Big Data and Artificial Neural Networks

Bharat Mallala
Indiana University
Bloomington, IN 47408, USA
bmallala@iu.edu

ABSTRACT

Big data is often referred as a problem of dealing with large data sets. With the advancements in computational science and the recent evolution of Artificial Intelligence(AI) and Machine Learning, huge volumes of data are being generated every day. Simultaneously the computational resources needed to process and analyze this data is trying to catch up with the rapidly growing data and for the most part have succeeded. In today's world, there is a large dependency on Neural networks for dealing with problems in AI and Data analysis. Big data and its applications can be used to address various issues that arise with Artificial Neural Networks(ANN).

KEYWORDS

I523, hid215, Artificial Neural Networks, Machine Learning, Artificial Intelligence, Data Analysis, Perceptron, Back propagation, Feed forward.

1 INTRODUCTION

Artificial Neural Networks are often referred as a Multi-layer Neural Network where each node in the network is a Perceptron. It often mics the human brain, i.e. it works in a similar fashion. Advancements in ANN's and its ability to solve complex problems at a relatively faster rate than the traditional approaches have made it the top choice for solving the usually NP-hard AI problems. "Visual analysis systems will all require a neural network behind them, and that involves a lot of computing power"[1] quoted Anderson. This explains the efficiency of Neural networks in solving problems and analysis. ANN's take a series of inputs from the users and map them accordingly to find reasonable patterns in data.

Certainly, with these advancements comes huge volumes of data which needs to be stored and processed efficiently. This is where Big data comes into picture with its ability to store and process large data sets of any kind, for example, audio, video, images, text etc in relatively less time. "s Big Data Analytics is an effective and capable way to, not only work with these data but understand its meaning, providing inputs for assertive analysis and predictive actions."[2] quotes Victor P Barros in his paper.

Artificial Neural Networks usually consists of three primary layers, an input layer, output layer, hidden layer. There may be multiple layers of perceptrons within the hidden layer. From the figure 1 we can see the three layers of the ANN. The input layer takes in the input as a set of features and its corresponding weights and the output layer returns a predicted value. All the calculations are done in the hidden layer. The ANN's typically use the feedforward algorithm combined with back propagation for its calculation. The figure below shows the structure of ANN with the three layers. The network initially feeds forward to the very end and generates an output from the initial set of features and weights. It then back propagates using Gradient descent and recalculates the wights for each iteration. The algorithm finally stops of the difference in weights from one iteration to the other is not greater than a predefined threshold. We then test this on the training set and evaluate the performance of the network.

2 ANN'S WITH BIG DATA IN PHARMACOLOGY

2.1 Pharmacology

ANN's with its advancements over the years has been one of the most effective ways in prediction and extracting useful features in the field of Pharmacology. "Pharmacology is the branch of medicine concerned with the uses, effects, and modes of action of drugs"[6]. In this study the data sets comprise of three parts, training data set, testing data set, and validation data set. The ANN's are used to differentiate between active compounds and inactive ones which help in the selection of compounds for use in life-saving drugs.[5]

2.2 History

A lot of research has been conducted in the field of Pharmacology from the late 1980's. Researchers have acknowledged the power of ANN's and have started using it for prediction and classification of robust compounds and materials.[5] But till the early 2000's the use of ANN's have slowed down due to the lack of enough data for training and testing the networks. But with the advancements in big data from the early 2000's the use of ANN's for predictive analytics have raised from the ashes.Big data provided a means of effectively storing huge volumes of the Pharmacological and bioinformatics data and to effectively process them. The effectiveness of big data lies in its ability to store and process the structured, semi-structured and unstructured data. Figure 2 shows the various stages of big data.

2.3 Method and Analysis

A series of processes are carried from the initial state till we get the final predicted outcome. It all starts with the selection of desired features based on the type of analysis to be carried out. This step is called Feature extraction. Hadoop(HDFS) with the ability to store large volumes of data is used to store the large data sets pertaining to Pharmacology. It is a distributed file system with a single name node and various data nodes. Feature extraction process is applied to data on HDFS and the apt features are extracted. The next phase is the initialization of weights, where we need to initialize the weights for the first iteration of the backpropagation algorithm. This is done using an unsupervised pre-training process.

Once the features of interest are extracted and we have the initial weights, then the ANN's are trained over these feature sets to

extract patterns from data and gain insights from them. In general ANN's are classified into a number of modules, for example, Deep Neural Networks(DNN), Convolution Neural Networks(CNN), Deep Recurrent Neural Networks(DRNN)[5].

DRNN is the most prominent when it comes to dealing with Pharmacology data. Generally, a ANN does not have loops in its architecture i.e. the output of each layer acts as input to the next layer in a chain-like manner. But in a DRNN there are loops within the network. i.e. the output of a layer will act as an input to the previous layer forming a loop, hence the name Recurrent. In Pharmacology DRNN's are used to predict the molecular properties or robust compounds. Figure 3 below shows the architecture of a DRNN.

2.4 Role of Big data

Due the vastness of the data sets the traditional algorithms needs to be modified to run the analysis more efficiently. The best approach to deal with this is parallel processing. Spark a component of big data has a set of Machine learning libraries through which the parallel processing is carried on multiple nodes of the network. Spark has an advantage over the other Big data components such as PIG, Hive, Map reduce with ease of writing code as well as user adoption towards it. The data here gets distributed across various data nodes in the cluster where the name node has the metadata pertaining to the data in the data nodes. So the ML algorithms are applied to the data from all the data node where the processing part is carried across multiple machines in a parallel manner. This helps in faster processing and lesser load on the machines. The below figure shows the big data architecture. [4]

3 ANN'S FOR PATTERN RECOGNITION IN CLOUD GAMING USING BIG DATA

3.1 Cloud based games

With the enhanced usage of smartphone and gaming consoles, cloud-based gaming is getting more and more prominent among many users from the past decade. All the traffic from such Cloud-based games and handles over the internet. With more intensive games being created, a large volume of different varieties of data is been generated and this data should be effectively handled for a smooth and pleasant user experience. In any Cloud-based game, there are mainly two players, the client and the server. The client sends requests to the server and the server replies with the relevant information. Some of the cloud-based gaming platforms are Gaikai and OnLive in the early 2000's, while today the most widely used ones are Sony's PlayStation and Microsoft's XBox.

3.2 Role of Big data

Since all these games are being played in real time, a lot of data is being generated every minute across all the major gaming platforms. This data comprises of audio, video, text which need to be stored, processed and analyzed in real time. This is stored information is helpful both for the company which made the game for further improvement of the game and the player's community for discussing tactics and comparing results. Figure 4 shows the Architecture and various components of big data.

Game Telemetry is the statistical data associated with Cloud-based games which most of the companies collect in a timely manner and conduct some rigorous analysis on. This is data is usually in large volumes which cannot be addressed by traditional methods. This where Big data comes into the picture with its ability to store and process such data. This is data also helps is finding patterns in data using some pattern recognition algorithm with ANN which is usually run on robust platforms, and gaining insights from the same. These are generally associated with the Hadoop atmosphere along with the integration of ANN's, where the Cloud-based gaming is carried out in parallel across multiple nodes[2].

3.3 Artificial Neural networks for Analysis

With the power and efficiency of ANN's in predictive analysis there most prominently used in Cloud-based gaming platforms. The dataset used for this study was taken from World of Warcraft game where each player is associated with an avatar which they play with. The ANN's are used analyze user behavior based on their avatar names in the games and classify them as low, medium and high. In this case, the ANN's are used for classification into the three classes stated. The approach involves analysis patterns on how players select their avatars and time they play with them along with the skills and attributes of each of the avatars using Big data tools. The networks used for this approach is the Multi-layer Feedforward network along with Backpropagation.The networks consist of 4 input neurons, 7 hidden neurons, and 1 output neuron. The output neuron is the class variable. The weights, in this case, are initialized at random and the Backpropagation algorithm recalculated the weights in each iteration. This ANN classified the class variable with an accuracy of 84 percent.

4 CONCLUSIONS

Artificial Neural Network is one of the most prominent Machine learning tool used both for classification and regression. Various modulations of ANN's are used based on the type of analysis to be carried out and the class variable. Some of them are Deep Neural Networks(DNN), Convolution Neural Networks(CNN), Deep Recurrent Neural Networks(DRNN). ANN's with its ability to solve complex problems needs a good amount of training data to train the model. Big data and its application and components provided the means to store and process large volumes of the various type of data. A lot of analysis has been conducted in various fields of study with the integration of Big data with ANN's, some of which have been covered in this paper. Results from these studies have proves that the networks can be trained to recognize various patterns in data and learn from the new as it becomes available.

ACKNOWLEDGMENTS

I would like to thank Dr. Gregor von Laszewski and the AI's for all the help they have provided for this paper.

REFERENCES

[1] David Anderson. 2017. President, SEMIfis Americas Operations. (2017). http://www.appliedmaterials.com/nanochip/nanochip-fab-solutions/july-2017/big-data-and-neural-networks

[2] Victor Perazzolo Barros and Pollyana Notargiacomo. 2016. Big Data Analytics in Cloud Gaming: Playersfi Patterns Recognition using Artificial Neural Networks.

IEEE International Conference on Big Data (Big Data) (2016). http://ieeexplore.ieee.org/document/7840782/

[3] Google. 2017. "Google images". (2017). https://www.google.com/search?q=artificial+neural+network+with+backpropagation&num=20&newwindow=1&rlz=1C1CHBF_enUS759US759&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjO7NLZwOLWAhUCzoMKHS9mCgEQ_AUICygC&biw=1536&bih=743&dpr=1.25#imgrc=BFQz8pDLT77p0M:

[4] Google. 2017. Google website. (2017). https://www.google.com/search?q=architecture+of+big+data&num=20&newwindow=1&rlz=1C1CHBF_enUS759US759&source=lnms&tbm=isch&sa=X&ved=0ahUKEwijm5Gu7-HWAhWi6oMKHVMBAw4Q_AUICigB&biw=1536&bih=743&dpr=1.25#imgrc=S0LkDdVUUpcpjM:

[5] Alejandro Pazos Lucas Antn Pastur-Romay, Francisco Cedrn and Ana Beln Porto-Pazos. 2016. Deep Artificial Neural Networks and Neuromorphic Chips for Big Data Analysis: Pharmaceutical and Bioinformatics Applications. *International Journal of Molecular Sciences* 17, 1313 (2016).

[6] Wikipedia. 2016. "Wiki website". (2016). https://en.wikipedia.org/wiki/Pharmacology

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

5 BIBTEX ISSUES

Warning–no journal in Anderson2017

Warning–no number and no volume in Anderson2017

Warning–page numbers missing in both pages and numpages fields in Anderson2017

Warning–no number and no volume in Barros2016

Warning–page numbers missing in both pages and numpages fields in Barros2016

Warning–page numbers missing in both pages and numpages fields in LucasAntonPastur-Romay2016

(There were 6 warnings)

6 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

6.1 Formatting

Incorrect number of keywords

6.2 Writing Errors

Spelling errors

Do not use phrases such as *shown in the Figure below.* Instead, use *as shown in Figure 3*, when referring to the 3rd figure

6.3 Citation Issues and Plagiarism

The citation mark should not be in the beginning of the sentence or paragraph, but in the end, before the period mark. example: ... a library called Message Passing Interface(MPI) [7].

Put a space between the citation mark and the previous word

6.4 Structural Issues

Abstract contains more of introduction material rather than a summary of the paper

6.5 Details about the Figures and Tables

The resolution of Figure 1 is too low

Figure 2,3 don't have a reference. In case you copied a figure from another paper you need to ask for copyright permission. In case of a class paper, you must include a reference to the original in the caption

Figures should be reasonably sized and often you just need to add columnwidth

e.g.

```
/includegraphics[width=\columnwidth]{images/myimage.pdf}
```

re

List of Figures

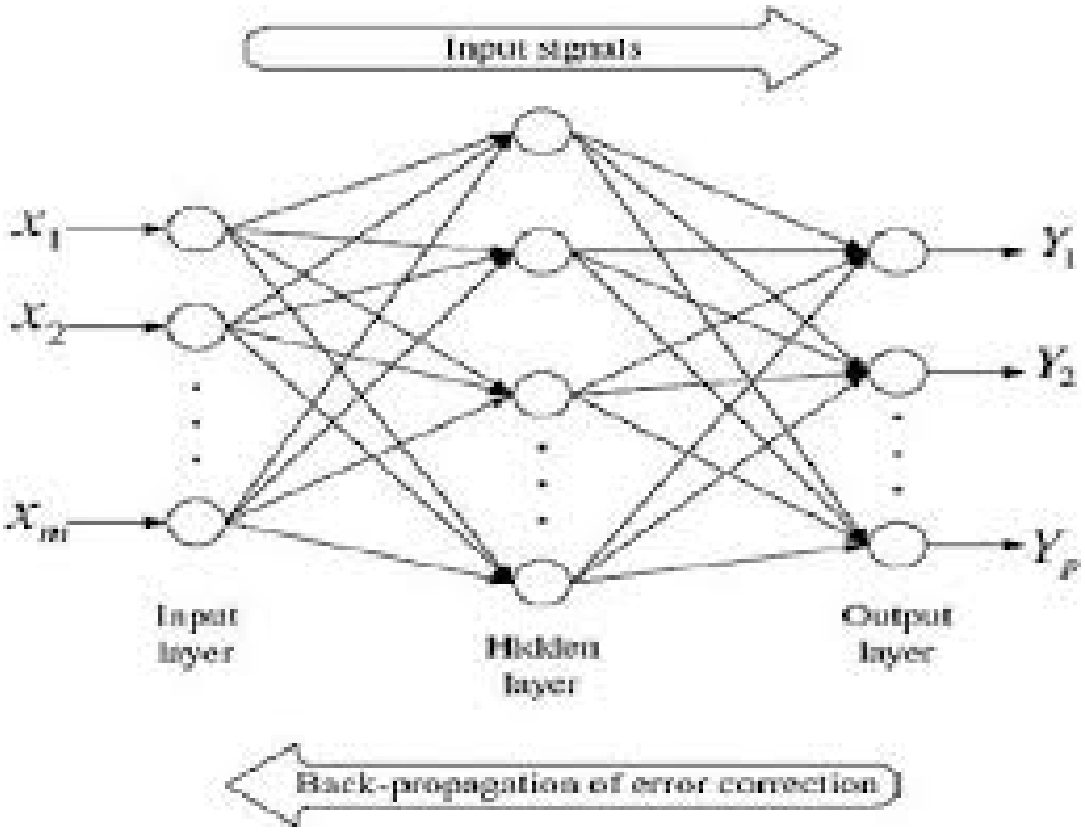

Figure 1: Neural Network Architecture [3]

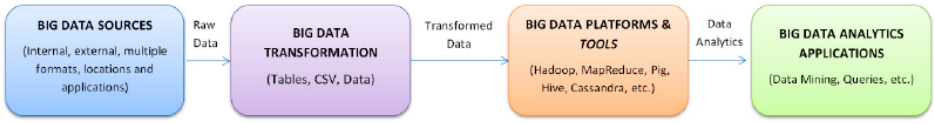

Figure 2: Stages of big data

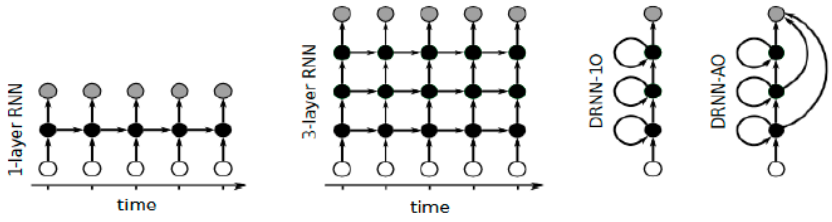

Figure 3: Architecture of DRNN

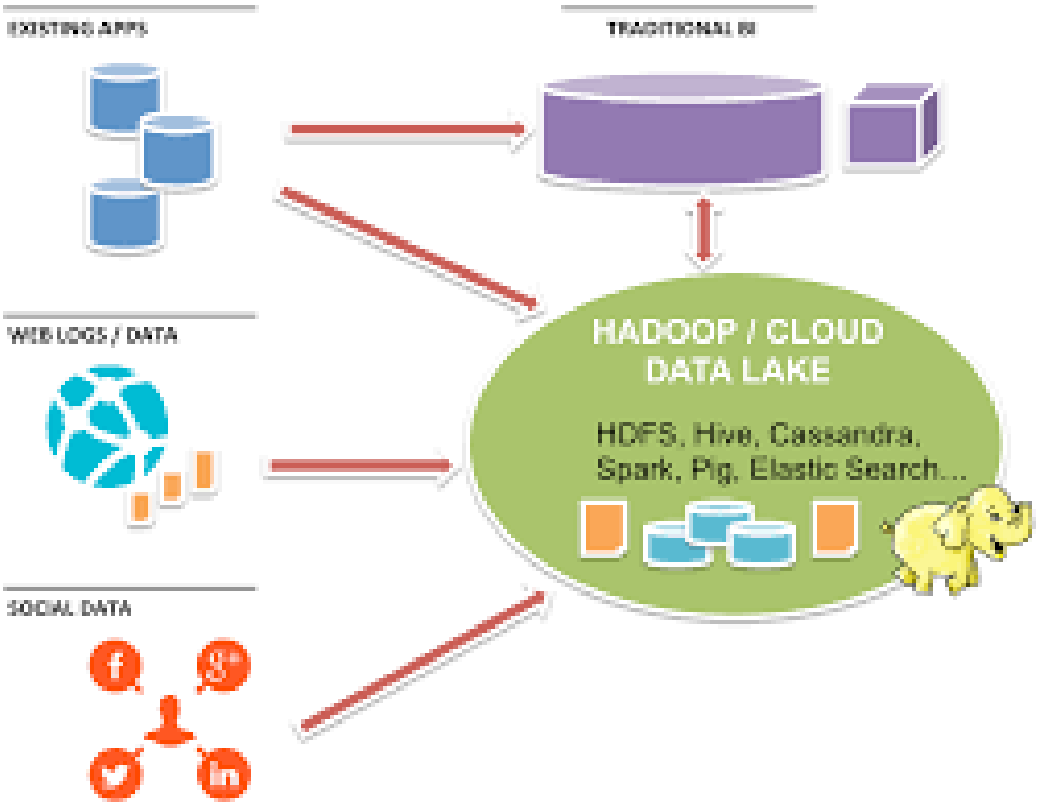

Figure 4: Big data Architecture and Components [4]