Searching for the Higgs Boson Particle using Data Analytics

Ananya Choudhury*
Dept. of Computer Science
Virginia Tech
Blacksburg, U.S.A
ananya@vt.edu

Vivek Bharath Akupatni†
Dept. of Computer Science
Virginia Tech
Blacksburg, U.S.A
vivekb88@vt.edu

Vignesh Adhinarayanan‡
Dept. of Computer Science
Virginia Tech
Blacksburg, U.S.A
avignesh@vt.edu

ABSTRACT
Abstract goes here

1. INTRODUCTION
Physicists use high-energy particle colliders such as the Large Hadron Collider (LHC) to produce new particles in order to understand the elementary structure of matter. In these colliders protons collide at high speed to produce several interesting particles which quickly decays into other stable particles. The particles produced in these experiments are identical in many aspects but differ in their kinematic features. Of particular interest in the Higgs-Boson particle which aids in understanding the fundamental structure of matter. Studying this particle is expected to provide insight into the functioning of universe and its possible fates. While the âĂIJreal worldâĂİ impact of the particle is not known yet, the scientific impact is so huge that billions of dollars has been invested in constructing colliders that produce this particle. Identifying the Higgs boson particle reliably and efficiently enables solving the bigger problem of studying the particle. One fundamental limitation of using magnetic detectors is their inability to distinguish signal (Higgs-Boson producing collision) from background noise (other collisions) when the signal-to-background ratio is too low. An alternative approach is to measure the properties of stable particles and infer the decaying particle originally produced by collision. For this purpose, scientists have created models of Higgs-Boson and other particles, simulate their decay using these models, and record the kinematic properties of the stable particles produced at the end. Based on these kinematic properties, they can create models to find the originally produced particles. This model can then be used in real accelerators for real-time detection of Higgs-Boson particle. Based on the above, we formally state our data mining task as follows. **Data Mining Task.** The problem of interest is classification. Given the kinematic features of stable particles produced in a collision, we classify it as signal (decays from Higgs-Boson) or background (otherwise). Given the rarity in occurrence of signal events and their importance in understanding the nature of matter, the primary goal is to identify most signal events correctly. Secondarily, since millions of events occur every second, we are also interested in identifying techniques that maximize the number of classifications per second. Our major contributions are as follows:

1. Implementation of several classifiers covering the following broad categories of classifiers for Higgs-Boson detection: Bayesian, tree-based, instance-based, deep learning, ensemble methods, meta-classifiers.
2. Comparison of classification accuracy for the different classification schemes.
3. Evaluation of throughput (Number of classifications per second) for the different classification schemes.

Our contributions help in identifying the appropriate classifiers for the given task and provides clarification on the complexity of features necessary for classification. Our major findings are as follows: <Fill in interesting results here> The rest of the paper is organized as follows. Section II provides the necessary background in theoretical physics. Section III describes the related work. Section IV describes our approach. We present our results in Section V and conclude in section VI.

2. BACKGROUND
Physics background

3. RELATED WORK
Detecting exotic particles in high-energy physics (HEP) using data analytics techniques instead of the traditionally used physical detectors [14] is not new. Cutts et al. are among the first to use neural networks to identify interesting events in HEP experiments [7]. This was quickly followed by several attempts in improving the classification accuracy of neural networks [11, 13]. Apart from the widely popular techniques based on neural networks, only decision trees were explored for a long time. Bowser-Chao and Dzialo used binary decision trees wto detect top quarks and compared

*Ananya worked on.
†Vivek worked on.
‡Vignesh worked on.

their results with neural networks [3]. The conventional wisdom was that neural networks were by far the best when it comes to classification in HEP until Roe et al. came along and projected boosted decision trees as an alternative to artificial neural networks [16]. By combining several *weak* classifiers, Roe et al. showed that it is possible to obtain better accuracy than a neural network. However, more recently, Baldi et al. showed that a deep learning neural network with several hidden layers outperforms the boosted decision tree [2]. While a number of classification techniques have been developed and applied over the last several years, the HEP community has so far explored only neural networks and boosted decision trees in any depth. This lead to the development of statistical packages for the HEP community such as StatPatternRecognition so that several other techniques based on Rotation Forest, Discriminant Analysis etc. could be explored [12]. Despite this effort, no documented work exists in HEP where alternative techniques are explored even though other techniques are thought to be inferior.

Since Baldi et al. [2] work is closest to ours, we describe it in detail here. The authors in this paper used deep learning methods of neural network to find exotic particles in high energy particle colliders. Deep learning models are neural networks with multiple hidden networks. Current technics like shallow models which are single hidden layer feedforward network fail to capture all features. The deep learning model here is used on 2.6 million training examples and 100,000 validation examples. The mode is a five-layer neural network with 300 hidden units in each layer, learning rate of 0.05, and a weight decay coefficient of 0.00001 . Testing is done on 500,000 examples. For Higgs benchmark, Area Under the Curve(AUC) - complete for deep neural network is 0.88 and for shallow neural network = 0.81.

4. METHODOLOGY

4.1 Data Preprocessing

In some situations, it becomes difficult to measure the physical properties of a particle such as momentum and energy accurately. In our original dataset, as many as 177000 out of 250000 instances had missing attributes. We considered three approaches to deal with missing values. First, we tried ignoring instances with missing attributes. But this resulted in very few useful instances. Second, we tried to replace the missing values with the mean/median. However, this biases the experiments. Finally, we decided to adopt a method known as multiple imputations [**?**] which replaces the missing values with a random number that follows the distribution for that attribute. We use the `Amelia` [9] package to perform this task.

4.2 Feature Engineering and Selection

The raw dataset includes 17 features. From these 17 basic features, 13 additional features were derived. These 13 derived features describe some property of the particle and requires knowledge of physics. The features are provided by physics. Their description is given in the appendix. From these 30 features, a subset is selected based on the following factors. First, we decide the ability of the feature/attribute to distinguish between signal an background. The distribution of different attributes is shown in Fig. 1 for signal and background separately. Second, we try to avoid using features correlating with each other while building the classifier. Fig. 2 shows the correlation matrix of the features. The final set of features selected differs for each classifier. The details are given in their respective subsections.

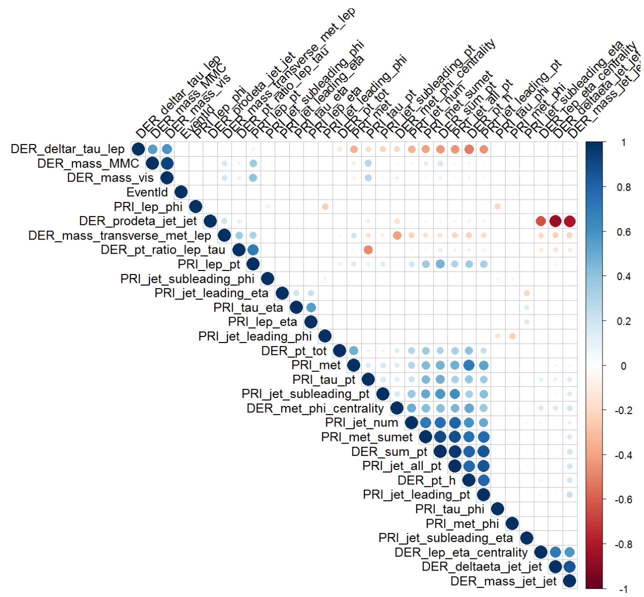

Figure 2: Correlation matrix. Dark blue indicates features that shows strong positive correlation. Dark red indicates features that show strong negative correlation.

4.3 Data Analytics Techniques

In this section, we describe the various classification schemes we explored. For each technique, we also describe the parameter settings explored in brief.

4.3.1 *Bayesian Classifiers*

Naive Bayes. This classifier is based on the Naive Bayes technique developed by John et al. [10]. In this technique <describe technique here>. <Describe parameter tuning here>.

4.3.2 *Functions-based Classifiers*

Logistic Regression. This classifier is based on the Ridge estimation technique developed by Cessie et al. [6]. In this technique <describe technique here>. <Describe parameter tuning here>.

Linear Discriminant Analysis

Quadratic Discriminant Analysis

4.3.3 *Tree-based Classifiers*

Simple CART. This classifier is based on the classification and regression trees (CART) technique developed by Brieman et al. [5]. In this technique <describe technique here>. <Describe parameter tuning here>.

4.3.4 Instance-based Classifiers

k-Nearest Neighbor. This classifier is based on the k-nearest neighbor (kNN) technique by aha et al. [1]. In this technique <describe technique here>. <Describe parameter tuning here>.

4.3.5 Deep Learning

4.3.6 Meta Classifiers and Ensemble Methods

Classification via Clustering and Regression. We simply cluster the raw dataset and mark certain clusters as signal and others as noise. Prediction based on the distance of the new data point to the centroid of the two cluster groups.

REP Tree with Bagging. Bagging technique developed by Brieman involves creating several models using different subsets of the training dataset [4]. Each model does its own classification and they all vote with equal weight to decide on the class.

Decision Stump tree with ADA Boosting. Based on ADA Boosting developed by Freund and Schapire [8]. Used in conjunction with Decision Stump Tree classifier.

Decision Stump tree with Multi-Boosting. Based on the MultiBoosting technique developed by Webb [17].

Rotation Forest. Rotation Forest is an ensemble method developed by Rodriguez et al. [15] where several classifiers are combined to produce a highly accurate classifier.

5. RESULTS

5.1 Results

Bayes Classifiers

Rule-based Classifiers

Tree-based Classifiers

Instance-based Classifiers

Deep Learning

Ensemble Methods

5.2 Discussion

6. CONCLUSION

Conclusion goes here

7. REFERENCES

[1] D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. *Mach. Learn.*, 6(1):37–66, Jan. 1991.

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

[3] D. Bowser-Chao and D. L. Dzialo. Comparison of the use of binary decision trees and neural networks in top-quark detection. , 47:1900–1905, Mar. 1993.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984. new edition [**?**]?

[6] S. L. Cessie and J. C. V. Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):pp. 191–201, 1992.

[7] D. Cutts, J. Hoftun, A. Sornborger, R. Astur, C. R. Johnson, and R. T. Zeller. The use of neural networks in the d0 data acquisition system. *Nuclear Science, IEEE Transactions on*, 36(5):1490–1493, Oct 1989.

[8] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, pages 148–156, San Francisco, 1996. Morgan Kaufmann.

[9] J. Honaker, G. King, and M. Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011.

[10] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338–345, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[11] L. Lönnblad, C. Peterson, and T. Rögnvaldsson. Finding gluon jets with a neural trigger. *Phys. Rev. Lett.*, 65:1321–1324, Sep 1990.

[12] I. Narsky. StatPatternRecognition: A C++ Package for Statistical Analysis of High Energy Physics Data. *ArXiv Physics e-prints*, July 2005.

[13] C. Peterson. Track finding with neural networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 279(3):537 – 545, 1989.

[14] J. Pinfold. Searching for exotic particles at the {LHC} with dedicated detectors. *Nuclear Physics B - Proceedings Supplements*, 78(1âĂŞ3):52 – 57, 1999. Advanced Technology and Particle Physics.

[15] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, 2006.

[16] B. P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research A*, 543:577–584, May 2005.

[17] G. I. Webb. Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, Vol.40(No.2), 2000.

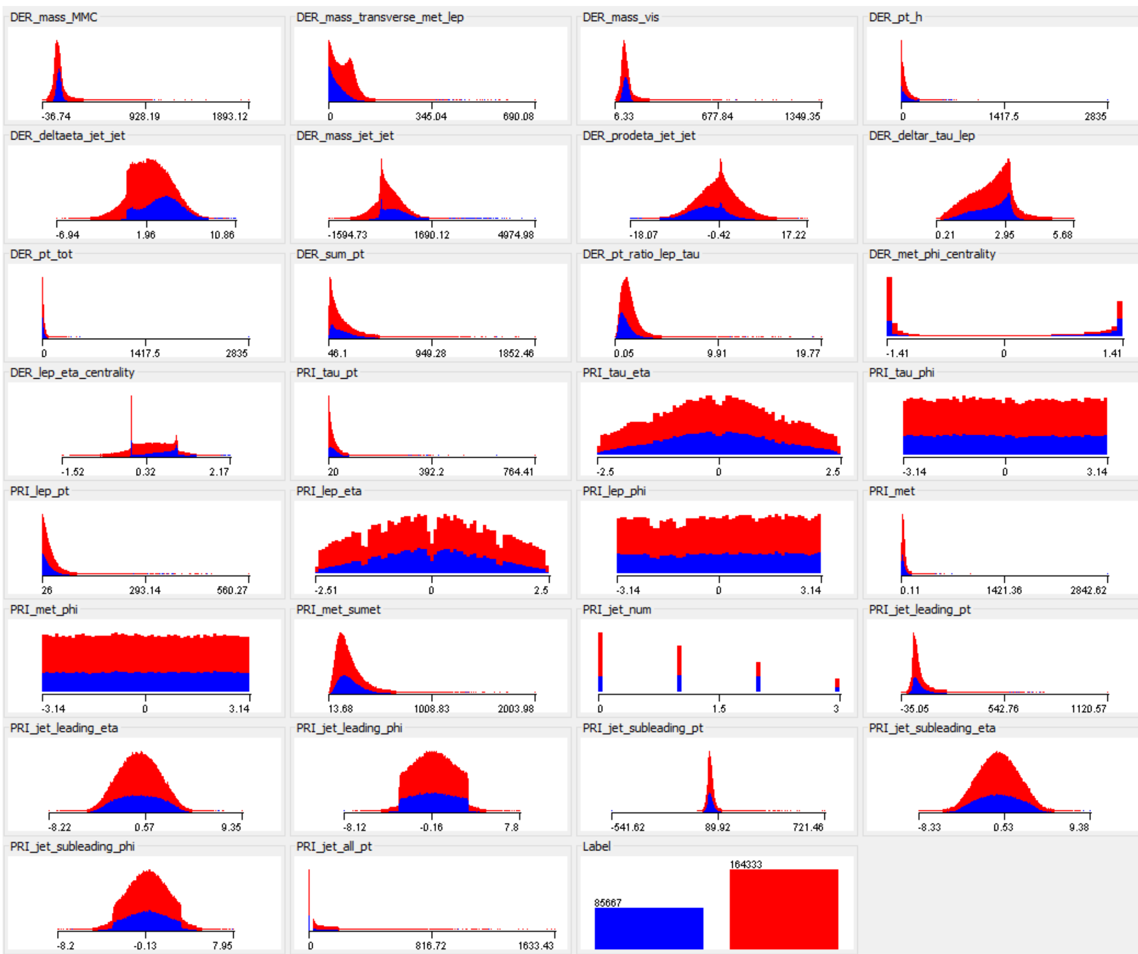

Figure 1: Distribution of different attributes for signal. Blue represents signal and red represents background.