ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

A Study on Weather Forecasting using Machine Learning in Big Data

Kavita Devi [1], Nandani shrama [2]

P.G. Student, Department of Computer Science & Engineering, SRCEM, Palwal, Haryana, India[1]

Assistant Professor, Department of Computer Science & Engineering, SRCEM, Palwal, Haryana, India [2]

ABSTRACT: Collecting, storing and processing of huge amounts of climatic data is necessary for accurate prediction of weather. Meteorological departments use different types of sensors such as temperature, humidity etc. to get the values. Number of sensors and volume and velocity of data in each of the sensors makes the data processing time consuming and complex. Leveraging Map Reduce with Hadoop to process the massive amount of data. Hadoop is an open framework suitable for large scale data processing. Map Reduce programming model helps to process large data sets in parallel, distributed manner. This project aims to build a data analytical engine for high velocity, huge volume temperature data from sensors using Map Reduce on Hadoop. However Weather forecasting has traditionally been done by physical models of the atmosphere, which are unstable to perturbations, and thus are inaccurate for large periods of time. Since machine learning techniques are more robust to perturbations, in this paper we explore their application to weather forecasting to potentially generate more accurate weather forecasts for large periods of time. The scope of this paper was restricted to forecasting the maximum temperature and the minimum temperature for five days with weather condition, given weather data for the past two years. A linear regression model and a variation on a regression model were used for the future implementation and research which will be able to capture trends in the weather and Support Vector Machine to originate the vectors to segregate the complexity based on hyper plane. Both of the models were outperformed by and adhered professionally for weather forecasting scenarios.

KEYWORDS: Weather Forecasting, Big Data, Hadoop, Map-Reduce, Linear Regression, Support Vector Machine.

I. INTRODUCTION

Machine Learning techniques are the process of extracting meaningful information from the large data set. The process of extract meaningful information described as knowledge discovery that can be applied on any large data set. The main Machine Learning techniques are Classification, Clustering, Association rules, Regression and Classification. The different techniques used for solving various forecasting problem using supervised and unsupervised learning techniques. Weather forecasting problem include prediction of temperature, rain, fog, winds, clouds, storm etc. Weather sensors collect data every hour at many locations and gather a huge data. Weather forecasting is always a big challenge because it is hard to predict the state of the atmosphere for the upcoming future because climate dataset is unpredictable and frequently change according to global climate changes. The data used is from the INDIA METEOROLOGICAL DEPARTMENT(IMD), the format of dataset support a rich set of meteorological elements, which are good candidate for analysis with big data because it is semi-structured and record oriented. The term Big Data came around 2005, which refers to datasets that are big, also high in variety and velocity, which makes them difficult to process using traditional tools and techniques. Big data created huge business and social opportunities in each field, enabling the discovery of previously hidden patterns and the development of new insights to make decisions, ranging from web search to content recommendation and computational advertising. The term Big Data is now used almost everywhere in our daily life and it is a current technology and also going to rule the world in future and has emerged because people and different companies makes increasing use of data-intensive technologies. Big data sizes are currently ranging from a Terabyte (TB or 1012 or 240) to Zettabyte (ZB or 1021 or 270) in a single data set. Like the physical universe, the digital universe is large. According to research conducted by IDC, from 2005 to 2020, the digital universe will grow from 130 Exabytes to 40,000 Exabyte's, or 40 trillion gigabytes. From now, the

DOI: 10.15680/IJIRCCE.2019. 0701025

digital universe will about double every two years until 2020. As stated by IBM, with machine-to-machine(M2M) communications, online/mobile social networks and pervasive handheld devices it creates 2.5 quintillion bytes of data in each day — so much that 90 percentage of the data in the world today has been produced in the last two years alone.

Characteristics of Big data– Big Data has many characteristics or properties mentioned by n V's characteristics. Set of V's characteristics of the Big Data were collected from different researcher's publications to have Nine V's characteristics (9V's characteristics). These 9V's characteristics are: (Veracity, Variety, Velocity, Volume, Validity, Variability, Volatility, Visualization and Value).

1. **Veracity**: Big Data veracity refers to the biases, noise, and abnormality in data.
2. **Variety**: Structured, semi-structured, and unstructured data besides text and more data types have emerged, such as record, log, audio, and hybrid data.
3. **Velocity**: The created information at a faster pace than before, in which the different channels of Big Data increase the output content.
4. **Volume**: the amount of data is known as volume of data, where the amount of data continues to explode. • Validity: the data is correct and accurate for the intended use. Clearly, valid data is the key to making the right decisions.
5. **Variability**: the data flows may be highly inconsistent with periodic peaks, daily, seasonal, and event-triggered peak data loads can be challenging to manage, especially with unstructured data involved.
6. **Volatility**: Once retention period expires, we can easily destroy it.
7. **Visualization**: means complex graphs that can include several variables of data while still remaining understandable and readable
8. **Value**: It has a low-value density as a result of extracting value from massive data. Useful data needs to be extracted from any data type and from a huge amount of data.

Hadoop : Hadoop is widely used in big data applications in the industry, e.g., spam filtering, network searching, click-stream analysis, and social recommendation. In addition, considerable academic research is now based on Hadoop. Some representative cases are given below. As declared in June 2012, Yahoo runs Hadoop in 42,000 servers at four data centers to support its products and services, e.g.,searching and spam filtering, etc. At present, the biggest Hadoop cluster has 4,000 nodes, but the number of nodes will be increased to 10,000 with the release of Hadoop 2.0. In the same month, Facebook announced that their Hadoop cluster can process 100 PB data, which grew by 0.5 PB per day as in November 2012. Some well-known agencies that use Hadoop to conduct distributed computation in addition, many companies provide Hadoop commercial execution and/or support, including Cloudera, IBM, MapR, EMC, and Oracle. According to the Gartner Research, Bigdata Analytics is a trending topic in 2014 and further years. Hadoop is an open framework mostly used for Big Data Analytics. Map Reduce is a programming paradigm associated with the Hadoop.

Big Data is that data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast and is unstructured and doesn't fit the structures of the architectures. To gain value from this data we need an alternative way to process it . Various fields for example that generate such large amounts of huge data are Facebook, Twitter ,Weather stations ,New York Stock Exchange , Worldwide electric transmissions etc. Thus in our project we are dealing with huge amount of unstructured weather data.

While under this scheme we focuses on the shifting of processes from single node data processing to Hadoop distributed file system for faster processing and the best technique to process the queries. Weather forecasting is always a big challenge for the meteorologists to predict the state of the atmosphere at some future time and the weather conditions that may be expected. It is obvious that knowing the future of the weather can be important for individuals and organizations.

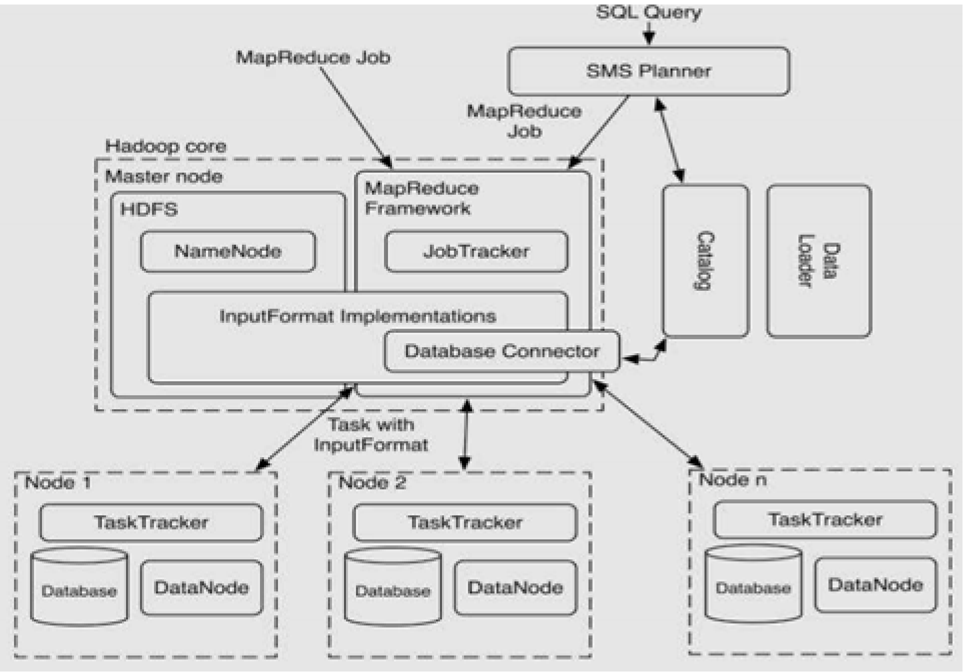

Figure 1: Hadoop Ecosystem.

Some of its characteristics of Hadoop are following. It is an open-source system developed by Apache in Java. It is designed to handle very large data sets. It is designed to scale to very large clusters. It is designed to run on commodity hardware. It offers resilience via data replication. It offers automatic failover in the event of a crash. It automatically fragments storage over the cluster. It brings processing to the data. Its supports large volumes of files into the millions. The Hadoop environment as shown in Figure 1.

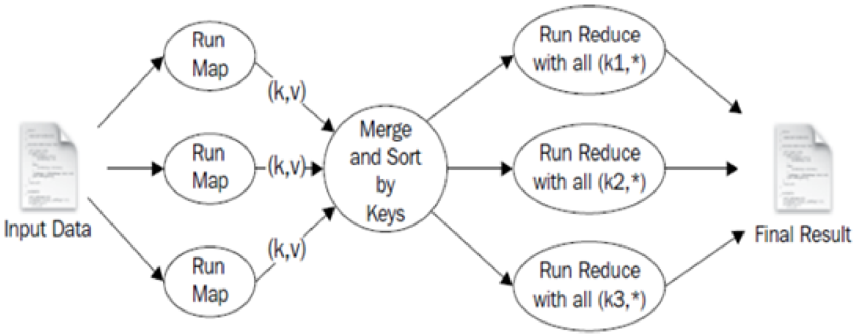

Figure 2: Map Reduce Framework

Map Reduce : Hadoop using HDFS for data storing and using Map Reduce to processing that data. HDFS is Hadoop’s implementation of a distributed file system. It is designed to hold a large amount of data, and provide access to this data to many clients distributed across a network. Map Reduce is an excellent model for distributed computing, introduced by Google in 2004. Each Map Reduce job is composed of a certain number of map and reduce tasks. The Map Reduce model for serving multiple jobs consists of a processor sharing queue for the Map Tasks and a multi-server queue for the Reduce Tasks. To run a Map Reduce job, users should furnish a map function, a reduce function, input data, and an output data location as shown in figure 2. When executed, Hadoop carries out the following steps: Hadoop breaks the

DOI: 10.15680/IJIRCCE.2019. 0701025

input data into multiple data items by new lines and runs the map function once for each data item, giving the item as the input for the function. When executed, the map function outputs one or more key-value pairs. Hadoop collects all the key-value pairs generated from the map function, sorts them by the key, and groups together the values with the same key. For each distinct key, Hadoop runs the reduce function once while passing the key and list of values for that key as input. The reduce function may output one or more key-value pairs, and Hadoop writes them to a file as the final result. Hadoop allows the user to configure the job, submit it, control its execution, and query the state. Every job consists of independent tasks, and all the tasks need to have a system slot to run. In Hadoop all scheduling and allocation decisions are made on a task and node slot level for both the map and reduce phases. There are three important scheduling issues in Map Reduce such as locality, synchronization and fairness. Locality is defined as the distance between the input data node and task-assigned node. Synchronization is the process of transferring the intermediate output of the map processes to the reduce process.

Linear Regression: Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeller might want to relate the weights of individuals to their heights using a linear regression model. Before attempting to fit a linear model to observed data, a modeller should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables. A scatter plot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).
The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

Example : The dataset "Televisions, Physicians, and Life Expectancy" contains, among other variables, the number of people per television set and the number of people per physician for 40 countries. Since both variables probably reflect the level of wealth in each country, it is reasonable to assume that there is some positive association between them. After removing 8 countries with missing values from the dataset, the remaining 32 countries have a correlation coefficient of 0.852 for number of people per television set and number of people per physician. The r² value is 0.726 (the square of the correlation coefficient), indicating that 72.6% of the variation in one variable may be explained by the other. (Note: see correlation for more detail.) Suppose we choose to consider number of people per television set as the explanatory variable, and number of people per physician as the dependent variable. Using the MINITAB "REGRESS" command gives the following results: The regression equation is People.Phys. = 1019 + 56.2 People.Tel.

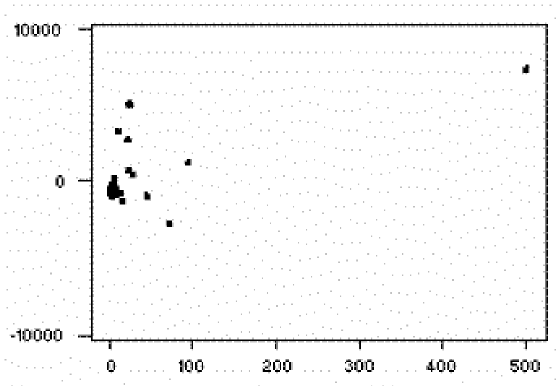

Figure 3: Linear Regression Depicting Residual Sum of Square and Regressed Sum of Square

Support Vector Machine: Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the illustration below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right is labeled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).

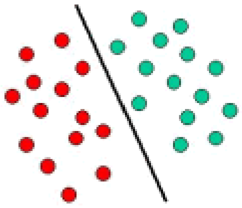

Figure 4: Decision Plane of Objects using Class Membership.

The above is a classic example of a linear classifier, i.e., a classifier that separates a set of objects into their respective groups (GREEN and RED in this case) with a line. Most classification tasks, however, are not that simple, and often more complex structures are needed in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). This situation is depicted in the illustration below. Compared to the previous schematic, it is clear that a full separation of the GREEN and RED objects would require a curve (which is more complex than a line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers. Support Vector Machines are particularly suited to handle such tasks.

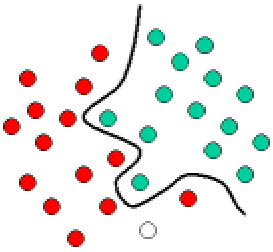

Figure 4: Classification of Vectors based on Hyper-plane.

The illustration below shows the basic idea behind Support Vector Machines. Here we see the original objects (left side of the schematic) mapped, i.e., rearranged, using a set of mathematical functions, known as kernels. The process of rearranging the objects is known as mapping (transformation). Note that in this new setting, the mapped objects (right side of the schematic) is linearly separable and, thus, instead of constructing the complex curve (left schematic), all we have to do is to find an optimal line that can separate the GREEN and the RED objects.

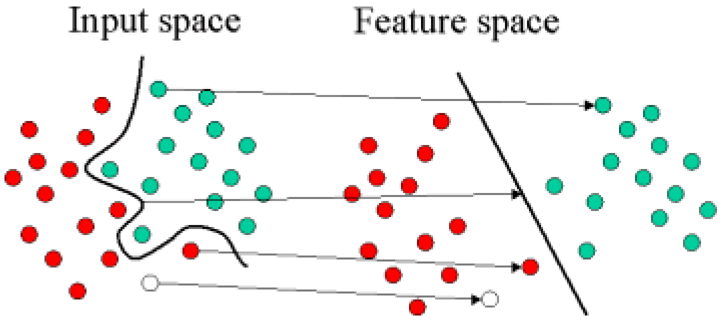

Figure 4: Classification of Vectors based on Hyper-plane and Kernels.

II. RELATED WORK

Surekha Sharad Muzumdar, Jharna Majumdar [1] depicts that 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.

Amir Ghaderi, Borhan M. Sanandaji, Faezeh Ghaderi [2] depicts that, the paper presents a spatio-temporal wind speed forecasting algorithm using Deep Learning (DL) and in particular, Recurrent Neural Networks (RNNs). Motivated by recent advances in renewable energy integration and smart grids, we apply our proposed algorithm for wind speed forecasting. Renewable energy resources (wind and solar) are random in nature and, thus, their integration is facilitated with accurate short-term forecasts. In our proposed framework, we model the spatiotemporal information by a graph whose nodes are data generating entities and its edges basically model how these nodes are interacting with each other. One of the main contributions of our work is the fact that we obtain forecasts of all nodes of the graph at the same time based on one framework. Results of a case study on recorded time series data from a collection of wind mills in the north-east of the U.S. show that the proposed DL-based forecasting algorithm significantly improves the short-term forecasts compared to a set of widely-used benchmarks models.

Hassani, H. & Silva [3] depicts that, the Big Data is a revolutionary phenomenon which is one of the most frequently discussed topics in the modern age, and is expected to remain so in the foreseeable future. In this paper we present a comprehensive review on the use of Big Data for forecasting by identifying and reviewing the problems, potential, challenges and most importantly the related applications. Skills, hardware and software, algorithm architecture, statistical significance, the signal to noise ratio and the nature of Big Data itself are identified as the major challenges which are hindering the process of obtaining meaningful forecasts from Big Data. The review finds that at present, the fields of Economics, Energy and Population Dynamics have been the major exploiters of Big Data forecasting whilst Factor models, Bayesian models and Neural Networks are the most common tools adopted for forecasting with Big Data.

Kiran Kumar & Usha Rani [4] depicts that, Weather is one of the most effective environmental constraints in every phase of our lives on the earth. We need to predict the weather such as temperature, rainfall, humidity etc to protect our self from abnormal conditions. The objective of our work is to design an effective rainfall prediction agent model using support vector machine and multiple linear regressions. To evaluate the proposed model, it implemented using MATLAB and compared with existing numerical models. Three quantitative standard statistical parameters, such as mean absolute error, root mean square error and Nash-Sutcliffe coefficient of efficiency are employed to compare with existing models. The experimental results reveal that the proposed model out performs the existing numerical rainfall prediction models.

III. CONCLUSION

Both Linear Regression and Support Vector Machine (Machine Learning Models) will be inculcated on weather forecasting scenarios over the Big-Data using Mapper and Reducer, although the discrepancy in their performance decreased significantly for future days, indicating that over longer periods of time, our models may outperform professional ones. Linear regression proved to be a low bias, high variance model whereas SVM proved to be a high bias, low variance model. Linear regression is inherently a high variance model as it is unstable to outliers, so one way to improve the linear regression model is by collection of more data and mapping and reducing using hadoop ecosystem. Regression, however, was high bias, indicating that the choice of model was poor, and that its predictions cannot be improved by further collection of data. This bias could be due to the design choice to forecast weather based upon the weather of the past two years, which may be too short to capture trends in weather that regression requires. If the forecast were instead based upon the weather of the past two or less than that, the bias of the SVM model could likely be reduced. However, this would require much more computation time along with retraining of the weight vector w, so this will be deferred to future weather forecasting therefore this scheme is just study to produce the effective and more accurate solution in out future or proposed research. Consequently, below we elaborate our study for the instruments and tools we are going to inculcate for future research.

REFERENCES

1. Surekha Sharad Muzumdar,Jharna Majumdar, Big Data Analytics Framework using Machine Learning on Multiple DatasetsInternational Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013)
2. Amir Ghaderi, Borhan M. Sanandaji, Faezeh Ghaderi, Deep Forecast: Deep Learning-based Spatio-Temporal Forecasting, Cornell University, arXiv.org Year 2017
3. Hassani, H. & Silva, Forecasting with Big Data: A Review, E.S. Ann. Data. Sci. (2015) 2: 5. https://doi.org/10.1007/s40745-015-0029-9
4. Kiran Kumar & Usha Rani, Weather Prediction through Machine Learning, South Asian Journal of Engineering and Technology Vol.2, No.38 (2016) 4–7 , ISSN Number (online): 2454-9614
5. Bélair, S., A. Méthot, J. Mailhot, B. Bilodeau, A. Patoine, G. Pellerin, and J. Côté, 2000: Operational implementation of the Fritsch–Chappell convective scheme in the 24-km Canadian regional model.Wea. Forecasting, 15, 257–274
6. Bousquet, O., C. A. Lin, and I. Zawadzki, 2006: Analysis of scale dependence of quantitative precipitation forecast verification: A case study over the Mackenzie River Basin. Quart. J. Roy. Meteor. Soc., 132, 2107–2125
7. Lorenz, E. N., 1963: Deterministic non periodic flow. J. Atmos. Sci., 20, 130-141.
8. Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. Tellus, 17, 321-33 3.
9. Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occcuring analogues. J. Atmos. Sci., 26, 636-646.
10. Hansen, L.K, & Salamon, P. (1990.) Neural Network Ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, 10, 993-1001.
11. ichardson, L. F. Weather Prediction by Numerical Process (Cambridge Univ. Press, 1922) R. E. Huschke (editor), Glossary of Meteorology, American, Meteorological Society, Boston,
12. Massachusetts, USA, pp 106, 419, 1959
13. R. L. Vislocky and J. M. Fritsch, An automated, observations based system for short-term prediction of ceiling and visibility, Weather Forecasting, 12, pp31–43, 1997.
14. Pandey GR, Nguyen VTV (1999) A comparative study of regression based methods in regional flood frequency analysis. Journal of Hydrology 225: 92–101.
15. T. W. Liao, Z. Zhang and C. R. Mount, Similarity measures for retrieval in case based reasoning systems, Applied Artificial Intelligence, 12, pp267–288,1998.