A Comparative Study of Neural Networks and Logistic Regression for High Energy Physics

Nigel Pugh, Joel Gonzalez-Santiago, Thomas Johnson
Center of Excellence in Remote Sensing Education and Research
Elizabeth City State University
Elizabeth City, NC 27909

Jerome E. Mitchell
School of Informatics and Computing
Indiana University
Bloomington, IN 27909

***Abstract*— Neural networks are programs that run based on machine learning algorithms and resources to mirror the function of the brain in its roughest capacity. Neural networks are used primarily for the management and manipulation of large quantities data to form classifications, more efficient searches, and prediction of the data. Neural networks exist as part of the larger field of machine learning that exists. Linear regression in turn serves as the statistics based solution to the classification issue, an alternative to neural networks that are also a form of machine learning. The focus of this research was to observe whether neural networks or linear regression models are more effective for classification of a supersymmetry dataset. The supersymmetry dataset is made up of the results gathered particle collision events within a particle accelerator. Supersymmetry itself is a theory within particle physics that suggests the particles that are absent in the standard model are symmetric, or balancing, counterparts to the particles that have been already discovered.**

Keywords—Neural Networks, High Energy Physics, Logistic Regression

I. Introduction

This research was performed for the purpose of analyzing whether backpropogation or logistic regression will produce more accurate results regarding analysis of the SUSY (supersymmetry) dataset. The fundamental concept that grounds the research endeavor is machine learning, a facet of computer science that deals with the usage of programs that can adapt their own algorithms to become more efficient at the goal that said program is meant to fulfill. Machine learning stands as not only the base of neural networks, but also as the base of search engines, artificial intelligence, virtual intelligence, etc. The machine learning algorithms are developed and tweaked to observe not only the effectiveness of backpropogation and linear regression at one structure, but how does the accuracy change as the structure of the neural network utilizing the backpropogation algorithm and the neural network using the linear regression algorithm alter with modification.

The SUSY dataset which has eight features and was attained through Monte Carlo simulations that are from kinetic impacts of particles within a particle accelerator. It contains eight features as well as five million pieces of data to utilize. The dataset was composed from data that was collected from the usage of particle colliders to examine the conditions in which supersymmetric particles may be present. The reason for this focus on the supersymmetry theory in particle physics is that

it provides answers to many of the questions that plague particle physics to this day, causing more and more theories to be proposed in the wake of a lack of definite answers. Supersymmetry, as it has been applied in particle physics, claims that in the Standard Model, the particles that have been discovered have counterparts that fill the absent particles. These supersymmetric particles, however, have not been identified with certainty which has led to the supersymmetry dataset being crafted from the research to observe anomalies that may indicate the presence of supersymmetric particles. Supersymmetry is critical as it could remove the plethora of mysteries that currently inhabit the field of particle physics and areas of the field of physics as well. The linear regression model is another construct of machine learning that utilizes data that it is given to estimate the result for any given input. Linear regression lacks the adjustments that backpropagation utilizes to rid the model of any significant biases. Linear regression relies heavily on the amount of data that is input for use by the linear regression model. Both the backpropagation and linear regression models were utilized with the data being standardized by a sigmoid function so that the data can be analyzed and results compared.

II. Related Literature

Deep Learning, Dark Knowledge, and Dark Matter is a paper that displays an attempt to utilize deep learning to analyze superparticles that are critical to supersymmetry in physics [1]. The goal was to use a Gaussian process and the usage of the Spearmint variation of Bayesian optimization with backpropagation absent as a baseline for comparison [1]. This leaves the question of whether the model can effectively classify the data and if there are overfitting or misclassifications issues that are overlooked due to the adaptive properties of backpropagation not being used to assist in measuring the accuracy of the proposed neural network model [1].

Searching for exotic particles in high-energy physics with deep learning is a research experiment that used the SUSY dataset for training the particle detectors so that leptons and transversal energy could be utilized for classification of the

signal and background conclusion of particle collisions[2]. The purpose of this is to display that the usage of such instruments as neural networks can provide the opportunity for retaining and analyzing more data than the methods that are typically implemented in physics[2]. The application of SUSY allows for the neural network to be prepared to better analyze the quantities of data that it will be given for further analyzing supersymmetric physics during particle collisions[2]. Backpropagation is avoided for the alternative of neural networks with at least five hidden layers under the assumption that the increased number of hidden layers combined with a function for optimization, in this case gradient descent, will yield better results than backpropagation rather than actually testing such an assumption to see if it stands true [2].

SCYNet: Testing supersymmetric models at the LHC with neural networks a research endeavor centered on utilizing a regression algorithm in combination with a chi-squared test for SUSY dataset phenomena [3]. The regression algorithm is utilized to evaluate the Standard Model of particles in physics to other theoretical or alternative models that are being compared to in observation to data from the Large Hydron Collider [3]. The usage of a regression algorithm alone does not allow for the neural network to be adapted and increase efficiency as the weights remain untouched and uncorrected unlike the backpropagation algorithm [3]. Therefore, there can be a higher chance of overfitting or misclassification in the linear regression algorithm used here [3].

Oblivious Multi-Party Machine Learning on Trusted Processors is a paper that addresses the issue of performing research in machine learning algorithms that are accessible to a number of groups [4]. The SUSY dataset is utilized to test the capability of Intel Skylake processors to override the traditional use of SGX-processors on the grounds of significant improvements of efficiency that contribute to the overall development of machine learning group endeavors [4]. A key portion of such is that the Intel Skylake processors were tested to display that they could preserve the privacy and security that was sought out in the SGX-processors [4].The SUSY dataset was exposed along with other datasets to machine learning algorithms that did not include backpropagation, leaving room for possible improvements that could have been observed thanks to usage of a backpropagation algorithm in competition with the other machine learning algorithms at work [4].

Data stream classification using random feature functions and novel method combinations is a study on the management of data streams by machine learning models that can make identifications with great celerity [5]. The SUSY dataset was utilized as a dataset that allows for comparison of the effectiveness of different methods when implemented for the purpose of classifying data at great speeds [5]. In observing such, notable differences were observed based on the machination that was implemented to allow the machine learning algorithm to build its function for the goal at hand [5]. The issue that arises here is that the measurement of data streams is somewhat compromised when backpropagation is proposed as backpropagation is influenced by a large number of factors within the neural network itself [5]. Backpropagation was disregarded in favor of utilizing other machine learning algorithms that are relatively isolated to research that requires neural networks, removing the usage of backpropagation to act as a an alternative measurement [5].Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. Methodology

The SUSY dataset is utilized to train a neural network that can accurately identify signal and background data that occur when the Large Hadron Collider is being utilized [6]. 1,000 examples were selected from SUSY randomly to train the backpropagation model while another 100 were utilized in testing [6]. In the backpropagation model, 10,000 training iterations were used to allow the backpropagation model to build an accurate function [6]. For each feature, if the mean was larger than one it was compacted to one with a standard deviation of one [6].

A. Neural Networks:

A Neural Neural network can be defined as “A computing system made up of simple, highly interconnected processing elements, which process information by their dynamic state response to external outputs” [7]. A neural network is modeled after the human brain. The brain has millions of neurons that are connected to each other by connections called synapses. These connections transfer signals back and forth from each neuron. These signals, activation levels, and thresholds will result in a neuron being able to send another signal to other neurons that are connected. A Neural networks highly connected neurons are organized in layers.

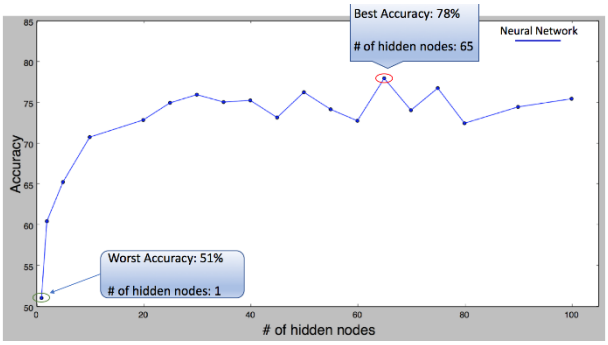

Fig. 1. Accuracy Results for Backpropagation Algorithm with varying hidden nodes

We have three classifications of layers: Input Layer, Hidden Layer, and Output Layer. Normally, three layers is the standard, however multiple hidden layers can be added to the network. Each of the layers are fully connected to the previous layer by weights which are equivalent to synapses in the human brain.

To determine what the initial output of the network, we will have to execute a forward propagation. The input layer takes in data that will be used on the neural network. The input layer is then multiplied by a set of initial weights (synapses). The initial weights are randomly generated. An activation function is applied to the sum of the input multiplied by the weights and stored in the hidden layer. The hidden layer nodes contain an activation function. An activation function takes a single number and performs a certain fixed mathematical operation. This mathematical operation can be the activation function chosen in our neural network was the sigmoid activation function. The mathematical representation of the sigmoid function is $\sigma(x) = \frac{1}{1+e-x}$.

The sigmoid activation function will accept the weighted sum of the inputs and output as a probability. The sigmoid functions outputs range from the 0 to 1. A zero will identifies the example is not a part of the class that it is trying to classify. A one identifies the example is a part of the class that it is trying to classify. In general, large negative numbers will become 0 and large positive numbers will become 1. From the hidden layer the weights are then multiplied by another set of weights and an activation function is applied to the sum of the hidden nodes multiplied by the weights. They are then stored in the output layer. The output layer is used to represent classification of the data that has been inputted. One node is present in the output layer per classification.

Backpropagation is a popular algorithm that is used to train neural networks. Backpropagation is a technique used to correct weights by reducing the error on each back pass. Backpropagation starts from the output layer, and works its way backwards to the input layer, adjusting weights. An error has to be calculated to determine how far the output is away from the target value. The target value can be classified as the desired output. The error equation we used was the squared error function: $Error = ½(target - output)^2$. After the error is calculated we will need to find out how much of the error is respective to each weight?

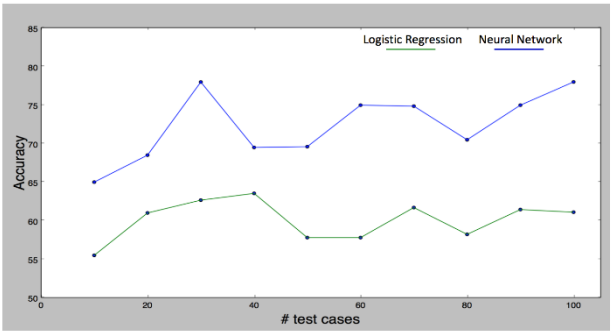

Fig. 2. Backpropigation v.s Logistic Regression Accuracy Results

Partial Derivatives are used to find the error with respect to each weight. (General form is: $\frac{\partial Error}{\partial w}$. After the backpropagation process is complete the weights are then updated. The entire process is then repeated, reducing the error each training iteration. In theory, the more training iterations that are executed, the smaller the error should be.

B. *Logistic Regression*

Architecture: Logistic regression is a formula applied for binary datasets. The values that are received from a logistic regression formula will always be within the range of zero and one. The constraints cause the logistic function graph to show an S shaped curved on the graph. Logistic function is represented by

$$(z) = \frac{1}{1+e-z}$$

Where theta is the output of the function and should be a value within zero and one. In machine learning the Logistic function is also referred to as the sigmoid function. Logistic regression is best used with simple datasets that do not have many input values.

IV. Evaluation

The first experiment conducted was testing the accuracy of the backpropagation algorithm with the SUSY data set. We tested various hidden nodes ranging from 1 to 100. The reason we tested different hidden nodes is due to there being no standard method to finding this optimal number of hidden nodes. The hidden nodes are so closely interconnected and there are many factors that attribute to the determining the accuracy for the optimal hidden nodes. Accuracy is the number of predictions that our model correctly identified, to the total number of test examples. Figure 1 is a graph to represent our experiment various accuracies with respect to number of hidden nodes. The X axis represents the number of hidden nodes, and the Y axis represents the accuracy. Sixty-five hidden nodes produced the best accuracy results at 78 percent. The worst accuracy was 51 percent and only one hidden was used. The next experiment conducted was comparing the backpropagation with the optimal number of hidden nodes (65 nodes in our experiment) against Logistic Regression (Figure 2). The X axis represents the number of test cases and the Y axis represents the Accuracy. It is known that 65 hidden nodes produced the best accuracy in our experiment. The team wanted to test would backpropagation perform well with various number of test cases ranging from 10 to 100 compared to Logistic Regression. According to our experiment, backpropagation outperformed logistic regression on all test case experiments.

V. Conclusion

The research yielded evidence that neural networks with a backpropagation algorithm were significantly more accurate means to classify the SUSY dataset than the logistic regression algorithm. The reason for the backpropagation algorithm being more effective than that of the linear regression algorithm is due to the backpropagation algorithm augmenting the weights in the training phase to yield a better learned model for testing while the linear regression model lacked such a property. Thus, the backpropagation algorithm could significantly increase the accuracy of the learned model before the testing phase versus the linear regression algorithm.

VI. FUTURE WORKS

Research opportunities arose in the possibilities of altering the number of hidden layers to observe if more hidden layers will affect the accuracy of the backpropagation algorithm. The neural network utilized in this research contained only one hidden layer to be utilized with a varying number of nodes in that single hidden layer. For the future, utilizing a neural network that has multiple hidden layers may affect the accuracy similarly to how different numbers of nodes in the hidden layer affected accuracy for the backpropagation algorithm when working with the SUSY dataset. Perhaps with experimentation, the increased number of hidden layers can yield more accurate testing results as suggested in Searching for exotic particles in high-energy physics with deep learning [2].

Altering the number of hidden layers could result in more stable measures of accuracy than altering the number of hidden nodes in a neural network with one hidden layer with the backpropagation and linear regression algorithms. Experiments can be conducted regarding the alteration of the number of hidden layers with a consistent number of nodes in each hidden layer. Their effects on the accuracy for classifying the SUSY dataset can be examined. After altering the number of hidden layers to analyze their effects on the accuracy of the model, the next test would be observing the accuracy of the model utilizing multiple hidden layers while altering the number of nodes within each hidden layer. The number of nodes in the hidden layers would remain consistent amongst the hidden layers, the consistent number of nodes in the hidden layers would be altered to examine if the multiple hidden layers in conjunction with augmenting the number of nodes in the hidden layers.

This research was limited to testing the accuracy of a backpropagation algorithm to a linear regression algorithm which leaves a multitude of alternative machine learning algorithms untested in their effectiveness. Measuring the other machine learning algorithms to backpropagation and linear regression against the SUSY may yield evidence as to which machine learning algorithm is the most effective choice for classification of the SUSY dataset. This research concludes that backpropagation is the better machine learning algorithm for classifying the SUSY data set. The machine learning algorithms that are available extend beyond backpropagation and linear regression algorithms. All of other algorithms offer their own respective advantages and disadvantages. An examination of the results of the backpropagation and linear regression in this research should be compared to determine if the algorithms will hold any significance in further testing.

The neural network that was utilized in this research had 1000 training examples used during the training phase for the model to develop a function through 10,000 training iterations. For the testing phase, there were 100 testing examples to determine the accuracy at which the learned model was classifying any given piece of data. In future projects, the number of testing examples, training examples, and training iterations could be modified to observe how the modification of each affects the accuracy of the model in classifying. The general rule with machine learning is that increased data during the training phase results in a better model which means a higher rate of accuracy for classification. Therefore, exploring the usage of models with an increase in training examples should present results of higher accuracy than the results that were presented in this research.

In addition, the accuracy results in this research need to be compared to accuracy results of other machine learning algorithms to determine if there is a more effective algorithm to classify the SUSY data set. This research was limited to testing the accuracy of a backpropagation algorithm to a linear regression algorithm which leaves a multitude of alternative machine learning algorithms untested in their effectiveness. Measuring the other machine learning algorithms against backpropagation and linear regression against the SUSY may yield evidence as to which machine learning algorithm is the most effective choice for classification of the SUSY dataset.

ACKNOWLEDGMENT

The authors would like to thank Dr. Linda B. Hayden for the opportunity to conduct this research as well as providing funding for it.

REFERENCES

[1] Sadowski, Peter, Julian Collando, and Pierre Baldi. ”Deep learning, dark knowledge, and dark matter”. JMLR Workshop And Conference Proceedings 42. Irvine, California: Journal of Machine Learning Research, 2017. 81-97. Web. 12 Mar. 2017.

[2] Baldi, P., P. Sadowski, and D. Whiteson ”Searching for exotic particles in high-energy physics with deep learning”. Nature Communications 5.4308 (2014): 9. Web. 13 Mar. 2017.

[3] P. Bechtle, S. Belkner, M. Hamer, T. Keller, M. Kramer, B. Sarrazin, J. Schutte-Engel and J. Tattersall, ”SCYNet: Testing supersymmetric models at the LHC with neural networks”, Eur. Phys. J. C manuscript, pp. 1-19, 2017.

[4] O. Ohrimenko, F. Schuster, C. Fournet, A. Mehta, S. Nowozin, K. Vaswani and M. Costa, ”Oblivious Multi-Party Machine Learning on Trusted Processors,” in 25th USENIX Security Symposium, Austin, Texas, 2016, pp. 618-636.

[5] D. Marrn, J. Read, A. Bifet and N. Navarro, ”Data stream classification using random feature functions and novel method combinations,” Journal of Systems and Software, vol. 127, pp. 195-204, 2017.

[6] D. Whiteson and M. Lichman, ”UCI Machine Learning Repository: SUSY Data Set”, Archive.ics.uci.edu, 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/SUSY. [Accessed: 03- Apr- 2017].

[7] “A Basic Introduction To Neural Networks”, Pages.cs.wisc.edu, 2017. [Online].Available:http://pages.cs.wisc.edu/ bolo/shipyard/neural/local.html. [Accessed: 09- Apr- 2017].