Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 170 (2020) 1141–1146

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Statistical Methods and Artificial Intelligence
(IWSMAI 2020)
April 6-9, 2020, Warsaw, Poland

Higgs Boson Discovery using Machine Learning Methods with Pyspark

Mourad Azhari[a,*], Abdallah Abarda[b], Badia Ettaki[a,c], Jamal Zerouaoui[a], Mohamed Dakkon[d]

[a]*Laboratory of Engineering Sciences and Modeling, Faculty of Sciences- Ibn Tofail University, Campus Universitaire, BP 133, Kenitra, Morocco*
[b]*Laboratoire de Modélisation Mathématiques et de Calculs Economiques, FSJES, Université Hassan 1er, Settat, Morocco*
[c]*Laboratory of Research in Computer Science, Data Sciences and Knowledge Engineering, Department of Data, Content and knowledge Engineering School of Information Sciences Rabat, Morocco*
[d]*Département de Statistique et Informatique de Gestion, Université Abdelmalek Essaadi,Tetouan, Morocco*

Abstract

Higgs Boson is an elementary particle that gives the mass to everything in the natural world. The discovery of the Higgs Boson is a major challenge for particle physics. This paper proposes to solve the Higgs Boson Classification Problem with four Machine Learning (ML) Methods, using the Pyspark environment: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) and Gradient Boosted Tree (GBT). We compare the accuracy and AUC metrics of those ML Methods. We use a large dataset as Higgs Boson, collected from public site UCI and Higgs dataset downloaded from Kaggle site, in the experimentation stage.

Keywords: Boson Higgs; Spark;Pyspark; Machine Learning (ML); Logistic Regression (LR); Decision Tree (DT); Random Forest (RF); Gradient Boosted Tree (GBT); AUC and Accuracy

1. Introduction

The discovery of Higgs Boson is a major challenge for the High Energy Physics area. Machine Learning methods are proposed to solve the separation problem of signal from background events. For explanation, the signal is the decay of Exotic Particles, a zone in feature space that is not explained by the background processes. The background consists of disintegrating Particles that have previously been discovered in precedent experiments. Many works relevant to Higgs Boson search have advanced in the particle physics domain. ATLAS detector physicists, as one of the major

[*] Corresponding author. Tel.: +212694324310
E-mail address: azharimourad@yahoo.fr

10.1016/j.procs.2020.03.053

experiments at the Large Hadron Collider (LHC) at the European Organization for Nuclear Reseach(CERN), tested the predictions of the Standard Model. In the context of Higgs Boson Machine Learning Challenge included in 2014 [1], 12 algorithms were implemented: Naive Bayes, Decision Tree, K-Nearest Neighbors, KMeans Clustering, Spectral Clustering, Support Vector, Machines Affinity Propagation, AdaBoost, Gradient Boosting, Bagging, Random Forest, and Gaussian Mixture Models. The classification accuracy has achieved 84% with Gradient Boosting Classifier and Support vector machines with linear kernels achieved an accuracy of 76%. Tianqi and Tong advanced a regularized version of the Gradient Boosting algorithm with a highly efficient implementation [2]. Baldi et al have implemented the Deep Networks Classifier on Higgs and Susy datasets [3, 4, 5]. Alves used Stacking Machine Learning classifiers to identify Higgs Bosons at the LHC, in a multivariate statistical analysis (MVA), outperforms Boosted Decision Drees and Deep Neural Network application in particle physics[6]. The following paper proposes to solve the Higgs Boson classification problem with four Machine Learning (ML) Methods with Pyspark environment: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) and Gradient Boosted Tree (GBT). We compare the accuracy and AUC metrics of those ML Methods using 11 million instances of Higgs dataset UCI and Higgs dataset Kaggle.
This paper is organised as follows: The 1st section present a related works relevant to exotic particle classification problem. While the 2nd section advances the spark environment. The 3rd section proposes Machine Learning Methods to detect Higgs particle. The 4th section describes Higgs dataset. In the end, the 5th section presents our experimental results and their consequent analysis.

2. Spark Environement

Apache Spark is a powerful tool of Big Data and an open-source distributed cluster-computing framework. It includes a common Machine Learning (ML) library (MLlib) designed to ML classifiers. Spark runs well in memory and supports many programming languages (java, scala, python, SQL, R) and works well with Python.

2.1. Components of Spark Ecosystem

Spark ecosystem constitutes of Spark core component, Spark SQL, Spark Streaming, Spark MLlib, Spark GraphX, and SparkR. In this work, we use Spark MLlib. MLlib is a scalable Machine Learning library that contains Machine Learning libraries with the implementation of various Machine Learning Algorithms [7, 8].

2.2. Features of Apache Spark

The important features of Apache Spark are [9]:

- Rapid Processing: Apache Spark presents high data processing swiftness(about 100x faster in memory and 10x faster on the disk);
- Dynamic: We can develop a parallel application in Spark(80 high-level operators available);
- In-Memory Computation: We can improve the processing speed in-memory;
- Reusability: We can easily reuse spark code for batch-processing.

3. Proposed Methods

This paper have applied different algorithms to this Higgs classification problem: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) and Gradient Boosted Tree (GBT).

3.1. Logistic Regression (LR)

Logistic Regression is a particular case of Generalized Linear Models (GLM) that predict a categorical response . In spark.ml logistic regression module, we predict a binary outcome using Binomial Logistic Regression [10].

3.2. Decision Tree (DT) Method

Decision Tree Method is an algorithm used to create a training model to predict classes or values of target variables by deduction of the Learning Decision rules from the training dataset [11, 12]. Decision Tree presents a simple visualization of results and it is a basic predictor for the Bagging model [13, 14].

3.3. Random Forest (RF)

Random Forest is a Machine Learning classifier that associates the Bagging classifier and a decision tree algorithm. This method reproduces a set of Decision Trees from a randomly chosen subset of training sampling. Then, it regroups the votes from various Decision Trees to detect the final class of the test item [15, 16, 17, 18]. Random Forest algorithm runs into two stages: The creation and prediction [19].

3.4. Gradient Boosted Tree (GBT)

Gradient Boosted Tree (GBT)is an ensemble method that trains iteratively Decision Trees in order to minimize a loss function [20]. The spark.ml implementation supports GBT for binary classification with categorical features.

4. Distribution of Higgs dataset

In this work, we use the Higgs dataset, downloaded from UCI site. We use 29 Higgs features as independent variables and a dependent response with two classes (signal and background events) [21]:

- features of the high energy (22 variables): 'lepton pT','lepton eta','lepton phi', 'missing energy magnitude', 'missing energy phi', 'jet 1 pt', 'jet 1 eta', 'jet 1 phi', 'jet 1 b-tag', ' jet 2 pt', 'jet 2 eta', 'jet 2 phi', 'jet 2 b-tag', 'jet 3 pt', 'jet 3 eta', 'jet 3 phi', 'jet 3 b-tag', 'jet 4 pt', 'jet 4 eta', 'jet 4 phi', and 'jet 4 b-tag'.
- features of the low energy (7 variables):'m_jj', 'm_jjj', 'm _lv', 'm_jlv', 'm_bb', 'm_wbb', and 'm_wwbb'.

The class label consists of 47% of the signal and 53% of background events. Then, we conclude that the distribution of the Higgs dataset is nearly balanced. We use also Higgs dataset collected from Kaggle site, it consists of 818238 examples with 66% of the signal and 34% of background events and 29 features (see Table 1):

- "DER_pt_h", "DER_deltaeta_jet", "DER_mass_jet", "DER_prodeta_jet", "DER_deltar_tau_lep", "DER_pt_tot","DER_sum_pt", "DER_pt_ratio_lep_tau", "DER_met_phi_centrality", "DER_lep_eta_centrality", "PRI_tau_pt", "PRI_tau_eta", "PRI_tau_phi", "PRI_lep_pt", "PRI_lep_eta", "PRI_lep_ph'i", "PRI_met", "PRI_met_phi", "PRI_met_sumet", "PRI_jet_num", "PRI_jet_leading_pt", "PRI_jet_leading_eta", "PRI_jet_leading_phi", "PRI_jet_subleading_pt", "PRI_jet_subleading_eta", "PRI_jet_subleading_phi", and "PRI_jet_all_pt".

4.1. Split data

Looking to get an unbiased estimation of the performances of the algorithms, Higgs dataset UCI and Higgs dataset Kaggle are divided into a training sample (70%) and test sample (30%). (refer to table 1)

4.2. Evaluation functions

We present some important metrics that are evaluated for comparison [22].

Table 1. description of Higgs datsets: Higgs dataset UCI and Higgs dataset kaggle

	Higgs dataset uci	Higgs dataset Kaggle
instances	11 000 000	818 238
features	29	29
class(signal,backgraound)	Label (s,b)	Label(s,b)
training data 70%	7 700 000	572 767
test data 30%	3 300 000	245 471
signal (s=1)	47%	34%
background(b=0)	53%	66%

4.2.1. Accuracy

Accuracy is the proportion of total examples classified correctly:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 1 - ErrorRate \tag{1}$$

Where:

- TP: The true positive is the number of instances that are Higgs Particles and are being classified as Higgs Particles;
- TN: The true negative is the number of instances which are non-Higgs Particles and being classified as non-Higgs Particles;
- FN: The false negative is the number of true Higgs Particles that are wrongly being classified as non- Higgs Particles;
- FP: The false positive is the number of non-Higgs Particles that are wrongly being classified as Higgs Particles.

4.2.2. AUC-ROC

Receiver Operating Characteristic (ROC) Curve allows comparing various Supervised Learning classifiers. It is especially useful for cases of skewed class distribution [23]. Area Under Curve (AUC) is an area equivalent to the probability that the algorithm will place a randomly selected positive example higher than a randomly selected negative example [24].

5. Experimental results and analysis

5.1. Experimental results

Table 2 Presents the evaluation metrics: AUC and Accuracy with Higgs dataset kaggle and tuning parameters using repeated cross-validation. Table 3 Presents AUC and Accuracy metrics with Higgs dataset UCI.

5.2. Analysis and discussion

For comparing different models on Higgs dataset UCI and Higgs dataset Kaggle in terms of AUC and Accuracy metrics, We analyze the accuracy metric of four classifiers on Higgs dataset with tuning and cross-validation way.
In the context of Higgs dataset UCI, The performance of the four methods varies between 64% and 70%. Figure 1 shows that the GBT method achieved higher accuracy and AUC (70%) in comparison to the Random Forest with 67%, and logistic regression (accuracy=68%, AUC=64%) that better than decision tree with AUC=63 %.
In the context of Higgs dataset Kaggle, The score of performance of these methods is improved. The accuracy metric outperforms 83% with GBT classifier and AUC achieved 82% with logistic regression. Therefore, we remark that

Table 2. AUC and Accuracy with Higgs dataset Kaggle and tuning parameters using repeated cross-validation

	Higgs dataset Kaggle		Tuning and cross-validation	
Classifier	AUC	*Accurcay*	AUC	*Accurcay*
Logistic Regression	0,8149	0,75	0,6348	0,7204
Decision tree	0,7496	0,8004	0,7422	0,7878
Random Forest	0,7708	0,812	0,765	0,797
Gradient Boosted Tree	0,796	0,8254	0,7941	0,8264

Table 3. AUC and Accuracy with Higgs dataset UCI

	Higgs dataset UCI	
Classifier	AUC	*Accuracy*
Logistic Regression	0,6843	0,6421
Decision tree	0,642	0,6357
Random Forest	0,6715	0,6764
Gradient Boosted Tree	0,705	0,7062

BGT score is better than Random Forest score with (AUC=81%, accuracy=77%) and decision tree with (AUC=80%, accuracy=75%) (see figure 2).
This conclusion is close to the repeated cross-validation performance (refer to table 2). Hence, our results are similar to the score accuracy obtained in Higgs challenge 2014 [2].

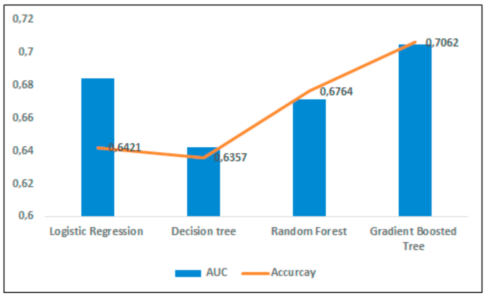

Fig. 1. AUC and Accuracy on Higgs dataset UCI

6. Conclusion

In the context of Higgs Boson classification, we can conclude that:

- As far as Higgs dataset Kaggle, The Gradient Boosted Tree (GBT) classifier runs well with accuracy achieved 83%. It works better than Random Forest (81%), Decision Tree (80%) and Logistic Regresion (75%);
- Concerning Higgs dataset UCI, The Gradient Boosted Tree (GBT) classifier works well with accuracy achieved 70%. It works better than Random Forest (67%), Logistic Regression (64%) and Decision Tree (63%);
- Tuning and repeated cross-validation confirm the ranking of the predictive power of those Machine Learning classifiers.

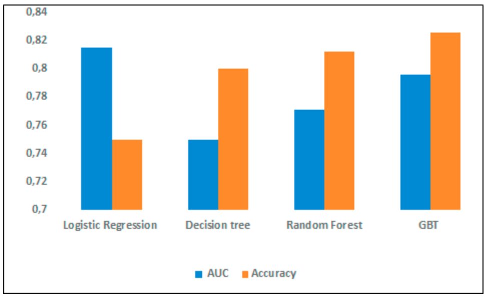

Fig. 2. AUC and Accuracy on Higgs dataset Kaggle

References

[1] Adam-Bourdarios, C., Cowan, G., Germain-Renaud, C., Guyon, I., Kégl, B., and Rousseau, D. (2015). ”The Higgs Machine Learning Challenge.*Journal of Physics: Conference Series”*, **634**: 072015. doi:10.1088/1742-6596/664/7/072015

[2] Chen, Tianqi, and Tong He.(2015) “Higgs Boson Discovery with Boosted Trees,” *JMLR: Workshop and Conference Proceedings* , **42** : 69–80

[3] Baldi, Pierre, Kyle Cranmer, Taylor Faucett, Peter Sadowski, and Daniel Whiteson.(2016)“Parameterized Machine Learning for High-Energy Physics.” *The European Physical Journal*, **76** : 235-241 https://doi.org/10.1140/epjc/s10052-016-4099-4.

[4] Sadowski, Peter J, Daniel Whiteson, and Pierre Baldi. (2014)“Searching for Higgs Boson Decay Modes with Deep Learning,”*Advances in Neural Information Processing Systems (NIPS)* **27**

[5] Sadowski, Peter, Julian Collado, Daniel Whiteson, and Pierre Baldi.(2015) “Deep Learning, Dark Knowledge, and Dark Matter.”,*JMLR: Workshop and Conference Proceedings*, **42** : 81–97

[6] Alves, A.(2017) “Stacking Machine Learning Classifiers to Identify Higgs Bosons at the LHC.” *Journal of Instrumentation* **12** : T05005–T05005. https://doi.org/10.1088/1748-0221/12/05/T05005.

[7] Meng, X., Joseph, B., Burak, Y, Evan, S., Shivaram, V., Davies, L., Jeremy, F., et al.(2016). ”MLlib: Machine Learning in Apache Spark”, *Journal of Machine Learning Research, Boston, MA* **17** : 1–7

[8] Assefi, M., Behravesh, E., Liu, G., and Tafti, A. P. (2017). Big data Machine Learning using apache spark MLlib. 2017 *IEEE International Conference on Big Data (Big Data), Boston, MA*: 3492—3498, doi: 10.1109/BigData.2017.8258338

[9] Armbrust, M. et al.,(2015) .“Spark SQL: Relational Data Processing in Spark,” *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD ’15, Melbourne, Victoria, Australi*: 1383—1394, doi: 10.1145/2723372.2742797.

[10] Peng, H. , Liang, D. , and Choi, C. , (2013). “Evaluating parallel logistic regression models,” *IEEE International Conference on Big Data, Silicon Valley, CA, USA*: 119—126, doi: 10.1109/BigData.2013.6691743.

[11] Loh,W.(2011)”Classification and Regression Trees” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**: 14–23.

[12] Azhari, M. ,Alaoui, A. , Achraoui, Z. ,Ettaki, B., and Zerouaoui, J. (2020). “Detection of Pulsar Candidates using Bagging Method ”*the International Workshop on Statistical Methods and Artificial Intelligence - IWSMAI 2020, Warsaw, Poland*

[13] Genuer,R., and Poggi, J.(2017)” Arbres CART et forêt aléatoires-Importance et sélection de variables”, HAL Id: hal-01387654, https:hal.archives-ouvertes.fr/hal-01387654v2.

[14] Abarda, A., Bentaleb, Y., El Moudden, M., Dakkon, M., Azhari, M., Zerouaoui, J., and Ettaki, B.(2018)”Solving the problem of latent class selection”*In Proceedings of the International Conference on Learning and Optimization,ACM Algorithms: Theory and Applications* **15**

[15] Breiman, L , (2001). ”Random Forests”, *Machine Learnin* **45**:5–32

[16] Azhari, M., Alaoui, A., Achraoui, Z., Ettaki, B., and Zerouaoui, J. (2019). ”Adaptation of the Random Forest Method”, *Proceedings of the 4th International Conference on Smart City Applications - SCA ’19*, doi:10.1145/3368756.336900

[17] Azhari, M., Alaoui, A., Abarda, A., Ettaki, B.,and Zerouaoui , J. (2020) ”Using Ensemble Methods to Solve the Problem of Pulsar Search” ,*In: Farhaoui Y, (eds) Big Data and Networks Technologies, BDNT 2019, Lecture Notes in Networks and Systemst,Springer* **81**: 183-189.

[18] Azhari, M., Alaoui, A., Abarda A., Ettaki, B.,and Zerouaoui, J.(2020): ”A Comparison of Random Forest Methods for Solving the Problem of Pulsar Search” *The Fourth International Conference on Smart City Applications, Springer, Cham.*

[19] Liaw, A and Wiener, M .(2002) ”classification and regression by Random Forest”,*R News* **2**: 18–22

[20] Ganjisaffar, Y. , Caruana, R. , and Lopes, C. V. , (2011). “Bagging Gradient-Boosted Trees for high precision, low variance ranking models” *in Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR ’11, Beijing, China*, doi: 10.1145/2009916.2009932.

[21] Baldi, P., P. Sadowski, and D. Whiteson, (2014) .“Searching for Exotic Particles in High-energy Physics with Deep Learning.”,*Nature Communications* , **5**: 4308, doi.org/10.1038/ncomms5308.

[22] Congalton, Russell G.(1991). “A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data.”*Remote Sensing of Environment*, **37**: 35—46, doi.org/10.1016/0034-4257(91)90048-B.

[23] Kumar, R., and Indrayan, A.(2011) ”Receiver operating characteristic (ROC) curve for medical researchers”*Indian Pediatrics* **48**: 277—287.

[24] Fawcett,T.(2006)”An introduction to ROC analysis” *Pattern Recognit* **27**: 861–874.