Discrimination of Signal-Background Events with Supervised and Semi-Supervised Learning in the Search for New Bosons Decaying to the $Z + \gamma$ Final State

Nkateko Baloyi[1,2], Bruce Mellado[2,3] and Xifeng Ruan[2]

[1] School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, Wits 2050, South Africa

[2] School of Physics and Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, Wits 2050, South Africa

[3] iThemba LABS, National Research Foundation, PO Box 722, Somerset West 7129, South Africa

E-mail: nkateko.baloyi@cern.ch

Abstract. The search for new particles beyond the standard model is vital for high energy physics experiments to answer fundamental questions concerning the laws of interactions and forces. This study compares supervised machine learning and semi-supervised machine learning performance in discriminating the signal and background events in the search for new bosons decaying into the $Z+\gamma$ final state. Boosted decision tree algorithm is employed in the supervised learning approach to discriminate signal and background events. The semi-supervised learning approach employs boosted decision tree following a weakly supervised learning approach to discriminate events from two different samples. The same classifier is then employed to discriminate signal and background events. The preliminary results show that the weakly supervised learning performance is similar to the performance achieved using supervised learning approach.

1. Introduction

The Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN) collides particles at extremely high energy and high luminosity. With the large luminosity accumulated by the LHC, new range of discoveries are possible, allowing the search for new physics beyond the Standard Model (BSM). The complex and high dimensional data collected in the LHC requires more advanced techniques with the potential to extract information produced during proton-proton (*pp*) collision. Statistical tools and advanced techniques, such as machine learning (ML) can be used to improve data processing, classification and regression tasks in particle physics data and directly improves the probability of discovering new particles. ML is developed with the concept of allowing computers to learn by themselves and improve efficiency with experience without being explicitly programmed and is applied in various professional fields and industries to improve the processing and the quality of results by finding patterns in data that lead to accurate decision making [1, 2, 3, 4, 5].

Various ML algorithms have been applied in binary classification tasks and have achieved

good accuracy. Boosting techniques and deep learning techniques have gained popularity in classification and regression tasks due to their ability to achieve better accuracy when applied on weak learning algorithms and trained on huge amount of data, respectively. Boosted decision trees and deep neural networks have been applied widely in different professions, including high energy physics data, and they have proven to be superior in terms of accuracy [2, 6].
This proceedings aims to highlights machine learning approaches that can be used to discriminate signal and background events on simulated data and help develop a model for real data. We develop two machine learning models following the supervised learning and the semi-supervised learning approach. Supervised learning trains a model on labeled data, where the model learns a function that maps the input (Variables) to the output (target labels). Semi-supervised learning is a combination of supervised and unsupervised learning, where labeled and unlabeled instances are used to train a model.

2. Data

This study employs the Monte Carlo data sample recorded from 2015-2017 in the pp collision at $\sqrt{s}$= 13 TeV with two different types of samples following the 2015+2016 data pileup distribution normalized to the luminosity of 36.1 fb^{-1} and the 2017 data pileup distribution normalized to the luminosity of 43.8 fb^{-1}, the mc16a and the mc16d data sample respectively. The signal data sample consist of Z boson produced from gluon-gluon fusion sample, real missing transverse energy (E_T^{miss}) from neutrinos produced by Z decay, and the heavy pseudo-scalar particle decaying to Z boson and heavy scalar sample. Missing energy is the energy which is not detected in a particle detector but is expected to be there due to the laws of physics and momentum, obtained from the negative vector sum in the transverse plane of the momenta of all particles detected. The background data is a mixture of the $tt\gamma$ and $\gamma -$ Jet background samples. The background contains fake E_T^{miss}. The data sample for the supervised approach consist of labeled signal (S) and background (B) events with 11730 S events and 151534 B events decaying to the $Z+\gamma$ final state. The second data sample used in this study consists of two different samples of unlabeled data. The first sample is the Monte Carlo events collected from 2015 to 2017 in the pp collision that consists of the $Z+\gamma$ events with the invariant mass outside 240-280 GeV. The second sample consist of the B mass-window events, between 240 GeV and 280 GeV, of the $Z+\gamma$ and the BSM monte carlo events. The first sample is made up of only B events, while the second sample consist of mixed S and mass-window B events. To evenly distribute the mixture of events in the second sample, we shuffle the sample to avoid having an entire minibatches of highly correlated examples. Table 1 below is the description of the variables used in this study. The variables used include two leptons, a photon and multiple jets.

Variable	Description
mu	Average bunch crossing
dphifjmet	The angular distance $\Delta(\phi)$ between forward jets and MET
dphisjmet	The angular distance $\Delta(\phi)$ between soft jets and MET
ssumpt2	Subleading vertex sum P_t squared
djpt	Scalar difference between the vectorial sum P_t of all the jets and leading vectorial sum P_t
dsumpt2	Difference between leading and subleading vertex sum P_t squared
dphirefjetmet	The angular distance $\Delta(\phi)$ between vectorial sum P_t of all jets and MET

Table 1. Input variables description

3. Methodology

The data sample is divided into the low (Low) category and intermediate (Int) category with respect to the transverse missing energy significance ($S_{E_T^{miss}}$). The Low category falls within the range $2.5 \leq S_{E_T^{miss}} \leq 3.5 \quad \sqrt{GeV}$ and the Int category is within the range $3.5 \leq S_{E_T^{miss}} \leq 5.5$ $\sqrt{GeV}$. The pre-selection cuts are applied to maximize S and minimize B, where the cut points are scanned regions of S to B efficiency.

This study implements two different approaches, the supervised learning approach and the semi-supervised learning approach. For the supervised learning approach, after the pre-selection cuts are applied, the data sample for each category is split into 70% training and 30% testing set. The training set consists of 114284 events with only 8211 S events and 106073 B, and the test set contains 48980 events with 3515 S events and 45461 B, before applying the pre-selection cuts. The Low category consists of 1336 S and 7875 B training events, and 529 S and 3369 B test events. The Int category consists of 3486 S and 3996 B events in the training set and only 1744 B and 1524 S events in the test set. The test set split before any pre-processing and saved on a different file in order to use the same test set for both the supervised and semi-supervised learning approaches. The training set is split into 80% training and 20 % validation in both categories. Synthetic minority oversampling technique (SMOTE) is applied on the training set of each category independently, to reconstruct the S events, thereby balancing the S to B ratio in the training set. SMOTE creates synthetic observations by finding the nearest neighbors of each instance of the minority class based on the euclidean distance between the instances in feature space. The distance is multiplied by a random number between 0 and 1 to create a new (synthetic) observation. The S is oversampled by creating synthetic observations to avoid training the classifier on skewed category distribution, to allow learning to be feasible and reduce bias in the classifier decision making. SMOTE is only applied on the training set to avoid having synthetic observations on the test set. Using K-fold cross validation and GridSearchCV, a process of performing hyper parameter tuning in order to determine optimal values, we find the optimal hyper parameters for each category since both categories consists of different number of events selected based on the $S_{E_T^{miss}}$. A 3-fold cross validation is used to first find the optimal maximum depth. The optimal maximum depth is then fixed, and a GridSearchCV hyper-parameter tuning is used to apply all the possible combinations of parameters provided through a list of dictionaries to find the optimal parameters to build a BDT classifier. The validation set is used to evaluate the performance of the classifier while fine tuning the parameters, to ensure there is little-to-no overtraining observed from the training and validation accuracy before applying it to the test set. An optimized Boosted decision tree (BDT) is trained on the labeled training set to discriminate the S and B events by learning a function that maps the variables to the target class (S or B). The test set is used to measure and evaluate the performance of the classifier using some performance metric.

The semi-supervised learning approach employed in this study is called the weakly supervised learning approach. This approach employs the unlabeled data samples of B events and the mixed S and B events sample. The two samples are assigned weak labels, where the B only events in sample 1 ($M1$) are labeled 0 and the mixed S and B events in sample 2 ($M2$) are labeled 1. The samples are split into 80% training and 20% validation. We use GridsearchCV and K-fold cross validation to find the optimal hyper parameters. An optimized BDT classifier is trained to discriminate events from $M1$ and $M2$. The same BDT classifier trained to discriminate $M1$ and $M2$ events, is used to discriminate the S and B in the test set used in the supervised learning approach. A Study in [7] shows that a weakly learning classifier trained to discriminate events from different samples can be employed to discriminate the S and B and still performs just as well as the supervised learning classifier.

The optimal hyper parameters used for the supervised and semi-supervised learning approach for both the Low category and Int category are listed in Table 2.

	Low category		Int category	
Supervised BDT	Hyper parameter	Sample Distribution	Hyper parameter	Sample Distribution
	N-estimators	700	N-estimators	800
	Max-Depth	3	Max-Depth	3
	Learning rate	0.01	Learning rate	0.01
	Subsample	0.75	Subsample	0.75
Semi-supervised BDT	Hyper parameter	Sample Distribution	Hyper-parameter	Sample Distribution
	N-estimators	900	N-estimators	1200
	Max-Depth	3	Max-Depth	3
	Learning rate	0.01	Learning rate	0.01
	Subsample	0.75	Subsample	0.75

Table 2. Supervised Learning and Semi-supervised Learning Hyper parameters

4. Results and Discussion

The performance of the BDT classifier is measured using a receiver operating characteristic curve and the distributions of the train and test output. Accuracy is not a reliable metric for this study since the data is imbalanced. The supervised learning BDT output distributions shown in Figure 1 shows how well the classifier can discriminate S and B events on the Low category. The Distribution on the left demonstrate the S and B discrimination achieved when training the classifier, and the distribution on the Right, demonstrate the S and B discrimination achieved on unseen test set. The blue distribution shows the B and the red distribution is the S. Both the train and test distributions show that the classifier achieved a good discrimination with the B shifted towards the Left and S shifted towards the Right of the plot. The train distributions shows a slightly better discrimination when compared to the test distributions.

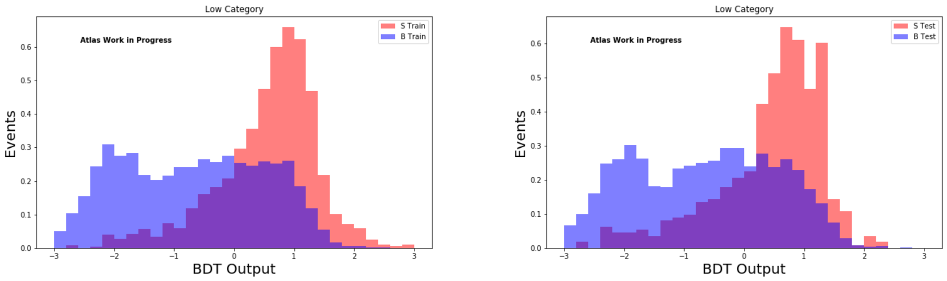

Figure 1. Low Category Train (Left) and Test (Right) Distributions

The distributions in Figure 2 show that the Int category classifier also achieves a good discrimination capacity on both the training and testing distribution, with the the testing distribution consistent with the shape of the training distribution. However, the training achieves better separation as compared to the test, which implies that the classifier is slightly overtraining.

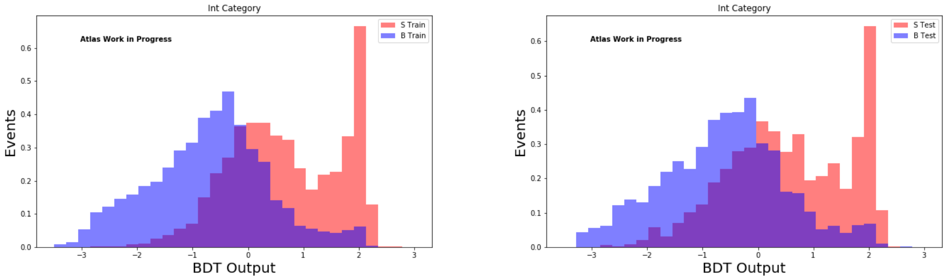

Figure 2. Int Category Train (Left) and Test (Right) Distributions

Both the Low category and Int category can generalize well when given unseen test set, however, both classifiers misclassifies some of the events, demonstrated by the overlapping of the S and B events. The test output distribution achieved by both the Low and Int categories, are similar to the output distributions of the classifier during training. A cut can be applied on the distributions to maximize S and minimize B without loosing a lot of S events.

The receiver operating characteristic (ROC) curve in Figure 3, is a plot of background rejection against signal efficiency, which measures how well the classifier can discriminate between the S and B events. The black dotted line is the non-discriminatory line, the red and blue curves represents the Int category and Low category, respectively. The non-discriminatory represent the behavior of a random classifier. The region below the non-discriminatory represent the performance of a poor classifier and the region above the non-discriminatory represent the performance of a good classifier, with a very good classifier extended towards the top right corner. The curves for both the Low category and Int category are extended towards the upper right, showing that the classifier is correctly classifying the test set with less uncertainty in the results, thereby maximizing background rejection. The BDT classifier has a significant performance from the non-discriminatory line, with both the Low category and Int category classifiers achieving area under the curve (AUC) of 79%, which represents the discrimination capacity. Both the classifiers achieved 79% discrimination capacity on test set, with 21% likelihood of misclassifying the events. This means that when the classifier is given new test set, it can correctly classify 79% of the events and misclassify 21% of the events, as seen from the output distributions demonstrated by the overlapping of the S and B events.

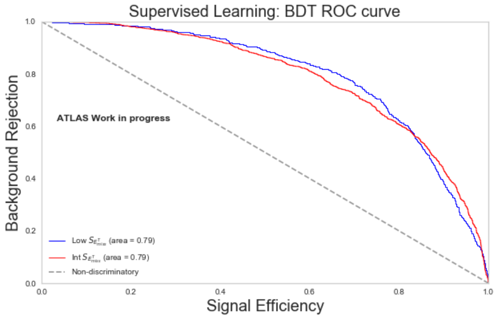

Figure 3. Supervised learning ROC curve. The red curve and blue curve represents the performance of the classifier on the Int category and Low category, respectively

The output distributions in Figure 4 and Figure 5 represent the weakly supervised learning BDT training output of $M1$ and $M2$ (Left) and the performance of the BDT classifier when predicting the test S and B (Right). The blue and red distributions on the Left are the $M1$ train and $M2$ train, respectively. The blue and red plots on the Right distributions shows the how well the classifier performed when discriminating B and S, respectively. The Low category output distributions in Figure 4 show that the classifier could slightly discriminate $M1$ and $M2$ during training, however, the performance of the classifier improved when same BDT classifier is used to discriminate S and B.

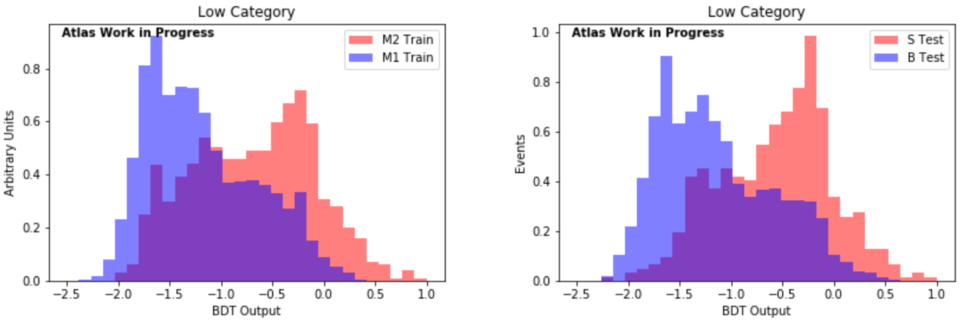

Figure 4. Weakly supervised BDT output for the Low category

The Int category output distributions in Figure 5 below show that the BDT classifier achieved a good separation when trained to discriminate $M1$ and $M2$, and achieved similar performance when given new test set of S and B events. The S and B output distributions on both the Low and Int categories shows that a classifier can be trained on weak labeled samples, and then used to discriminate S and B events and still achieves performance similar to that of supervised learning. The distributions show that a cut can be applied on the S and B distributions to maximize the S and minimize B without loosing a lot of S events.

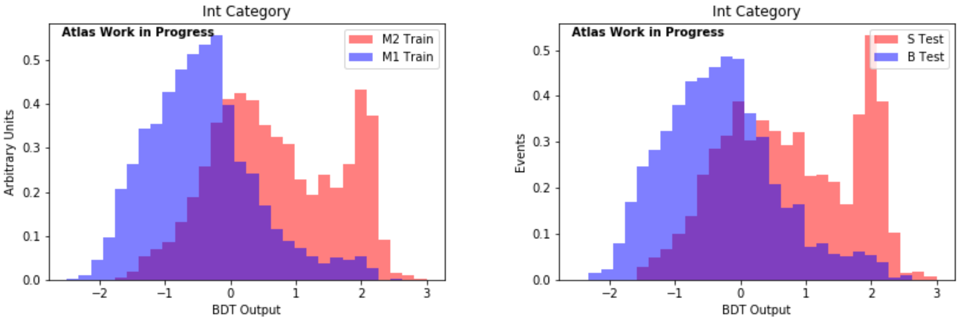

Figure 5. Weakly supervised BDT output for the Int category

The ROC curve in Figure 6 shows the performance of the weakly learning classifier when classifying the S and B events. The same weakly learning classifiers trained to discriminate $M1$ and $M2$ events on the Low category and Int category are applied to discriminate the S and B events from the supervised learning test set. The ROC curves show that the weakly supervised learning approach performs well in classifying S and B on the Low and Int categories, with AUC of 78% and 78%, respectively. The ROC AUC show that the two classifiers can achieve 78% discrimination capacity on unseen test set, which means there is a 22% likelihood of misclassifying the S and B events.

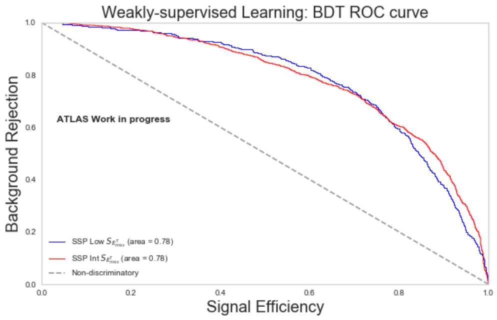

Figure 6. Weakly supervised ROC curve

The output distributions and ROC curves on the Low and Int categories show that the classifiers can discriminate S and B events with 79% and 78% discrimination capacity using supervised learning and weakly supervised learning, respectively. This means that a cut can be applied where there is maximum B and less S to disregard the B without loosing a lot of S events. An ideal discrimination capacity means that when a cut is applied to maximize S and minimize B, only insignificant amount of S statistics will be lost and less B events overlapping the S region, thereby improving the purity of the S. The good performance achieved with the weakly supervised learning shows that the approach can perform almost as good as the supervised learning approach. However, the good performance may be due to the weakly supervised learning classifier looking at the statistical fluctuations of the S events in $M2$, thereby classifying correctly new S events. This means that with a significant amount of S events in $M2$, the classifier may learn to discriminate $M1$ and only S in $M2$, since both B events in $M1$ and $M2$ share similar statistics. To further check if the performance of the weakly supervised classifier is similar to the performance of the supervised learning, we will inject 1/10 of the S on the mass-window B. This way, the signal is insignificant and the classifier will learn to discriminate the mass-window in $M2$ events from the side-band B in $M1$. A good performance with a small amount of S injected, will prove that the weakly supervised learning approach performance matches the supervised learning approach performance.

5. Future Work

To improve the performance of the semi-supervised approach, deep learning algorithms will be applied to cluster the S and B events. Deep neural network will be implemented to train the weak labeled samples, and the same classifier trained on weak labeled samples will be applied to discriminate signal and background from the supervised learning test set. Four vectors of the particles will be added as variables to improve the performance.

Acknowledgement

Gratefully acknowledges the funding I received from DST-CSIR, NRF and SA-CERN. I would also like to extend my sincere gratitude to the NETTP for their constant support.

References

[1] Aad G, Butterworth J, Thion J, Bratzler U, Ratoff P, Nickerson R, et al. The ATLAS experiment at the CERN large hadron collider. Jinst. 2008;3:S08003.

[2] Baldi P, Sadowski P, Whiteson D. Searching for exotic particles in high-energy physics with deep learning. Nature communications. 2014;5:4308.

[3] Katal A, Wazid M, Goudar R. Big data: issues, challenges, tools and good practices. In: 2013 Sixth international conference on contemporary computing (IC3). IEEE; 2013. p. 404–409.

[4] Michie D, Spiegelhalter DJ, Taylor C, et al. Machine learning. Neural and Statistical Classification. 1994;13.

[5] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016.

[6] Roe BP, Yang HJ, Zhu J, Liu Y, Stancu I, McGregor G. Boosted decision trees as an alternative to artificial neural networks for particle identification. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. 2005;543(2-3):577–584.

[7] Metodiev EM, Nachman B, Thaler J. Classification without labels: Learning from mixed samples in high energy physics. Journal of High Energy Physics. 2017;2017(10):174.