediction and black dots are the actual data. The background $Z\gamma$ mass spectrum is smoothly falling and can be parameterized with $f(x) \approx (1 - x^{1/3})p_1 x^{p_2}$ with $x = M_{Z\gamma}/\sqrt{s}$. Figure 4b on page 6 shows the local p-value of the $M_{Z\gamma}$ data with respect to the background-only hypothesis. By comparing the data and the background, the largest local deviation is around $M_x = 268 \text{GeV}$ where the local significance is 2.2 $\sigma$.

Upper limits on $\sigma(pp \to X) \times BR(X \to Z\gamma)$ are set using a profile likelihood method. Figure 5 on page 7 shows the observed (solid lines) and median expected (dashed lines) 95% confidence level limits on the product of the production cross section times the branching ratio for the decay to a Z boson and a photon of a narrow scalar boson $X$, $\sigma(pp \to X) \times BR(X \to Z\gamma)$. The green and yellow band indicate 1 and 2 standard deviation intervals. A modified frequentist (CLs) method is used to set upper limits on the product, by identifying the value of $\sigma \times BR$ for which CLs=0.05. The observed $\sigma \times BR$ limits vary from 88 fb at $M_X = 250 \text{ GeV}$ to 2.8 fb at $M_X = 2.40 \text{ TeV}$. There is not a significant deviation from the observed $\sigma \times BR$ limits.

4.4 Decay Polarizations

The identification of W/Z boosted jet is affected by $\Delta R(q\bar{q})$ distributions and hence the polarizations of the boson decay products. In order to quantify such decay, this paper defines a helicity frame and analyzes the $q\bar{q}$ decay angular polarizations in the frame.
Figure 6 illustrates the definition of the helicity frame. The helicity frame is the W/Z boson rest frame with its Z axis defined to be along the same direction as the W/Z boson decay from the rest frame of massive boson X. We can reach the helicity frame by performing two consecutive sets of rotation and boost from the proton-proton rest frame. In the helicity frame, the $q\bar{q}$ decay products are back to back due to conservation of momentum. This paper studies the $\cos \theta$ distribution to quantify the polarizations.

Theoretically, the $\cos \theta$ distributions are different for the different W/Z production channels. Figure 7 shows the expected functional forms of the decay angular distributions from Standard Model electroweak decay of the W boson.

![Helicity Frame Definition](image)

**Figure 6: Helicity Frame Definition**

**Figure 7: Functional forms of decay polarizations for different W/Z production channels**

- For $W$ production, $h = +1$, $\frac{dN}{d\cos \theta} \sim (1 - \cos \theta)^2$
- For $\bar{s}W$ production, $h = -1$, $\frac{dN}{d\cos \theta} \sim (1 + \cos \theta)^2$
- For $sW$ production, $h = 0$, $\frac{dN}{d\cos \theta} \sim \sin \theta^2 = 1 - \cos \theta^2$, Longitudinal Polarization
- For $\bar{s}W$ production, $h = \pm 1$, $\frac{dN}{d\cos \theta} \sim 1 + \cos \theta^2$, Transverse Polarization
(a) $\cos \theta$ Distributions for purely longitudinal (blue) and purely transverse (red) polarizations

(b) $\cos \theta$ Distributions for three channels at $M_X = 5\text{TeV}$

Figure 8: Theoretical $\cos \theta$ distributions and an example distribution

The $\cos \theta_{\bar{q}d}$ distributions are fitted to the functional forms of $A + B \cos^2 \theta = A(1 + \frac{B}{A} \cos^2 \theta)$. The fitting parameter $\frac{B}{A}$ is hence used to quantify the polarization. $\frac{B}{A} = 1$ indicates a purely transverse polarization whereas $\frac{B}{A} = -1$ indicates a purely longitudinal polarization. Figure 8a illustrates the distributions for purely longitudinal and purely transverse polarizations. Figure 8b shows an example of $\cos \theta$ distributions at $M_X = 5\text{TeV}$. It clearly demonstrates that the spin 0 and spin 2 productions are much more transversely polarized, while the spin 1 production is much more longitudinally polarized. The same fitting process is repeated for all $M_X$ slices and the $\frac{B}{A}$ parameters as a function of $M_X$ are plotted in Figure 9. Spin 0 and spin 2 productions are almost purely transversely polarized while the spin 1 production is almost purely longitudinally polarized.
We use fat jet selection to identify W/Z boosted jet. This is using a cone radius of $\Delta R_{q\bar{q}} < 1$, which is defined to be

$$\Delta R_{q\bar{q}} = \sqrt{\Delta \eta_{q\bar{q}}^2 + \Delta \phi_{q\bar{q}}^2}$$  \hspace{1cm} (1)$$

Figure 10 shows the $\Delta R_{q\bar{q}}$ distributions for two $M_X$ at 1TeV and 4TeV. We obtain good containment on the parton level. However, real data has extra inefficiencies due to presence of jets.
5 Standard Model Signal and Background

The signal of this thesis is the standard model $W + \gamma$ decay, whereas the background is the standard model $\gamma +$ jets decay.

5.1 Baseline and D2 Cuts

In ATLAS experiment, there is a set of baseline and a D2 variable cut applied as the first round of event selection. The baseline selections are:

1. To only use the detector region with the best detection: $E_{T}\gamma > 250 \text{GeV}$, $|\eta_\gamma| < 1.37$

2. To only use the detector region with the best detection: $P_{T}\gamma > 200 \text{GeV}$, $|\eta_{jet}| < 2$

3. To remove photon tracks disguised as jets: $\Delta R_{\gamma-Jet} > 1$, where $\Delta R_{\gamma-Jet} = \sqrt{\Delta \eta_{\gamma-Jet}^2 + \Delta \phi_{\gamma-Jet}^2}$

D2 is a variable specifically designed to differentiate single jet showers from di-jet showers. As illustrated in Figure 11, the D2 algorithm takes a unit radius circle in the $\varphi, \eta$ space of jet showers and calculates the variable as:

$$D_2 = ECF_3 \times \frac{ECF_1^3}{ECF_2^3} \quad (2)$$

The D2 distributions of the signal and background events are shown in Figure 12. The $ECF_2$ variable could be 0, in which case we designate the D2 to be -1.
Figure 11: D2 algorithm illustration.

\[ \Delta R = \sqrt{\Delta \eta_{qq}^2 + \Delta \varphi_{qq}^2} \]

Figure 12: D2 distributions in signal and background

(a) Signal
(b) Background

5.2 Background and Signal Features

The important evidence that any classification mechanics works is whether it could successfully retain the signal features while suppressing the background features. In our search for new resonances, invariant mass is apparently the most important feature to retain. Figure 13 shows the jet mass of the highest pt jet in the standard model $\gamma + \text{jets}$ decay background. Figure 14 shows the jet mass in the $W + \gamma$ signal.
Figure 13: Jet mass of highest Pt jet in background.

Figure 14: Jet mass of highest Pt jet in signal.

Other event features that we took into account to develop the multi variant analysis
method are:

1. $Pt$, $\eta$, $\varphi$, mass, $D_2$, number of sub-jets, $ECF_2$ and $ECF_3$ for the 3 highest $Pt$ jets.
2. $Pt$, $\eta$, $\varphi$ of the highest $Pt$ $\gamma$.

5.3 Features Preprocessing

We have altogether 27 features for each event to form our feature space. However, since photon jets removal removes some jets from events, it is sometimes the case that an event does not even have 3 jets. In that case, we manually assign pseudo feature parameters that are clearly outside the range of distributions of real jets. The reason to do this is because our multi variant method demands consistent features in all events. Figure 15 demonstrates a manual addition of fake jets. In our signal events, there is a large number of events without even a second jet. In those cases, we manually assign an $\eta$ value of -5, which is definitely outside the range of $\eta$ distributions of the real jets. By having manual fake jets addition, we provide the possibility that the different number of jets in events could also be indirectly utilized as a differentiation parameter, since the fake jets greatly distort the feature distributions.

![Figure 15: Manual fake jets addition.](image)

Besides manual fake jets addition technique, we recognize that different features vary greatly in ranges. For example, the $\eta$ variable typically is in the range of $\pm$4, while the mass of jets could go as high as 200GeV. However, the huge difference in ranges does not
mean that one feature is much more important than the other. Our assumption is that all features should be treated equally. Therefore we also adopted the feature normalization technique. For each feature column, we find the maximum and minimum feature and normalize the features as:

\[
\bar{x}_{i,j} = \frac{x_{i,j} - \text{min}_j}{\text{max}_j - \text{min}_j}
\]  

where \(x_{i,j}\) is the \(j^{th}\) feature in the \(i^{th}\) event. And \(\bar{x}_{i,j}\) is the normalized feature, \(\text{min}_j\) and \(\text{max}_j\) are the extremums of the \(j^{th}\) feature across all events, including both the signal and the background events. The normalization procedure preserves the distribution of features and restrict the range of values of all features to be from 0-1. This prevents our classification structure to be misled into thinking that mass of jet is million times more important than \(\eta\) of jets.

6 Traditional Box Cuts

7 Support Vector Machine

8 Random Forest

9 Neural Net

9.1 Structure and Terminology

Artificial neural net is widely used as a classification structure. The structure and working principles of a neural net is illustrated in Figure 16. A basic neural net comprises of three parts, the input layer, the hidden layers and the output layer. Each layer has a certain number of computation units called neurons. Each neuron is nothing more than a matrix multiplication and an addition. Upon taking an input vector, each neuron first calculates a weighted sum of the vector by multiplying the input vector by a weight vector according to

\[
a = w^T \bar{x}
\]  

where \(w\) is the weight vector and \(\bar{x}\) is the input feature vector. The scalar product \(a\) is then offset by a number \(b\). Therefore, the final output from each neuron is

\[
a = w^T \bar{x} + b
\]  

Both the weight vector and the offset constant are parameters obtained from the training of the neural net.

Take the neural net structure in Figure 16 as an example, the input vector to the input layer neurons is the feature vector, in our case, it is a 27 by 1 vector. Each one of the three
neurons in the input layer then calculates a different scalar output, based on what weight vector and offset constant they have.

Figure 16: Artificial neural structure.

This vector then undergoes a function called the activation function. There are many models of activation functions, one example is the ReLU as shown in Figure 17. In the ReLU activation function, if the number output from a single neuron (which is always just a scalar), is less than 0, the neuron will simply outputs 0. If the scalar is larger than 0, the neuron will output the scalar itself.

Figure 17: ”ReLU” activation function.

The result from the input layer is hence a 3 by 1 vector. This vector is then fed into each one of the four neurons in hidden layer 1. The result from the hidden layer 1 is hence
a 4 by 1 vector, which is then fed into hidden layer 2 to produce another 4 by 1 vector. The output layer eventually computes a scalar product of this 4 by 1 vector and offset it to obtain a single scalar number, which becomes the output of this neural net example. Take note that the application of activation function happens for all layers.

9.2 Methodology

Different neural net structures greatly affect the performances. In order to look for a good structure, I implemented a random structure test. The number of neurons in each layer is randomly chosen to follow Random structure tests. ROC, signal/background.

9.3 Test Set

The important metric to test the performance of the classification is to look at the test set jet mass distributions. The mixed signal and background data this thesis looks at is divided into three categories. The first category uses all the preprocessed features as described in section 5.3, and is labeled as "all". The second category uses only the high level features, and is labeled as "high_level". The last category uses all the features except the D2 variable, and is labeled as "no_D2". Since each category draws its test set randomly, the test set jet mass distributions vary to a small extent. Figure 18 shows the random test sample distribution in all three categories. The area under graph has been normalized to 1.
Figure 18: Unfiltered Jet Mass Distributions in 3 Categories

(a) Mass of highest Pt jet, all features
(b) Mass of highest Pt jet, high level features
(c) Mass of highest Pt jet, absent D2
Figure 19: Constant learning rate.
9.4 Results

9.4.1 All Features

Figure 20: 3-layer Filtered Jet Mass exp lr, all.
9.4.2 High Level Features

Figure 21: 3-layer Filtered Jet Mass exp lr, high level.
9.4.3 Absent D2

Figure 22: 3-layer Filtered Jet Mass exp lr, no D2.
10 Dynamic Learning Rate

10.1 Stair Case Decay

Figure 23: Stair case learning rate.
10.2 Exponential Decay

Figure 24: Exponentially decayed learning rate.
10.3 Triangular

![Evolution of learning rate](image)

Figure 25: Triangular learning rate.
10.4 Stochastic Gradient Descent with Warm Restarts (SGDR)

Figure 26: SGDR learning rate.

11 Deep Neural Net

12 Ensembles

13 Conclusion

14 Discussion

15 References

References
