Application of Artificial Neural Network(ANN) and Feature Selection Algorithm(FSA) on the ATLAS Experiment Data to Identify Higgs Boson

Haozhan Tang
The High School Attached to Northwest Normal University
No. 21, Shilidian South Street, Anning District, Lanzhou City, Gansu Province, China
*Author’s e-mail: zhouhang@niesl.com.cn

Abstract. From the 1950s, physicists started to smash particles together to temporarily form new smaller particles for observing and studying. With further development and enormous experiments, physicists had successfully built the Large Hadron Collider, which became the key hardware for proton colliding. The historical research and experiment in this field have provided us a number of referring data from the LHC experiment and the detailed information of the LHC upgrade to LH-LHC during the previous several years. The Higgs Boson detected by ATLAS Detector in 2013 was significant to scientific research, and its associated dataset could form a new hypothesis to predict the universe rule for small particles. In this study, both Artificial Neural Network(ANN) and Feature Selection Algorithm(FSA) are applied to the ATLAS experiment data to identify Higgs Boson, including the characteristics and fitness for the current model of nature. The study is based on the experiment result of head-on collisions of protons of extraordinarily high energy. The study is aimed to explore the potential of both methods of ANN and FSA to improve the significance of the experiment study and discovery. It will benefit further study to develop the full potential and the enormous scope of physics opportunities given by LHC.

1. Introduction
In 2012, a new particle with a mass region around 125 GeV was observed by CERN's CMS and ATLAS experiment. It was close to the Higgs Boson's mass but needed further study to confirm its identity[5]. Nevertheless, this was a significant breakthrough in finding the evidence to support the Standard Model and stabilized the confidence of more experiments on LHC to further study the fundamental of the universe. To have better performance in the experiment, the LHC came to a period of closing and upgrading[1]. As one of the four major experiments at LHC at CERN, ATLAS is a general-purpose particle physics experiment run by an international collaboration and, together with CMS. In this study, both ANN and FSA are applied on the ATLAS experiment data to identify Higgs Boson, including the characteristics and fitness for the current model of nature. This project resorted and analyzed the data from the LHC experiment, where both ANN and FSA are applied on the ATLAS experiment data to identify Higgs Boson, including the characteristics and fitness for the current model of nature. Furthermore, this study is also aimed to explore the potential of both methods of Artificial Neural Network (ANN) and Feature Selection Algorithm(FSA) to improve the significance of the experiment study and discovery, such as finding the various rare decays. It will benefit further research to develop the full potential and the large scope of physics opportunities given by LHC.
In 1964, physicists found that protons and neutrons were not the fundamental particles. The outcome of the experiments made physicists believe that there are smaller particles that constitute those protons and neutrons, which are named Quark by the physicists. In a while, physicists strengthened their confidence in the study Standard Model, which is a theory stated that there are 12 fundamental particle types and their antiparticles, combine with the extra gluons, photons, W bosons, Z bosons, and the particle of carrying fundamental forces. The physicists today are trying to find a better theory that can better explain the universe by extending Standard Model[2].

However, when the LHC is operating, the high energy environment is unique and is affected by a number of factors. The controlling technologies on those factors such as radiation damage, temperature difference, and the rigidity of the inner track, are not well established and perfectly working in controlling affecting factors [8]. Furthermore, the LHC is the largest and highest-energy particle accelerator, and the high energy particles' collisions happening inside the LHC may lead to big safety issues, and the LHC Safety Assessment Group paid huge effort to avoid safety issue in 2003[3]. However, the LHC had achieved several breakthroughs in the exploration of the universal rule, such as discovering evidence of the existence of the Higgs boson, also called god particle, which brought the first Nobel Prize in Physics to the LHC experiments in 2013[6]. In the future, the LHC undoubtedly has great prosperity with its new version- the HL-LHC.

2. Motivation
It appears that the particle physics experiment is randomly colliding and form particles, especially from the Pb element, and analyze any result come out of those experiments. The outcome of the experiment almost totally relies on natural probability. However, the experiment is expensive though those experiments are funded by central funds from several European countries[4]. Thus, if some prediction about the outcome of the experiment could come out without experimentation but only data analysis, the funds required for research in particle physics can be reduced. One way to experiment more efficiently is to let the computer help physicists to analysis simulated and previous data and find the potential relationship, which can give physicists ideas about how to design further experiments. In other words, it led the computer to play a more efficient human brain role as an assistant for the particle physicists.

3. Methodology
The project is started with raw data from past years Large Hadron Collider's experiments. Those raw data are typically measured by the detectors inside the inner track of the LHC. Different types of data are sorted and labeled with their specific feature, such as units. Data usually contains the particle's mass, azimuth angle as phi, pseudorapidity as eta, the auto-calculated momentum, and energy value from collisions. From those raw data combined with the additional data generated by the Monte Carlo simulation, the original data set is formed and is put in the analysis process. In the LHC detectors, each layer in the detector has its corresponding layer, and the corresponding layer together with this layer form an event, where each event generates more than thousands of eta and phi. For that phi and eta raw data, the data amount is enormous, and most of those data are not considered to be valuable for analysis. Therefore the raw data such as eta and type are filtered by the process of Data Preference, which divides the raw data of eta and phi into the signal area and background area. The following analysis will only run base on the data in the signal area, which is the area of data that is considered valuable. Feature extraction basically plots different types of data into different types of graphs, typically histogram, to search for the possible hidden feature of the specific data type. The input data set and output dataset are also put into Principal Component Analysis to search for the possible relationship between those data set. Finally, the data is analyzed by the Self-organizing Map, which lets the computer study the data set in a human neural brain way but with significantly faster speed.
4. Data Processing and Results

From using Monte Carlo simulation and refer to the past years' LHC experiments to generate a large number of collisions and sorting data sets into different types, there are 33 types of data set. Except for the data set of Event ID and label, there are 31 types of data sets used for further analysis by graphs, and each dataset variable contains 250000 amount of data. Furthermore, all of the angles relate datasets are in the unit of radian, and all mass, momentum, and energy datasets are in the unit of GeV by converting from natural units. Each event generated from the Monte Carlo simulation is grouped into two areas, signal and background. Only the data in the signal will be considered valuable for data analysis. In the project, only 34% of the events from the raw data events are considered valuable, or signal, for processing.

The Figure 2 contains all types of data sets included in this project. The x-axis is the different types of datasets. Every kind of dataset has its specific inner y-axis, which displays the numerical values in the dataset and to visualize the range of the dataset. The width of the blue area with respect to the area between each dataset boundaries is the frequency of the specific value of the data in the corresponding inner y-bins. Every three different datasets are grouped into one unit to predict one of the random type datasets inside those three different types of datasets. Ten of those units of datasets together form the confusion matrix. The confusion matrix here is aimed to find the correlation between every three types of datasets by using their values to predict each other’s values. A higher diagonal value usually indicates...
more correct or close prediction, which means a higher possibility of correlation to occur. After several arrangements, the blue type area has a significantly higher diagonal number than other colored areas. The right side of the graph shows the diagonal number proportion of the blue type area to all areas and the other type area to all areas in the corresponding line. The bottom side of the graph shows the diagonal number proportion of the blue type area to all areas and the other type area to all areas in the corresponding row. The blue type is occupied most of the diagonal number, which indicates a high correlation of variables in those areas.

PCA is the Principal Component Analysis, which can be used in making predictive models and analyzing exploratory data. Datasets are kept in lower-dimensional with as much as possible variance being maintained. Figure 3 are the distribution of the PCA score of all 31 types of datasets, which shows the Principal component scores matrix. The x-axis displays the number corresponds to the 31 datasets. The y-axis is the value of the representatives of the corresponding dataset in the principal component space. The color bar on the right side of the graph is the color referring values with normalized. The graph also gives the Principal component scores matrix of 31 types of datasets, with the two dimensions. The color bar of this graph shows the color referring values without normalized. After the basic PCA analysis, analysis is running on finding the direct possible linear relationships on all 31 variables. The graph of Figure 4-5 gives the direct possible correlation relationships of 31 types of datasets, with the three dimensions. This 3D displaying also contains 31 rows and 31 columns, which compare each variable to all variables. A taller or more outstanding component of this 3D graph indicates a higher correlation between two variables.

After the PCA analysis, a Self-Organizing Map analysis is applied to those variables to find the potential relationships between those variables. A 31x31 matrix of the variables is input into the SOM analysis with their weight and position analyzed by the neurons in the SOM. SOM input plane after 2 cycles of machine learning in Figure 6-7 shows there is no strong connection of each variable to other variables detected by the SOM, since there are only 2 ANN learning cycles ran to form this input plane. After 15 cycles of machine learning, the input plane of the SOM in Figure 6-7 tends to have more areas in those subplots with a major color other than red. The left lower part of those subplots tends to have more variance color, which indicated more complex relationships. About 50% of the area of those subplots are occupied by a complex color with complex relationships detected by ANN. More and more cycles of Self-Organizing Map were ran, and the input plane of the SOM kept changing as new relationships being detected, and almost 300 cycles of SOM were ran in this project to evaluate the potential of ANN in finding relationship between datas and using it to produce new virtual data.
ANN is utilizing the relationship it found to reversely predict the values of the datasets and check the predicted data with the existing true data to see whether the predicted data is correct. If the predicted data and the true data’s variance is relatively low, and the accuracy is relatively high, the corresponding relationship is confirmed and keep into the next training cycle to further reveal other potential relationships. Figure 8 is a variance trend of the predicted data with the true data graph. The red line of the graph is showing the trend of error, which is the variance between predicted data and the true data. As the red line is shown, the variance is decreasing since the relationships stored in the ANN is more and more reliable after running more cycles of training. The relationships between variables recognized by the ANN is predicting the data with higher accuracy. As Figure 9 shown, the ANN is predicting data more and more correctly and with increasing precision. Therefore, the Artificial Neuron Network did study the 31 types of variables, with 250000 data in each type, in the correct way as expected, and the ANN successfully discovered some potential relationships between variables, which make is predicted unknown values of the dataset with higher and higher accuracy.
5. Conclusion
The ANN and FSA can lead the computer to play an intelligence brain role as an assistant for particle physicists as it revealing features and potential relationships of variables in LHC experiment. From reading and sorting raw data from past years Large Hadron Collider's experiments combined with the data generated by the Monte Carlo simulation under qualified physics model such as the Hard Sphere model, 33 types of variables is formed, with 250000 datas in each variable dataset, and is put into the analysis process.

The outcome of the data processing shows that the potentially complex relationships of those variables are successfully discovered by the ANN, and the ANN can use those recognized relationships to predict more same type of data with increasing accuracy and decreasing variance with the true data. Additionally, the FSA solidifies the base in finding those potential relationships between variables, which select the fitted data for constructing the model in each of the neurons of the SOM in each cycle of training. This method also challenges the relationships or statements of the feature of variables again and again in each cycle by predicting more unknown data and compare them to the true data at the place that is previously “unknown.” In this way, the true relationships or statement of the feature of variables are maintained through cycles of learning, and the wrong part are eliminated from this Artificial Neuron Network. The potential of both methods of Artificial Neural Network and Feature Selection Algorithm to improve the significance of the experiment study and discovery is confirmed through this project. It will benefit further study to develop the full potential and the large scope of physics opportunities given by LHC.

References
[1] Baer, H., Chen, C. H., Paige, F., & Tata, X. (1996). Signals for minimal supergravity at the CERN Large Hadron Collider. II. Multilepton channels. Physical Review D, 53(11), 6241.
[2] Allanach, B. C., Odagiri, K., Palmer, M. J., Parker, M. A., Sabetfakhri, A., & Webber, B. R. (2003). Exploring small extra dimensions at the large hadron collider. Journal of High Energy Physics, 2002(12), 039.
[3] Aiba, M., Fartoukh, S., Franchi, A., Giovannozzi, M., Kain, V., Lamont, M., & Calaga, R. (2009). First β-beating measurement and optics analysis for the CERN Large Hadron Collider. Physical Review Special Topics-Accelerators and Beams, 12(8), 081002..
[4] Plehn, T., Rainwater, D., & Zeppenfeld, D. (2002). Determining the structure of Higgs couplings at the CERN Large Hadron Collider. Physical review letters, 88(5), 051801.
[5] CERN Accelerating science. (n.d.). Retrieved from https://home.cern/science/physics/higgs-boson
[6] Pierog, T., Karpenko, I., Katzy, J. M., Yatsenko, E., & Werner, K. (2015). EPOS LHC: Test of collective hadronization with data measured at the CERN Large Hadron Collider. Physical Review C, 92(3), 034906.
[7] Wilkerson, J. F., Aguayo, E., Iii, F. A., Back, H. O., Barabash, A. S., Beene, J. R., & Busch, M. (2012). The MAJORANA demonstrator: A search for neutrinoless double-beta decay of germanium-76. In Journal of Physics: Conference Series (Vol. 375, No. 4, p. 042010). IOP Publishing.
[8] Banner, M., Madsen, B., Chollet, J. C., Siegrist, J. L., HäNNI, H., Di Lella, L., & Parrou, G. (1983). Observation of single isolated electrons of high transverse momentum in events with...
missing transverse energy at the CERN collider. *Phys. Lett. B*, 122(CERN-EP-83-25), 476-485.

[9] Bonn-Brussels-Cambridge-CERN, U. C., Alpgård, K., Ansorge, R. E., Åsman, B., Berghlund, S., Berkelman, K., & Burow, L. (1983). Forward-backward multiplicity correlations in pp-collisions at \( s = 540 \) GeV. *Physics Letters B*, 123(5), 361-366.