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 3, March 2019

Review on Weather Forecasting using Linear Regression and SVM 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: Weather or Climate forecasting is very much considered as the most requesting issue both hypothetically and experimentally by the world in the most recent decade. This, in the long run, came about into an extraordinary interest for creating models which help towards a viable forecast of the climate forecasting information. A decent number of the meteorologists have made critical walks in anticipating the climate utilizing models dependent on time arrangement. In the greater part of the models dependent on time arrangement, the examination of climate information is done by thinking about a couple of factors for the appraisal of the information. Be that as it may, the characteristics of the climate assume a significant job in climate forecasting. The proposed scheme or model will look into reality and develop the sophisticated model with comprehensiveness for examining the climate conduct and change of the last few years to predict the weather or future weather or climate. using machine learning in the most recent decade in science and innovation helped in proposing dynamic methodologies for the forecast of climate by utilizing experimental methodologies. Enormous information relating to the meteorology is accessible for use in various arrangements. This information is created both from the surface perception stations and flying investigation stations. With the expansion in the number of climate stations, gigantic information is accessible regularly, day by day, week by week, month to month and yearly premise and the information is put away exponentially. This information is put away and is made accessible for the successful investigation of climate forecasting, fiasco anticipating and for the use of different divisions. The investigation of the related information from this gigantic information is of essential significance and requirements mining procedures wherein using Big Data, Map Reduce, Linear Regression and SVM the future weather prediction will be made with high accuracy and results.

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

I. INTRODUCTION

In the barometrical sciences, meteorological information is very rich and esteemed, which requires a mass of logical processing, and gives administrations to the networks. With the further extension of meteorological administrations and the improvement of the modernization standard in meteorology, a lot of meteorological data has been gathered and gathered in meteorological administrations, research, and the executives exercises. Elite PCs or machines with enriched computational powers are required to process these information, yet little associations and units can't manage the cost of the high cost of superior PCs. Distributed computing using Hadoop innovation gives shoddy processing administrations to the Meteorological Organization with higher proficiency, lower cost, and lower carbon. Atmosphere information are drastically expanding in volume and multifaceted nature since clients of these information in established researchers and the open are quickly expanding. Looked to such vast scale meteorological information, high-proficient processing power (in excess of a trillion times) is earnestly required. Along these lines, setting up a distributed computing climate data preparing framework is significant and critical. MapReduce is a key innovation of utilizing distributed computing to process a lot of information. It is a parallel programming model and a related execution for handling and creating extensive datasets in a wide assortment of true errands proposed by Google. It isn't just a programming model yet additionally an undertaking booking model. It is made out of two key capacities: Map and Reduce, characterized by clients. A Map work is to deal with a key/esteem pair to deliver the middle of the road key/esteem pair. A Reduce work is indicated to join the majority of the moderate an incentive with a similar center key.

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 3, March 2019

MapReduce is regularly used to perform conveyed figuring on groups of PCs. Google's MapReduce abstracts the circulated figuring from its perplexing subtleties; to such an extent that software engineers can deal with vast dispersed framework assets with no experience about a parallel computing or conveyed framework using Hadoop Distributed File System . Along these lines, the impact initially accomplished just by costly elite PC can be accomplished by minimal effort registering administrations. As we probably am aware, not all information mining calculations can be parallelized to deal with substantial datasets as of now. A few calculations can't be parallelized in principle. Some should be adjusted to exploit the productivity of parallelization. In this paper, we use the Support Vector Machine calculation in the MapReduce structure. In particular, we receive the Support Vector Machine calculation in an open-source programming structure: Hadoop and apply the parallel MapReduce with SVM to group the vast scale climate information.

II. RELATED WORK

Related works incorporated a wide scope of and captivating frameworks to endeavour to perform atmosphere figures. While a great deal of current deciding advancement incorporates re-enactments in light of material science and differential conditions, various new philosophies from electronic thinking used basically AI methodologies, for the most part neural frameworks while some drew on probabilistic models, for instance, Bayesian frameworks. Out of the three papers on AI for atmosphere desire we examined, two of them used neural frameworks while one used help vector machines. Neural frameworks give off an impression of being the conspicuous AI model choice for atmosphere deciding by virtue of the ability to get the non-direct states of past atmosphere examples and future atmosphere conditions, not at all like the straight backslide and useful backslide models that we used. This gives the upside of not tolerating fundamental direct states of all features over our models. Of the two neural framework approaches, one [3] used a blend exhibit that used neural frameworks to demonstrate the material science behind atmosphere evaluating while the other [4] associated adjusting even more explicitly to envisioning atmosphere conditions. In like manner, the methodology using support vector machines [6] Additionally associated the classifier direct for atmosphere estimate yet was more limited in degree than the neural framework approaches. Distinctive procedures for atmosphere anticipating included using Bayesian frameworks. One charming model [2] used Bayesian frameworks to show and make atmosphere desires, nonetheless, used an AI figuring to find the best Bayesian frameworks and parameters which was computationally expensive because of the considerable proportion of different conditions yet performed incredibly well. Another methodology [1] focused on an increasingly specific occurrence of foreseeing extraordinary atmosphere for a specific geological territory which confined the necessity for adjusting Bayesian framework conditions anyway was compelled in degree.

Number	Name	Value
1	Classification	Clear
2	Maximum	Temperature (F) 57
3	Minimum	Temperature (F) 33
4	Mean Humidity	Humidity 43
5	Mean Pressure	Atmospheric Pressure in 30.13

Table 1: Sample data from January 1, 2017, with the number, name, and value of each of the five features.

III. LITERATURE SURVEY

A. Adamu Galadima portrays a short take a gander at the Arduino microcontroller and some of its applications and how it can be utilized as a part of learning. Arduino is an open source microcontroller utilized as a part of electronic prototyping. Arduino equipment and its segments might be taken a gander at. Programming and the Environment that Arduino keeps running on are both taken a gander at as well. A few applications will be taken as illustrations that can

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 3, March 2019

help make learning Arduino additionally fascinating. This can be utilized as a noteworthy method to urge understudies and others to take in more about gadgets and programming.

B. Jeffrey Cohen display information parallel calculations for advanced factual systems, with an emphasis on thickness strategies. At last, he responds on database framework includes that empower deft outline and adaptable calculation improvement utilizing both SQL and Map Reduce interfaces over an assortment of capacity instruments.

C. Brian Dolan display the outline rationality, methods and experience giving MAD examination to one of the world's biggest promoting systems at Fox Audience Network, utilizing the Green plum parallel database framework. We depict database plan approaches that help the light-footed working style of examiners in these settings.

D. R. P. Singh clarify why a cloud-based arrangement is required, depict our model usage, and investigate some case applications we have executed that show individual information proprietorship, control, and examination. He address these issues by outlining and executing a cloud-based engineering that furnishes buyers with quick access and fine-grained control over their utilization information, and also the capacity To break down this information with calculations of their picking, including outsider applications that investigate that information in a protection saving style.

E. Jeffrey Dean depicts the essential programming model and gives a few cases. Numerous genuine errands are expressible in these models. Usage of Map Reduce keeps running on an extensive bunch of ware machines and is exceptionally adaptable: a regular Map Reduce calculation forms numerous terabytes of information on a huge number of machines. Software engineers and the framework simple to utilize: several Map Reduce programs have been actualized and upwards of one thousand Map Reduce employments are executed on Google's bunches each day.

F. Panagiotis D. Diamantoulakis implements the Big Data Analytics for Dynamic Energy Management in Smart Grids. The smart electricity grid enables a twoway flow of power and data between suppliers and consumers in order to facilitate the power flow optimization in terms of economic efficiency, reliability and sustainability. This infrastructure permits the consumers and the micro energy producers to take a more active role in the electricity market and the dynamic energy management (DEM). The most important challenge in a smart grid (SG) is how to take advantage of the user's participation in order to reduce the cost of power.

G. L. Aniello investigate the possibility of a structure utilizing various information sources to enhance assurance capacities of CIs. Difficulties and openings are examined along three fundamental research bearings: I) utilization of particular and heterogeneous information sources, ii) checking with versatile granularity, and iii) assault demonstrating and runtime mix of various information examination procedures

IV. PROPOSED WORK

The most outrageous temperature, slightest temperature, mean clamminess, mean barometrical weight, and atmosphere gathering for consistently in the years 2011-2015 for Delhi, India were gained from Weather Underground. [7] Originally, there were nine atmosphere orders: clear, scattered fogs, to some degree shady, generally shady, dimness, overcast, rain, tempest, and snow. Since an extensive parcel of these requests are practically identical and some are meagrely populated, these were diminished to four atmosphere groupings by joining scattered fogs and not entirely shady into sensibly shady; generally shady, foggy, and shady into extraordinarily shady; and rain, tempest, and snow into precipitation. The data from the underlying four years were used to set up the counts, and the data from the latest year was used as a test set and the alluded data for January using the table 1 depicted parameters.

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 3, March 2019

Number	Name	Value
1	Classification	Clear
2	Maximum	Temperature (F) 57
3	Minimum	Temperature (F) 33
4	Mean Humidity	Humidity 43
5	Mean Pressure	Atmospheric Pressure in 30.13

Table 1 : Parameters for Regression and Classification

The essential count that was used was straight backslide, which tries to suspect the high and low temperatures as an immediate blend of the features. Since straight backslide can't be used with gathering data, this computation did not use the atmosphere course of action of consistently. As needs be, only eight features were used: the best temperature, minimum temperature, mean moistness, and mean climatic weight for each of the past two days. In this way, for the I-th join of consistent days, x (I) ∈ R9 is a nine-dimensional component vector, where x0 = 1 is portrayed as the square term. There are 14 adds up to be expected for each join of consecutive days: the high and low temperatures for each of the accompanying seven days. Let y (I) ∈ R 14 imply the 14-dimensional vector that contains these sums for the I-th match of progressive days utilizing direct relapse and further utilizing help vector machine arrangement limit the blunder work utilizing:

$$\frac{1}{2} w^T w - v\rho + \frac{1}{N}\sum_{i=1}^{N} \xi_i$$

subject to the constraints:

$$y_i\left(w^T \phi(x_i) + b\right) \geq \rho - \xi_i, \xi_i \geq 0, i = 1,\ldots, N \text{ and } \rho \geq 0$$

For this type of SVM the error function is:

$$\frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i + C\sum_{i=1}^{N} \xi^{\bullet}{}_i$$

which we minimize subject to:

$$w^T \phi(x_i) + b - y_i \leq \varepsilon + \xi_i^{\bullet}$$
$$y_i - w^T \phi(x_i) - b_i \leq \varepsilon + \xi_i$$
$$\xi_i, \xi_i^{\bullet} \geq 0, i = 1,\ldots, N$$

Below is the proposed workflow scheme comprising of Rawdata source for meteorological department, Pre-processing technique, Migrating data to Hadoop Distributed Filesystem using Hadoop, thereinafter integrating data with schema model using HIVE which is Object Relation Database Management System, Subsequently using MapReduce, therein using Linear Regression to find the Intecept, Slope, Residual Sum of Square, Regressed Sum of Square as regressed values finally using Support Vector Machine for classification and results.

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 3, March 2019

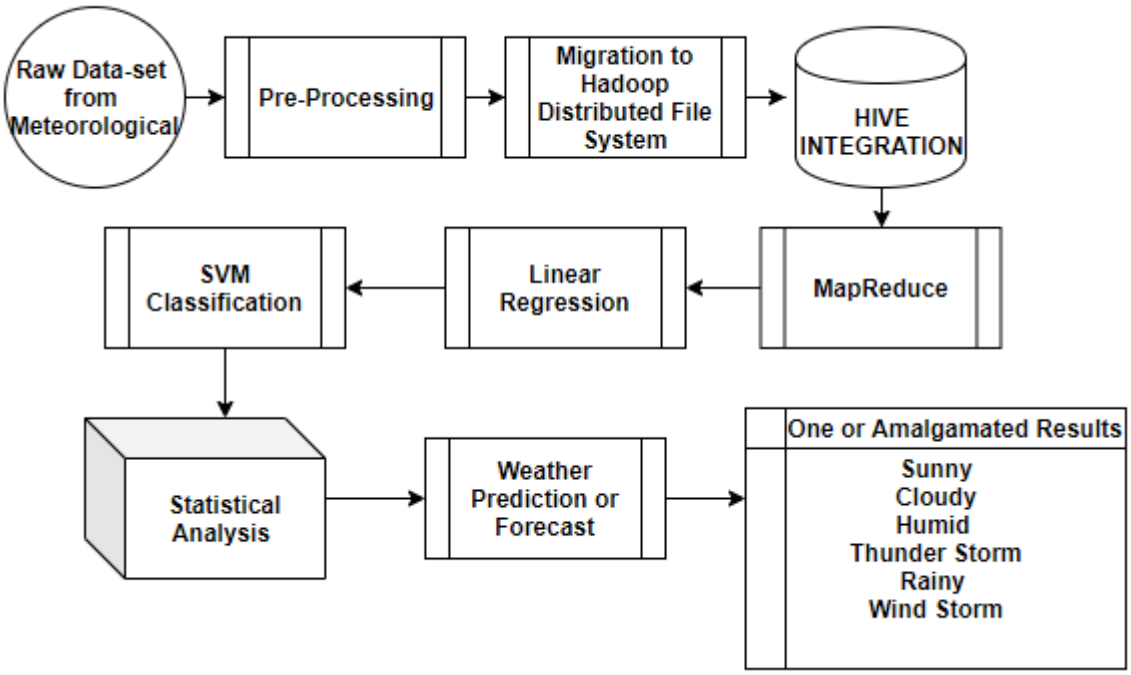

Figure 1 : Proposed Scheme comprising of Workflow Techniques using Hadoop Ecosystem (HDFS,HIVE and MapReduce) along with Machine Learning Models (Linear Regression and Support Vector Machine) for Weather Forecasting

REFERENCES

1. Abramson, Bruce, et al. "Hailfinder: A Bayesian system for forecasting severe weather."International Journal of Forecasting12.1 (1996): 57-71.
2. Cofno, Antonio S., et al. "Bayesian networks for probabilistic weather prediction."15th Eureopean Conference on Artificial Intelligence (ECAI). 2002.
3. Krasnopolsky, Vladimir M., and Michael S. FoxRabinovitz. "Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction."Neural Networks19.2 (2006): 122-134.
4. Lai, Loi Lei, et al. "Intelligent weather forecast."Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on. Vol. 7. IEEE, 2004. Ng, Andrew. "CS229 Lecture Notes Supervised Learning" 2016.
5. Radhika, Y., and M. Shashi. "Atmospheric temperature prediction using support vector machines."International Journal of Computer Theory and Engineering1.1 (2009): 55. "Stanford, CA" in Weather Underground, The Weather Company, 2016. [Online]. Available: https://www.wunderground.com/us/ca/paloalto/zmw:94305.1.99999. Accessed: Nov 20, 2016.
6. Stern, H. (2008), The accuracy of weather forecasts for Melbourne, Australia. Met. Apps, 15: 65?71. doi:10.1002/met.67
7. Wang Y. and Banavar S. "Convective Weather Forecast Accuracy Analysis at center and sector levels", NASA Ames Research center, Maffett Field, Califonia Weather.com,http://www.weather.com AccuWeather.com,http://www.accuweather.com
8. Anad M. "Prediction and Classification of Thunderstorms using Artificial Neural Network", International Journal of Engineering Science and Technology (IJEST), Vol.3 (5) May 2011.
9. Wiki (2013). Applications and organizations using hadoop. http://wiki.apache.org/hadoop/PoweredBy
10. Gartner Research Cycle 2014, http://www.gartner .com
11. K. Morton, M. Balazinska and D. Grossman, "Paratimer: a progress indicator for MapReduce DAGs", In Proceedings of the 2010 international conference on Management of data, 2010, pp.507–518.
12. Lu, Wei, et al. "Efficient processing of k nearest neighbor joins using MapReduce", Proceedings of the VLDB Endowment, Vol. 5, NO. 10, 2012, pp. 1016-1027.
13. J. Dean and S. Ghemawat, "Mapreduce: simplied data processing on large clusters", in OSDI 2004: Proceedings of 6th Symposium on Operating System Design and Implementation.12, pp267–288,1998.