Applications of Machine Learning Methods on the Classification of Higgs Boson in the ATLAS Experiment

Chengjing Gong cgong29@wisc.edu
Kenny Jia hjia38@wisc.edu
Sherry Yang xyang467@wisc.edu

Abstract

Higgs boson is one of the most important portal in not only precision physics in the Standard Model (SM) but also in new physics beyond SM. However, extracting the Higgs boson signal from different channel of backgrounds remains a challenge. We study how different machine learning (ML) algorithm improves the classification accuracy with data from the A Toroidal LHC ApparatuS (ATLAS) full-detector simulation.

1. Introduction

In this study, we have study applications of different ML methods on classifying the Higgs boson decays to two tau ($H \rightarrow \tau\tau$) signal and SM background process. Both are generated by the official Monte Carlo simulation of the ATLAS detector response [1]. For this short report, here is an introduction of the minimal knowledge needed here for audience without basic knowledge of particle physics and quantum field theory.

1.1. The ATLAS Experiment

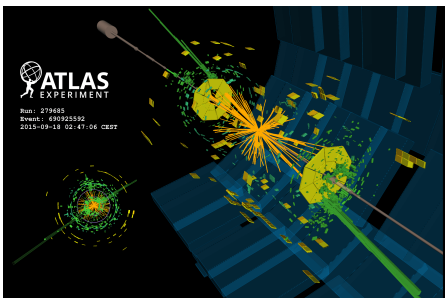

Figure 1. Visualization of a collision inside ATLAS

ATLAS is one of the general-purpose detectors in the Large Hadron Collider(LHC). At the center of the ATLAS detector, proton beams are accelerated to more than 99.999999% of the light speed and collide. One of its goal is to search for the Higgs boson. Here is a visualization of a real collision event in ATLAS 1.

1.2. Higgs Boson, tau, and the Standard Model

The Standard Model of particle physics (SM) is a theory that describes the known matter in terms of its elementary constituents and their interactions.

The Higgs boson is a elementary particle, the carrier particle for the Higgs field, a field which gives particles their mass. The more a particle interacts with the Higgs field, the higher its mass.

On 4 July 2012, the ATLAS collaboration at the LHC announced they had observed a new particle which is consistent with the description of the Higgs boson in SM [2]. The Higgs boson lifetime is $\sim 10^{-22}$ s. Consequently, we could only measured the physical properties of the final state particles, then reconstruct the Higgs boson from data. Tau is an elementary particle similar to the electron, but with much higher invariant mass.

2. Dataset summary

Our dataset is retrieved from the "Higgs Boson Machine Learning Challenge: Use the ATLAS experiment to identify the Higgs boson" on Kaggle [3]. We only use the "training" dataset since the official "test" dataset does not include the true classification label. We have choose a random sample of 10k events from 250k events as our primary dataset due to limit computing power and time.

Our primary dataset has an ID column, 30 feature columns, a weight column and a label column. Among these features, 17 are primitive features which are "raw" features from the simulation (physical properties measured), and 13 are derived features computed from the primitive features. Missing values have been set to -999 [1]. The label is binary ("s" or "b"), representing the event is either "signal" or "background". The weight column are an artifact of the way the simulation works, thus will be drop by us. One-hot encoding is applied on the label feature.

3. ML Methods Application

3.1. Feature Engineering

As shown in Figure 2, different features have large difference in ranges. In order to have a similarity measures for

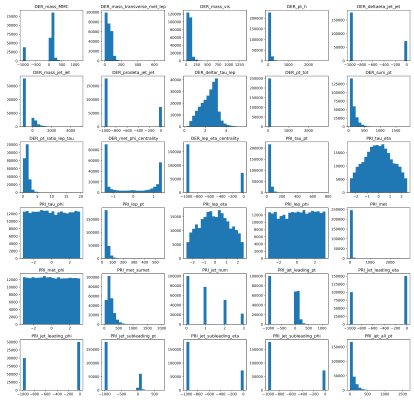

Figure 2. Histograms of 30 Features

all variables' contribution, we rescale these features using standardization. Then, we split the dataset 3 subsets consisting of 75%, 12.5%, 12.5%, as training, validation, and test dataset.

3.2. Learning Algorithm

We choose Random Forest, k-Nearest Neighbor (kNN), Gradient Boost Decision Tree, and Extreme Gradient Boost Decision Tree (XGBoost) as they outperforms others on large scale dataset.

3.3. Hyperparameter Tuning

For each algorithm, we use a grid search method for hyperparameter tuning with three fold cross validation. Best combination of hyperparameter along with its score are listed below:

- Random Forest: Accuracy = 0.83.

```
max_depth = 8
```

- k-Nearest Neighbor: Accuracy = 0.78.

```
n_neighbors = 45,
metric = 'manhattan'
```

- Gradient Boost Decision Tree: Accuracy = 0.84.

```
learning_rate = 0.1,
max_depth = 6,
n_estimators = 500
```

- XGBoost: Accuracy = 0.84.

```
learning_rate      = 0.01,
max_depth          = 7,
n_estimators       = 1000,
objective          = 'binary:logistic',
base_score         = 0.5,
booster            = 'gbtree',
colsample_bytree   = 0.7,
gamma              = 0,
min_child_weight   = 0.0001,
eval_metric        = 'rmse',
tree_method        = 'gpu_hist',
gpu_id             = 0,
subsample          = 0.8
```

3.4. Test on Validation Data

We test Accuracy and AUC for each algorithm on validation data, the results are listed below: As shown, XGBoost gives best performance among four algorithms with relative longer runtime.

Methods	Accuracy	AUC
Random Forest	0.84	0.87
kNN	0.77	0.84
GBDT	0.83	0.89
XGBoost	0.84	0.9

4. Result

XGBoost gives a complicated 7-depth model. As different decay modes of the tau pairs makes it harder to train with a single estimator, the classification efficiency is expected to improve with much more computing power, time, and feature engineering based on physics analysis.

5. Contributions

Kenny proposed the idea of this study. Sherry and Chengjing wrote the script for data reading, pre-processing, and the training of Random Forest, kNN, and GBDT. Kenny have done all hyperparameter tuning and add the XGBoost model.

6. Acknowledgment

We would like to appreciate the input from the instructor, Professor John Gillett and TA Jitian Zhao who both have provided constructive feedback on the report.

References

[1] Claire Adam-Bourdarios, Glen Cowan, Cecile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. Learning to discover: the higgs boson machine learning challenge - documentation, 2014.

[2] ATLAS Collaboration. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1–29, 2012.

[3] ATLAS collaboration. Dataset from the atlas higgs boson machine learning challenge 2014.