Journal of Data Mining and Management
e-ISSN: 2456-9437
Volume 3 Issue 3

Mining Big Data: Future Forecast of Weather

[1]Apurva Sehgal, [2]Kanika Khurana, [3]Cherry Singh, [4]Ankur Kr Aggarwal
[1,2,3]UG Students, [4]Assistant Professor
Department of Computer Science and Technology,
Manav Rachna University, Faridabad, Haryana, India
Email: *[1]sehgal.apurva10@gmail.com, [2]kanika20khurana@gmail.com, [3]rtk082@gmail.com,*
[4]ankur@mru.edu.in

DOI: *http://doi.org/10.5281/zenodo.1745433*

Abstract

Climate gauging is a shot by meteorologists to anticipate the state of the air at some future time and expecting the atmosphere conditions. Climate gauging is the single 'most' imperative pragmatic purpose that meteorology exist as a science. Learning of climate information or atmosphere information in a locale is basic for business, society, agribusiness and vitality; utilized for both residential and business purposes .The estimate can advise a rancher the best time to sow the seeds for germination, it also helps to select the relevant data which has to be send to the airplane and also helps to determine that at what time the airplane has to take off or arrive. It acts lucrative for the occupants of the waterfront districts too, making them mindful of the events of storms. Climate information is considered utilizing information digging method for example 'The Clustering Technique'. By utilizing this procedure, we can locate the shrouded designs in the extensive dataset gathered and after that change the recovered data into the usable data for expectation of the climatic conditions. An assortment of information mining instruments and strategies are accessible in the business, however with a restricted use in the meteorological business.

Keywords: *Climate Data, Synoptic Data, Genetic Algorithm, ANN, Nearest Neighbor, Rule Induction.*

INTRODUCTION

Data mining is the eradication of the masked data from the voluminous databases. It is an extraordinaryrecent modernization which can examine the vital data in the databases. Weather data is of two types [1]: The first is Climate data- it is the official information record, which is given after some quality check is performed and the second type is, Synoptic data- it is the real time data that is used for the forecasting modeling. Atmosphere and climate influences the human culture in all the conceivable ways. For instance trim creation in agrarian, which is dependent on the water resources, i.e. Rain and the rain is one of the element of weather.

Climate estimating was done with the help of the barometric pressure, prevailing climate conditions and sky conditions with standard figuring's when the technology was not developed. In any case, now-a-days information mining, delicate registering is utilized to decide the future conditions. Yet at the same time the human sources of info are required to pick the most ideal conjecture which includes the example acknowledgment abilities, teleconnections and so forth.

There are two types of data mining the first one is **Descriptive data mining**- it depicts the assignment applicable informational collections in the brief, condense and useful shape. And the second one is **Predictive data mining**- it is based upon the data and analysis. In this the model is constructed for the database and then the weather is predicted based upon that model. [1]

Data mining performs the useful information extraction operation using different techniques proposed in a literature that includes six different techniques. [2]

Genetic algorithm
In the field of climate gauging, this technique is a search technique that mime the method of the survival of the fittest. This belongs to the upper category of biological process algorithms that produce solutions to improvement issues exploitation techniques impressed by natural growth.

Artificial neural network[3]
An artificial neural network is derived by the technique cellular sensual systems, for example, the cerebellum. It is formed of countless interconnected preparing components (neurons) working in at the same time to take care of particular issues. And is designed for a settled application, for example, information characterization, through a picking up embraced.

Nearest neighbor
It is otherwise called closeness look, likeness inquiry or nearer point seek.As the name suggests, it is the methodology problem for discovering the closest points.Closeness is typically shows in terms of a variation function. Explanation of the nearest search problem is suppose there is a given set S which contains points in the memory M and there is a query point in the set S which is Q .Now, we have to find the point which is closest to the query point Q. This is the nearest neighbor problem.

Rule induction
It is a coherent procedure in which the different premises are accepted to be valid or discovered genuine more often than not, and all these genuine articulations are combined to secure a particular end. This is frequently utilized as a part of the applications that include expectation, anticipating or conduct.

Memory based reasoning[4]
In this framework, the thinking is specifically done by utilizing databases in this manner have numerous preferences like no information from the master is vital. Embedding's and taking out learning and clarification to the appropriate responses can be effortlessly done.

Decision tree[5]
Also known as the choice tree which has the tree like structure. As there is a node and the branch in the tree where the node depicts the test and the branch depicts the outcome of that test. By using the decision tree the result can be easily computed.

BIG DATA ON WEATHER FORECASTING

Weather forecasting involves the prediction that how the climatic conditions will be in the near future.The present climatic conditions are acquiredfrom satellites radars.This collected information is send to meteorological centers and then this collected information is analyzed by using computers as computer draws charts, map out of those data collected.

Computers not only prepare charts and map but also predict that how those map looks in future. It is one of the utilizations of big data. Big Data is another term which is utilized to recognize the datasets that can't be overseen by the present techniques or information mining programming apparatuses because of their extensive size and mythologies.

There are 3 pillars to big data also known as V's - volume, variety, velocity. Relations of these V's with the weather forecasting are: **VOLUME**–the amount of data collected on daily basis from various sensors (temperature, humidity, air flow, etc).**VARIETY**- variety in the weather forecasting includes that which type of data is collected as in related to the temperature patterns, rainfall patterns etc.**VELOCITY**- it refers to that with what speed the data is collected and used to perform various prediction of weather forecast.

Gartner [18] in his manuscript have discussed about the big data as elevated volume, variety and velocity of data or information that demands commercial, novel types of techniquefor improved acumen and opinion making process system. Few of the applications of Big data is as follow [19]:

- Smart cities: wise management of natural resources, cities centered on viable advancement.
- Technology that may cut back the interval of knowledge.
- Business that may find churn and personalization of client wants.
- Health sector to enhance health condition of human by observance and mining polymer of every person.
- Agriculture by providing inputs to farmer about the crop cultivation and weather condition prior to any damage.

Weather Forecast Data acquisition Tools

Forecaster uses various sensor and tools to acquire the records and use thesedatasets to predict weather are thermometers, barometers, sling psychomotor, and many more. [7].

Thermometer: It is used to gauge the air temperature. The thermometer is a closed tube which contains the fluidic component like mercury. Whenever the air temperature rises the fluidic component rises accordingly and then the temperature is noted down at which the component stops. **Barometer:** It quantifies gaseous tension. It reveals to us climate or not the weight is rising or falling. A rising indicator implies bright or dry conditions while a falling gauge implies stormy or wet condition.**Sling Psych-meter:** It quantifies relative stickiness utilizing the cooling impact of dissipation. Two thermometers are utilized as a part of this. In this they wet the material of one of the thermometers and swing the psychomotor around the few times. Water vanishes from the material making the temperatures on the thermometer be lower than the other.**Rain Gauge:**It gauges the measure of rain that has fallen over a particular timeframe. **Wind Invane**: It is an instrument that decides the heading from which the breeze is blowing. **Anemometer**: It quantifies wind speed. The glasses get the breeze turning a dial joined to the instrument. The dial demonstrates the breeze speed. **Weather Maps**: It shows the environmental conditions over a substantial bit of earth surface. Meteorologists utilize climate maps to estimate the climate. **Hygrometer**: It gauges the water vapor substance of air or the mugginess. **Weather Balloon**: It quantifies the climate condition higher up in the air. **Compass**: It is a navigational instrument for discovering headings. **Weather Eather Satellites:** They are utilized to photo and track expansive scale air developments. At that point meteorologist incorporates and breaks down the information with the assistance of PCs.

FORECAST TYPES

A forecast type is a forecast that's designed to utilize a precise type of data. Kalaiselvi, in her manuscript have discussed four such types [2].

Synoptic weather prediction

Succinct alludes to the examination of the distinctive climate components inside the particular time of perception. This is otherwise called the customary approach. To monitor the changing climate the meteorological focus readies a diagram regular and it information from which this graph is made is acquired from a huge number of climate stations.It is still used today for short term predictions.

Persistence
Assurance determining depends on the idea that the present climate conditions can help us to anticipate the tomorrow's climate. They make the examination utilizing thermometers and barometers to get to the climate. This anticipating strategy works best in regions with predictable climate, for example, tropical zone or a cold district.

Statistical weather prediction
It allows meteorologists to make prediction based on previous trends. In this the forecasters mull over verifiable information about the average temperature, high temperature and low temperature. Also they search for authentic tempests records and precipitation sums and utilize these information for the premise of estimating.

Computer modeling
These forecast methods are the most advanced methods which are based on the mathematical formulas that if the meteorologist inputs the recent weather data then they can calculate the future conditions.

Numerical weather prediction
This forecast utilizes the energy of PCs to anticipate the climate and the intricate PC programs which keeps running on the super PCs.

FORECASTING METHODS

Ashwini Mandale, Mrs. Jadhawar in their manuscript have discussed about various forecasting methods [8].The methods in weather forecast are two as below:

Qualitative methods
These types of gauging techniques depends solely on the assessments, intuitions' and some sort of guessing. In these types of methods there is no sort of any scientific calculation involved.

Quantitative methods
These sorts of estimating techniques depend solely on the numerical models. They depends only on the calculations. These are of two kinds: time series method and associative method.

Time series method: Time series methodtakes the past patterns and with the help of these patterns the future patterns are predicted. **Associative method**: Associative method presumes that the variable being gauge is dependent on the different factors on the earth.

Time series method is further categorized into four categories:
Naïve Method- In this method the data recorded in the past serves as the basis of the future weather. It is typically used to check the results of more sophisticated forecasting method.

Simple mean (Average) –Utilization of a normal of every single pastdata as a forecast. In this method we take the average of the actual data of the previous year. This average will serve as the forecasted value. Eg (table 1) for year 2 the average will be the actual value of the previous year as it is. Now for the 3 year the average previous 2 year will serve as the forecasting value (365+310/2 = 337.5, 337.5 is the forecast value for the 3 year.)

Simple Moving Average –In this method the gauge for next specified period will be equivalent to the normal of a predefined period of the latest perceptions, with every perception getting a similar accentuation (weight).This method is somehow similar to the mean method but in this we take the actual value of the previous 2 year only.

But in the mean method we take the value of all the previous actual values. (table 1) For 4 year we take the actual value of 3 and 2 year which will serve as the forecast value of the 4 year (365+396/2=380, 380 is the forecast value of this year).

Weighted moving average – utilizes a normal of a predefined number of the latest perceptions, with every perception getting an alternate accentuation (weight). In this outline we use the 3-year weighted moving normal. As there is no starting point for us, we considered the year 1 gauging be (300).After that , we utilized a gullible technique to make a gauge for the two consecutive years as 310 and 365 respectively After this we had adequate information to let our 3year weighted moving normal conjectures unfurl consistently.

Table: 1. *Forecast data example*

Year	Actual Demand	Forecast	Note
1	310	300	We presume this forecast
2	365	310	This forecast is made using Naïve technique
3	395	365	Using a naïve approach
4	415	369.000	Using 3yr moving avg approach
5	450	399.000	

ALGORITHMS USED

Various algorithms are there to perform the operation to predict the weather forecast based on the input data, few such algorithms are discussed in manuscript by Rohit & Vikrant [21].

K-means: Weather foretelling, use K-Mean clump technique on needed data. For example in meteorology, if any new weather knowledge values comes, then we are able to use progressive K-means on that knowledge values. In this algorithmic rule, we have a tendency to deals with the reason of modification in the data worth that provides higher results than the progressive K-means clump that deals with the dynamical threshold values. K-mean clump algorithmic rule is applied to an efficient knowledgebase; the data might be usually updated. Progressive Kmean clump algorithmic rule is applicable on the dataset that's updated on regular basis and some new values square measure usually more to that. Therefore it deals with an oversized range of updates. Artificial neural network is additionally one in all the approaches for weather prediction.

Incremental Clustering:[20]Clustering is an unsupervised learning technique which partitions data into groups based on their similarities. The main function of incremental clustering is to assort unseen data samples into clusters based on their features.

Incremental K-means : K means technique is used to group the related data values in clusters. Firstly, start from the K cluster center, after that make clusters according to properties of the k Cluster centers. Next calculate themean valueof all the clusters, named as K-means. The main advantageof this technique is that we can add data at any time. When any new data is entered into the database then insert it into the nearest cluster. Then calculate the new mean value of that cluster.

HMM: HMM algorithm is a statistical model in which the states are presumed to be concealed. In these the relevant state is generated using probability association.

TOOLS FOR WEATHER DATA ANALYSIS

The big data event is basically identified with the software's that are openly available. There a substantial number of organizations as Facebook, Yahoo, Twitter, LinkedIn, MapR and Cloudera who are contributing devices for the information examination of huge information as open source ventures. Huge information foundation manages SAS, R, or other similar software's.

Apache Hadoop: is a product of Apache Hadoop [10] which is based on the MapReduce model and HDFC which is distributed file system. Apache Hadoop allows writing small set of code thatprocess huge amount of forecast information in a distributed set of clusters of computing nodes. The main concept of Hadoop (MapReduce) is that it divides the complete data set into small independent subsets of data that are processed by independent Map task on every node in parallel. The result of every node is reduced to a single level by Reduce task to obtain a final result over a distributed cluster system. The advantage of having a MapReduce job as independent task on node reduces the time to achieve a final conclusion for problem.

Apache S4: [11] there are several situations where forecast data is generated in a continuous rate or data streams. Apache S4 is specifically designed to deal with the data stream processing. Apache S4 is designed to join data streams from several data source and perform processing of data in real time.

Forecast or any other data set can be analyzed in big data mining; there are several application which are openly available to perform such analysis. Few of them are discussed below:

Apache Mahout: [13] it's a data mining open source and scalable machine learning software which is based on Hadoop. Mahout provide wide range of implementations of algorithms for machine learning and data mining.

R:[14] R is a software which is an open source programming language and condition which is particularly intended for measurable registering and representation of information.

Pegasus: [17] is a tool which is designed on top of MapReduce framework with big graph mining system. It is used to perform analysis of data to find the relevant patterns.

There are many more tools like **MOA, GraphLab.**

STEPS FOR FORECASTING DATASET SELECTION

Mandale, Jadhawar[9] in their literature discussed about the various steps that are involved in performing collection, cleaning, selection, transform and mining.

Data Collection: This is the first phase. In this phase the data is collected form the specific area for which the gauging has to be done

Data Cleaning: In this phase the relevant data is collected and the remaining ambiguous data or the missing data is ruled out.

Data Selection: At this phase, the data which is relevant to the analysis is selected from the dataset. There are 10 attributes in the meteorological dataset which are given in Table2.

***Table: 2.** Attributes of meteorological dataset*

Attribute	**Type**	**Description**
Year	Numerical	Year which is being considered
Month	Numerical	Month which is being considered
Wind speed	Numerical	Wind travelled in kilometers
Vaporization	Numerical	Vaporization
Cloud formation	Numerical	The mean of the cloud formation amount
Radiation	Numerical	The chunk of radiation
Sunshine	Numerical	The chunk of sunshine
Min temp	Numerical	The month wise minimum temperature
Rainfall	Numerical	Total rainfall in a month
Max temp	Numerical	The month wise maximum temperature

Data Transformation: This is also called data combination. As the name suggests that the data is transformed or changed into the types which are relevant for the data mining.

Data Mining Stage: All the above stages are for the preparation for this final stage in this stage the data is mined or we can say that the most relevant data is selected and remaining data is discarded

FUTURE SCOPE

The development of advanced forecasting system for severe weather events is ongoing. Researchers are exploring several approaches to the problem of very short range forecast that are highly specific in time. These approaches includes expert systems, numerical modeling of storm cells, etc.

CONCLUSION

As the qualitative and quantitative ways for the foretelling are mentioned however out of those 2 quantitative is that the best approach as in qualitative is just primarily based upon the opinions whereas in qualitative approach visionary do calculations by adopting the various ways like naïve methodology, exponential smoothing methodology, etc. and predict the atmospherical conditions.

REFERENCES

1. Gaurav J. Sawale, Sunil R. Gupta, Use of artificial neural network in data mining for weather forecasting, International journal of computer science and applications, April 2013.
2. P. Kalaiselvi, Weather forecasting – a survey, International Journal of Modern Computer Science (IJMCS), August, 2016.
3. "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies" /"I.J. Information Engineering and Electronic Business", 2012, 1, 51-59 Published Online February 2012 in MECS (http://www.mecspress.org/) DOI:10.5815/ijieeb.2012.01.07.
4. Divya Chauhan and Jawahar Thakur, "Data mining techniques for weather prediction: A review " / International journal on recent and innovation trends in computing and communication, (volume:2issue :8) IJRITCC| August 2014.
5. Han, J. Michelin K," Data Mining: Concepts and Techniques"/, San Francisco, CA: Morgan Kaufmann publishers, 2007.
6. Mining big data: current status, and forecast to the future bywei fan and albertbifet.
7. www.weatherkids.com
8. http://mech.at.ua/Forecasting.pdf
9. Ms. Ashwini Mandale, Mrs. Jadhawar B. A., Weather forecast prediction: a DataMining application**,** International Journal of Engineering Research and

General Science, March-April, 2015.
10. http://hadoop.apache.org.
11. [L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed Stream Computing Platform. In ICDM Workshops, pages 170–177, 2010.
12. http://storm-project.net.
13. http://mahout.apache.org.
14. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.
15. [A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa. cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.
16. C. Bockermann and H. Blom. The streams Framework. Technical Report 5, TU Dortmund University, 12 2012.
17. U. Kang, D. H. Chau, and C. Faloutsos. PEGASUS: Mining Billion-Scale Graphs in the Cloud. 2012.
18. http://www.gartner.com/it-glossary/bigdata.
19. Intel. Big Thinkers on Big Data, http://www.intel.com/content/www/us/en/bigdata/big-thinkers-on-big-data.html, 2012.
20. Big Data And Analysis Of Weather Forecasting System Simranjot Kaur Research student, M. Tech. (CE), Department of Computer Engineering, Punjabi University Patiala , Er. Sikander Singh Cheema Assistant Professor, Department of Computer Engineering, Punjabi University Patiala.
21. A Weather Forecasting Model using the Data Mining Technique, Rohit Kumar Yadav Vikrant Institute of Technology & Management, Indore RGPV University, Bhopal, Madhya Pradesh, India ,Ravi Khatri Vikrant Institute of Technology & Management, Indore RGPV University, Bhopal, Madhya Pradesh, India.

Cite this article as: Apurva Sehgal, Kanika Khurana, Cherry Singh, & Ankur Kr Aggarwal. (2018). Mining Big Data: Future Forecast of Weather. Journal of Data Mining and Management, 3(3), 10–17. http://doi.org/10.5281/zenodo.1745433