Big Data Analytics Framework using Machine Learning on Multiple Datasets

Surekha Sharad Muzumdar [1], Jharna Majumdar [2]

[1] M-Tech Scholar, Dept of CSE (PG), NMIT, Bangalore 560064, Karnataka, India

[2] Dean R&D, Professor and Head CSE (PG), NMIT, Bangalore 560064, Karnataka, India

Abstract: ***Over 2.5 quintillion bytes of data have been created in last two years alone. These kinds of data comes from various sources such as healthcare informatics, weather information, sensors data, cell phone GPS signals, social media, digital images and videos, transactional information, etc. Big Data refers to huge collection of data sets that are so complex that it becomes so difficult to process using traditional data processing applications. Therefore it requires new set of framework to manage and process Big Data. Map Reduce plays a significant role in processing Big Data. In this paper, the multiple datasets such as data from healthcare organization, weather dataset and movie ratings dataset are stored and organized directly to distributed file system like HDFS. Then finally data is analyzed using Apache Hive for faster query access. In this paper Machine learning techniques are used to solve a big data analytics in a better and simple way.***

Keywords: Big data, Hive, Hadoop, HDFS, Machine Learning, COBWEB

1. Introduction

In today's day-to-day life Internet has led to huge amount of information being available online. Over 2.5 quintillion bytes of data have been created in last two years alone. These kinds of data comes from various sources available everywhere such as healthcare informatics, weather information, sensors data, cell phone GPS signals, social media, digital images and videos, transactional information, etc [1]. Big Data refers to huge collection of data sets that are so complex that it becomes so difficult to process using traditional data processing applications [2]. Therefore it requires new set of framework to manage and process Big Data. Big data analytics is the process of examining huge data sets containing a variety of data types to uncover unknown correlations, hidden patterns, customer preferences and other useful business information.

MapReduce plays a significant role in processing Big Data. MapReduce programming model has two functions map() and reduce() which has high degree of elasticity and fault tolerance but performance is slower compared to two parallel database processing applications [6]. Once the data is collected from Healthcare Organization, movie dataset or weather dataset it can be stored and organized directly to distributed file system like HDFS (Hadoop Distributed File System) or GFS (Google File System). Then finally data is analysed using components of Hadoop Ecosystem such as Flume, Sqoop, Pig, and Hive which has inbuilt map reducers. These components analyses the data so that it can be accessed faster and easily and query responses also becomes faster [2]. Data from different sources of social media are acquired from Flume. Flume gathers log files from different systems. Data from other traditional database systems can be loaded using Sqoop. Pig has a high-level language and an execution environment. Pig uses data flow language called Piglatin. Hive is datawarehouse software which provides query language similar to SQL called HiveQL which manages querying over large datasets [2].

1.1 Characteristics of Big Data

Big Data has different characteristics such as Volume, Velocity, Variety, Varacity and Value as shown in Figure 1.

Volume: Volume refers to large amount of data available in an organization. At present if the data is present in Terabytes and it is supposed to increase to Exabyte in near future. Therefore it is difficult to handle by traditional data processing applications [3].

Velocity: Velocity refers to data arriving at high speed. Big Data Analytics are near real-time hence high performance and high computing facility is needed [4].

Variety: Variety refers to data coming from heterogeneous resources. Data can either be in DBMS table format as structured form or in a semi-structured form or in the unstructured form such as email attachments, images, audios, videos, or medical records such as MRIs, ECGs, X-rays etc [5].

Varacity: Due to increase in velocity, data cannot be cleaned before using it. Hence for decision making and business there must be a mechanism to deal with imprecise data. Big Data is the combination of precise, imprecise, accurate, inaccurate form of data [7].

Value: Value refers to hidden information from big data. The challenge here is to identify, extract, transform and analyse this information to find the hidden value from it [4].

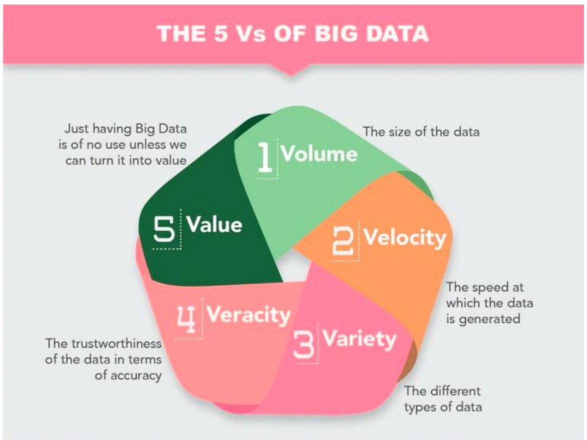
THE 5 Vs OF BIG DATA
1 Volume
The size of the data
2 Velocity
The speed at which the data is generated
3 Variety
The different types of data
4 Veracity
The trustworthiness of the data in terms of accuracy
5 Value
Just having Big Data is of no use unless we can turn it into value

Figure 1: Characteristics of Big Data

1.2 Various Datasets in Big Data Analytics

1.2.1 Healthcare

Healthcare stores large amount of data such as patient's medical history, medication, billing and drug manufacturing company. These data are very complex in nature and sometimes practitioners cannot correlate with other information, thus important information in results remains hidden. By applying predictive analytics techniques, these hidden information can be extracted, which results in personalized medication.

1.2.2 Weather Forecasting

Weather Forecasting is the application of science and technology to predict the state of atmosphere for a future time at a given location. Weather dataset is collected from NCDC (National Climatic Data Centre). Numerical weather prediction includes a highly developed framework that analyses the weather data by combining observations with short range (i.e. hourly, daily and monthly) forecasts of the weather.

1.2.3 Movie Lens Data

Movie Lens data is collected from GroupLens Research lab in the Department of Computer Science and Engineering at the University of Minnesota [15]. It is made available rating data sets from the MovieLens web site (http://movielens.org) [15]. The size of the movie lens data are available in 100k, 1M, 10M and 20M which is composed of 100,000 ratings from 1000 users on 1700 movies, 1 million ratings from 6000 users on 4000 movies, 10 million ratings and 100,000 tag applications applied to 10,000 movies by 72,000 users, 20 million ratings and 465,000 tag applications applied to 27,000 movies by 138,000 users respectively [15].

2. Apache Hive

MapReduce programming model requires developers to write a custom program which are more than 1500 lines of code for every simple operation and thus it is difficult to maintain and reuse [8]. Hence Hive is used for faster query access. Hive was developed by Facebook and later made an open source data warehousing solution built on top of hadoop [8]. Hive supports SQL like declarative language called Hive Query Language (HiveQL) [8]. HiveQL are compiled by Map Reduce jobs that are executed on Hadoop. Hive also includes system catalog called Metastore db which is used for query optimization, data exploration and query compilation [8].

Syntax of the Hive QL queries:

Create Table:

```
CREATE TABLE a (k1 string, v1 string);
CREATE TABLE b (k2 string, v2 string);
CREATE TABLE c (k3 string, v3 string);
```

For joining 2 tables:

```
SELECT  k1, v1, k2, v2
FROM  a  JOIN  b  ON  k1 = k2;
```

For joining more than 2 tables:

```
SELECT  a.v1, b.v2, c.v3
FROM  a  JOIN  b ON  (a.k1 = b.k2)
JOIN  c  ON  (c.k3 = b.k2) ;
```

Load the data:

```
LOAD   DATA   INPATH   'hdfs:/small/data.txt'    INTO
TABLE data;
```

Save the Results in HDFS:

```
INSERT  OVERWRITE  DIRECTORY '/small/data'
SELECT  k1, v1, k2, v2
FROM  a  JOIN  b ON  k1 = k2;
```

3. Machine Learning Algorithms

Machine learning is a field of computer science that explores the construction of study algorithms which can learn from and make predictions on data. Such kind of algorithms builds a model on example inputs and makes data driven predictions, decisions.

3.1 Regression

Regression analysis is used when the values are ordinal rather than categorical. Regression analysis is also called as number prediction. Regression is the task of predicting continuous values for the given input. This model is used for the prediction of one or more variables where one variable is predictor variable and the other variable is the response variable.

3.1.1 Logistic Regression with Regularization

In classification the output is binary rather than numeric thus to generalize it to numeric 0 to 1 range, we apply logit function for the linear regression [13].

$$y = \frac{1}{1+e^{-(a+bx)}} \qquad (1)$$

where in equation (1), the value of y is in the range 0 to 1

x = input numeric value

a = the intercept

b = the slope

In Equation (1) the value of *a* and *b* is calculated by the method of least squares.

$$a = \bar{y} - b\,\bar{x} \qquad (2)$$

$$b = \frac{\sum xy - n\,\bar{x}\,\bar{y}}{\sum x^2 - n\bar{x}^2} \qquad (3)$$

where in equation(2),

$\bar{x} = \frac{\sum x}{n}$ is the mean of x data

$\bar{y} = \frac{\sum y}{n}$ is the mean of y data

For Logistic Regression with regularization we can apply Equation (4) where the value of Equation (1) is applied for non-regularization cost [13].

$$\text{Cost} == \text{Non-regularization-cost} + \lambda\,(\alpha.\Sigma\,\|\Theta i\| + (1-\alpha).\Sigma\,\Theta_i^2) \quad (4)$$

The equation (4) [13] avoids an over fitting problem LI regularization is also called lasso regression ||Θi|| and L2 regularization is also called Ridge regression Θ_i^2 is added to the cost along with logistic regression [13].

3.2 Classification

Classification is similar to regression whereas it is used to predict the categorical values of the given input. Based on some specific set of rules defined by certain algorithm classifier are formed. The efficiency of the classifier depends on how well the input data is classified into certain class.

3.2.2 Neural Network – Back Propagation Algorithm

Back propagation is the neural network learning algorithm which comprises the set of neurons. Each network input or output neurons are associated with weights [14]. During the training phase the neurons adjusts the weights so as to predict the correct class labels [14].

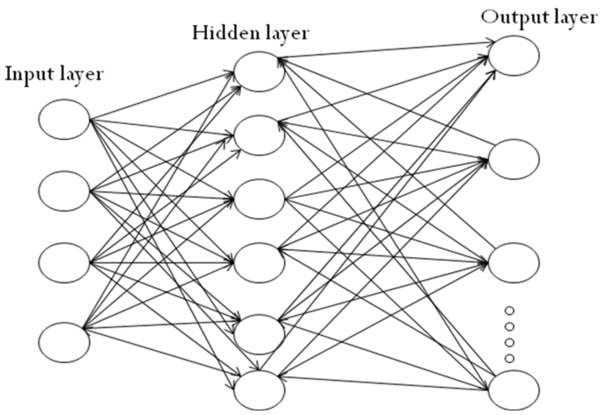

Figure 2: Architecture of Neural Network

The Algorithm uses feed forward and back propagation methods. For the feed forward method, first compute the net input value, this is done by multiplying the input unit and the corresponding weight associated with it [14] as shown in equation (5).

$$I_j = \sum w_{ih}\, O_i \quad (5)$$

The output unit is computed by taking the sigmoid function of the net input value [14] as shown in equation (6).

$$O_j = \frac{1}{1+e^{-I_j}} \quad (6)$$

The error is back propagated and the error at the output unit [14] is calculated by the equation (7).

$$Err_j = O_j(1 - O_j)(T_j - O_j) \quad (7)$$

The error at the hidden layer is computed [14] as shown in equation (8).

$$Err_j = O_j(1 - O_j)\sum_k Err_k W_{jk} \quad (8)$$

Weights are updated by the following equations which are as shown in equation (9) and (10) where Δw_{ij} the updated weight of is w_{ij} [14].

$$\Delta w_{ij} = (l) Err_j O_j \quad (9)$$

$$w_{ij} = w_{ij} + \Delta w_{ij} \quad (10)$$

3.3 Association

Association shows the frequency of the attribute value pairs for the given set of inputs. Association rule mining is usually used where frequent item set mining is used. Usually this type of mining is used in transactional databases where frequent items keep occurring throughout the database. Apriori Algorithm, FP-Growth is the types of Association rule mining.

3.3.1 Apriori Algorithm

Mining frequent item sets lead for the discovery of association rule mining in the transactional and relational datasets. Large datasets are collected thus mining such frequent patterns are important [14].

The set of items I = {$I_1, I_2, \ldots I_n$} are called item sets. The set of these items occurs in a database are called as transactions T which is a subset of item I [14]. Support and confidence is calculated for each item set. Support s, is the probability that the transaction T contains both A and B i.e. (AUB). Confidence c is the probability of A will also contain B in the transaction T [14].

$$\text{support}(A \rightarrow B) = P(A \cup B) \quad (11)$$

$$\text{confidence}(A \rightarrow B) = P(B|A) \quad (12)$$

The minimum support is the threshold which is given either in percentage from 0% to 100% or in 0 to 1.0 forms [14].
The k-itemset is an itemset that contains k items. Conventionally we use 1-itemset, 2-itemset and so on [14]. By the equation (12) we have,

$$\text{confidence}(A \rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support\ (A)} \quad (13)$$

Equation (13) is used to find the frequent itemset. Therefore the apriori algorithm runs on finding all the possible frequent itemsets and then finding the frequent itemset which is more than the min_support count [14].

3.4 Cluster Analysis

Cluster analysis is the process of analysing the objects which are similar to each other into one cluster and the objects which are dissimilar forms a different cluster. Conceptual Clustering is type of cluster analysis where clusters are formed for each concept. K-means, DB Scan are the clustering methods which form iteratively whereas

COBWEB is the conceptual clustering algorithm which keeps incremented rationally [12].

3.4.1 COBWEB Algorithm

COBWEB is a type of conceptual clustering algorithm. It generates a concept descriptor for each cluster. The cluster of COBWEB algorithm forms a tree like structure where root node represents entire dataset and the leaves represent individual concept and the branches represents the hierarchical clusters of the dataset. The total number of clusters depends on the size of the datasets. Each node of the tree represents certain concept of the dataset [11]. Figure 3 represents an example of COBWEB algorithm where C0 refers the probability of the entire dataset and C1, C2 and C3 represents the probability of the individual concept [11].

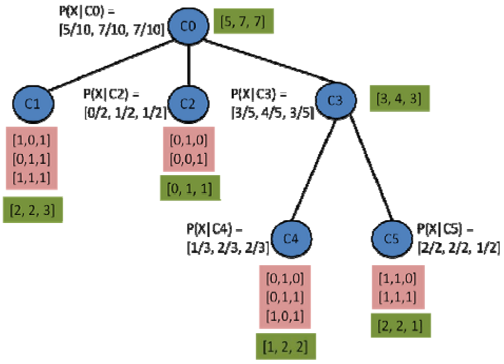

Figure 3: An example of the COBWEB Algorithm

4. Experimental Results

At first three different datasets are collected such as Healthcare dataset, movie lens dataset and weather dataset from different sources. These datasets are then pre-processed. Using Apache Hadoop and Apache Hive as SQL like query language these datasets are processed individually to find out the analytics and the results are stored back into HDFS (Hadoop Distributed File System). Using machine learning algorithms these results which were stored in HDFS are applied to find out the analytics of the datasets and to find out the accuracy of Machine learning algorithms.

4.1 Accuracy of Machine Learning Algorithms in Healthcare

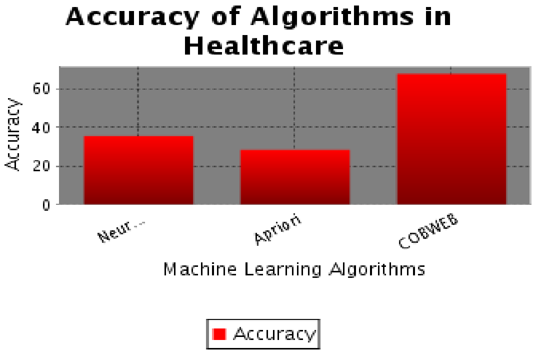

Figure 4: The accuracy of Machine Learning algorithms for healthcare dataset

In the Figure 4 the Neural Network –Back propagation algorithm, Apriori – Frequent item set and COBWEB algorithm when applied to healthcare dataset, the accuracy of COBWEB algorithm shows 68% of accuracy which is found better among the three.

4.2 Accuracy of Machine Learning Algorithms in Movie Dataset

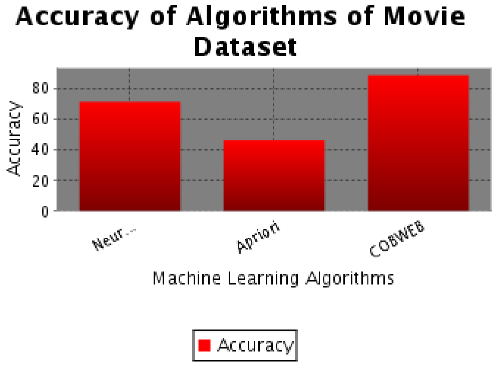

Figure 5: The Accuracy of Machine Learning Algorithms for movie lens dataset

In the Figure 5 the Neural Network –Back propagation algorithm, Apriori – Frequent item set and COBWEB algorithm when applied to movie lens ratings dataset, the accuracy of COBWEB algorithm shows 88% of accuracy which is found better among the three.

4.3 Correlation of Logistic Regression on Multiple Datasets

Correlation among all the three datasets such as healthcare, movie lens and weather data set are found to be same which is 0.99 which means that both the values *x* and *y* are dependent with each other which is shown in Figure 6.

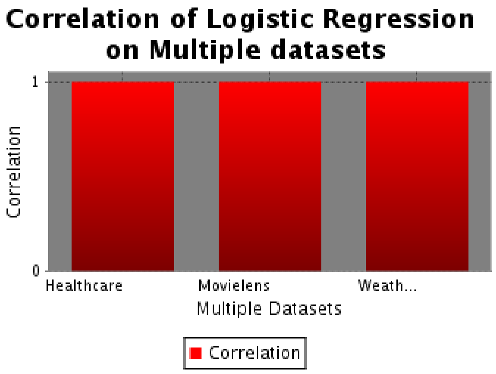

Figure 6: Correlation of Logistic regression on multiple datasets

4.4 Accuracy of Neural Network algorithm on multiple datasets

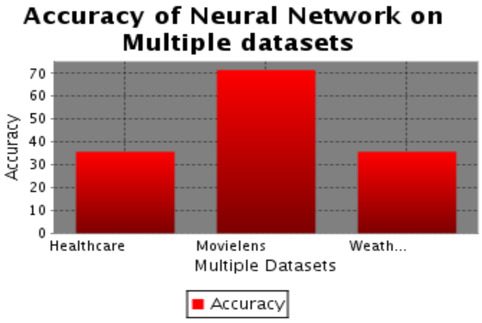

Figure 7: Accuracy of Neural Network Algorithm on Multiple Datasets

In the figure 7 the Neural Network – Back Propagation algorithm when applied to healthcare dataset, movie lens rating dataset and weather dataset it is found that the algorithm shows 71% of accuracy for movie lens ratings dataset.

5. Conclusion and Future Scope

In this paper big data analytics framework using machine learning on multiple datasets, different machine learning algorithms such as regression, classification, association and clustering algorithms are implemented on multiple datasets such as Healthcare Informatics, Movie Lens Rating and Weather Forecasting. The datasets are processed using Apache Hive for faster query access and the results of the queries are stored back in HDFS. Then the machine learning algorithms are applied for these datasets. The efficiency of the algorithm depends on the datasets used. The accuracy and the execution time are the metrics used.

For better accuracy this work can be extended using Mahout which is a part of Hadoop ecosystem. Mahout has inbuilt machine learning algorithms. Due to the complexity of the usage of mahout tool, this was not used in our work. In future mahout can be used as machine learning tool for Big Data for better algorithm efficiency.

References

[1] Shamil Humbetov, “Data-Intensive Computing with Map-Reduce and Hadoop”, IEEE, 978-1-4673-1740-5 /12, 2012
[2] Sanjay Rathee, “Big Data and Hadoop with components like Flume, Pig, Hive and Jaql”, International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.
[3] Avita Katal, Mohammad Wazid, R H Goudar, “Big Data: Issues, Challenges, Tools and Good Practices”, IEEE, 978-1-4799-0192-0/13, 2013.
[4] Zaiying Liu, Ping Yang, Lixiao Zhang, “A Sketch of Big Data Technologies”, 2013 Seventh International Conference on Internet Computing for Engineering and Science.
[5] Puneet Singh Duggal, Sanchita Paul, “Big Data Analysis: Challenges and Solutions”, International Conference on Cloud, Big Data and Trust 2013, Nov 13 -15, RGPV.
[6] Shweta Pandey, Dr.Vrinda Tokekar, “Prominence of MapReduce in BIG DATA Processing”, 2014 Fourth International Conference on Communication Systems and Network Technologies.
[7] Parth Chandarana, M. Vijayalakshmi, “Big Data Analytics Frameworks”, 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA).
[8] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu, Raghotham Murthy, “Hive –A Petabyte Scale Data Warehouse Using Hadoop”, Facebook Team, ICDE, 2010.
[9] Wullianallur Raghupathi, Viju Raghupathi, “Big data analytics in healthcare: promise and potential”, Health Information Science and Systems 2014.
[10] Divya Tomar and Sonali Agarwal, “A survey on Data Mining approaches for Healthcare”, International Journal of Bio-Science and Bio-Technology, 2013
[11] Yuni Xia, Bowei Xi, “Conceptual Clustering Categorical Data with Uncertainty”, 19th IEEE International Conference on Tools with Artificial Intelligence, 2007.
[12] D. Karina Trejo T., Ma. Auxilio Medina N, Jorge de la Calleja M., Érika A. Martínez M, J. Alfredo Sánchez, “An approach to visualize unorganized collections of documents”, IEEE, 978-1-4799-3469-0/14, 2014.
[13] Ricky Ho, “Big Data Machine Learning: Patterns for Predictive Analytics”, Dzone Refcardz
[14] Jiawei Han and Micheline Kamber, “Data Mining concepts and techniques”, Elsevier, Morgan kaufmann publishers, second edition, 2006
[15] http://grouplens.org/datasets/movielens/

Author Profile

Surekha Sharad Muzumdar received the B.E in Information Science and Engineering in 2012 and currently pursuing M.Tech in Computer Science and Engineering at Nitte Meenakshi Institute of Technology, Bangalore. Her areas of Interest are Big Data Hadoop and its ecosystem, Java, Android.

Dr. Jharna Majumdar currently working as Dean R & D and Professor and Head of Computer Science and Engineering (PG) at the Nitte Meenakshi Institute of Technology, Bangalore. Prior to this Dr. Majumdar served Aeronautical Development Establishment, Defense Research and Development Organization (DRDO), Ministry of Defense, Govt. of India as Research Scientist and Head of Aerial Image Exploitation Division, Bangalore. Dr. Majumdar has 40 years of experience in R & D and Academics in the country and abroad. She has published large number of papers in National, International Journals and Conferences. Her Project with team of students from 7 engineering colleges and ISRO for building the first smallest satellite in India (a PICO satellite of weight less than 1 kg) had taken off to the orbit successfully by mid May 2010. Nitte Amateur Satellite Tracking Centre (NASTRAC) developed by a team of students from NMIT, Bangalore, under her guidance is the first Tracking Station of Small Satellites developed in the country. The research team of robotics under her guidance has developed a Robot with Innovative Vision System and installed at the Birla Science Centre, Hyderabad as the first Robotics Exhibit in Indian Museums. Dr. Majumdar has 3 patents with students of NMIT, Bangalore in last 4 years.