DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4149

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

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

Abstract: The whole world is suffering from the changing clement and their side. To reduce this side effects up to some extent there are many techniques and algorithms through which we can predict the weather on the basis of given data. After doing the analysis of existing techniques we conclude that data mining technique enable us to analyze the given set of data and extract the useful information from the given data. Therefore in order to understand the fluctuating patterns of weather conditions, a predictive model is studied. In this paper, we are using incremental K-means cluster algorithm which groups the same type data sets together and to prefigure the data we are using R-tool that gives structural data. At the end, result will be calculated on the basis of some mathematical conditions.

Keywords: Big Data, Clustering, Data Mining, Prediction, R-tool, Weather

1. INTRODUCTION

Big Data contains vast amount of data in the structured, semi-structured and unstructured form. That's why it is very difficult to process, manage and store to this type of data. In recent years different types of tools and techniques are there to handle Big Data. Data mining is one of them which we have used in this paper to manage weather related data. In this paper we have used this data mining technique in the prediction of weather [1].

Now days, people of India suffering from changing clement and their side effects. Normally in agriculture field, farmers are facing many problems due to unexpected weather conditions [21]. Weather forecasting is directly depend upon the natural molecules present in the air like Ozone(O3), Nitrogen dioxide(NO2), Carbon Dioxide (CO2), Sulfur dioxide (SO2) etc [19]. In this paper we have focused on specific region in Punjab [2]. To reduce these side effects up to some extent there are many techniques and algorithms through which we can predict the weather on the basis of given data. Data mining technique is used in Weather prediction process. Weather is most effective environmental constraint in every phase of our human life. So weather forecasting is going too used in many fields like Food security disasters, Agriculture and science. In earlier years we have not any exact idea about weather conditions. So in those days, we faced many problems in food management process, industry and agriculture field. But, now in the era of advancement we have many ways to find weather conditions. This is the reason behind applying data mining techniques to find the weather conditions [6].

Under the section 2 of this paper we deal with previous work used to prefigure weather conditions. Further we have discussed the basic techniques which have applied in this paper. These techniques help in predicting the weather conditions of "Punjab" city. Different matters and their effects on weather are described in the next section. Then we have described the algorithm used in this paper. After that the application of new approach on air pollution data and analysis of the resultant prediction are described. Overall work and future scope is concluded in the last section [3].

As we know that not every system or predictive technique gives us 100% accurate results. Data Mining just helps us in Decision making, but the final decision always depends on the user. So it is not completely reliable [4].

Related Work

Many different applications, like vote prediction, population census, weather forecasting; use K-Mean clustering technique on required information. For example in weather forecasting, if any new weather data values comes, then we can use incremental K-means on that data values. In this approach, we deals with the reason of change in the data value that gives better results than the incremental K-means clustering which deals with the changing threshold values. K-mean clustering algorithm is applied to a dynamic database; the data may be often updated. Incremental K-mean clustering algorithm is applicable on the dataset that is updated on regular basis and some new values are often added to it. So it deals with a large number of updates. Artificial neural network is also one of the approaches for weather predictions [7].

R tool

R is said to be a very powerful statistical programming language which is free of cost. R tool consists of broad range of graph-drawing tools that helps us to produce standard graphs of any data [11].

Incremental Clustering

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 [17].

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 the mean value of all the clusters, named as K-means. The main advantage of 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. Incremental clustering is designed using clusters metadata captured from results of the K-Means [13].

2. AIR MOLECULES EFFECT ON WEATHER

NO2: It emits high temperature combustion when reacts with O2,
NO2 + O2 → NO + O3 or NO2 → NO + O, then it increase in the amount of NO in air molecules which makes weather cold and dry.

SO2: The industrial processes like burning of coal and petroleum release sulphur in the air. If it increases in excess, the atmosphere could be smoggy, fog and chance of acid rain, and also cold air or calm winds blow with high humidity [12].
CO2: Carbon Dioxide acts as main pollutant element in the air molecules. With the increase of CO2, the atmosphere becomes smoggy, humid and hot. So when carbon dioxide reacts with electron particles, it can create carbon monoxide that affects the weather conditions also.
O3: Ozone is basically formed on hot and sunny days when the air is dynamic.
PM 10: These are small particles of diameter 2.5 to 10 micrometers and contribute in increasing pollution as dust particles.

3. SYSTEM DESIGNED

Weather forecasting system provides us the information about future weather conditions for a particular region, locality over a specified time period. Weather directly depends upon the air molecules which can absorb high frequency sunrays. The air molecules data is collected by the system periodically after every one hour [5]. The R tool uses these raw data to bound large information to find the "Structural Air Pollution Database". Then we have applied k-mean algorithm to this database. Then we have stored the resultant data in the main database. So accordingly, we classified the main database into four regions that are grouped according to the direction of wind flow over the year [10]. First region includes December, January, February; second region includes March, April; in the third region May, June, July; and in the last region August, September, October, November are included. First region is considered as winter region. Second and fourth are known as temperate region. Third is called summer region. When we have to search any data, then we can search it in its particular domain. In the k-mean algorithm we have organized data into clusters according to the region. Incremental k-mean algorithm is used for the addition of any new data and it makes the data fit into the proper cluster. The initial cluster center is generated by using the genetic algorithm. When the user enters data to the system, then it is compared with the previous set of data using the priority based algorithm. Multiple year data is stored in the database, now with the help of [(1/3-α), (α), (1/3)] we can use the order of priority [9]. With the help of the record of statistics, we can conclude that weather basically depends on 2nd last year. So choose priority (1/3-α), (α) and (1/3) for the last year, 2nd last year (highest priority) and 3rd last year respectively. Prediction calculation has been done based on three years, where α is taken as a constant variable [14].

4. ALGORITHM

1. Collect data (The values of NO2, O3, CO2, and SO2) after every one hour and save it in the original database [18].
2. After every two hours, previously stored data is converted into structural data by using R tool. All type of data that is stored in database is finally stored in the modified "structural air pollution database".
3. "Structural air pollution database" is further divided into four sub-databases on the basis of weather separation [8].
4. Cluster centers are produced with the help of genetic algorithm (GA) after applying K-Mean clustering algorithm to structural data.
5. Whenever any new data is added into the database then use incremental K-Means clustering to handle the new data addition.
6. Find the resulting clusters.
7. By using priority based algorithm, prediction of results can be done for different years (max three years).
8. By using threshold temperature value ranges, we can forecast the probable weather condition for a particular time period [15].

5. RESULTS AND ANALYSIS

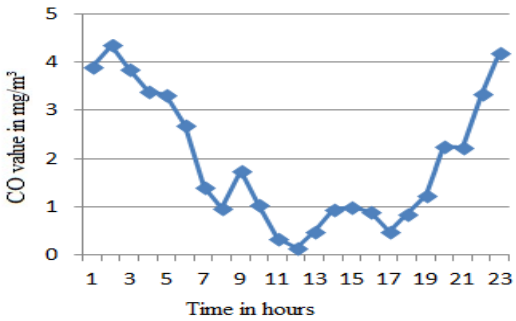

Fig: Co values after every one hour

Table: 1 CO value every one hour 02.01.17

Time	**mg/m³**
01:00	3.90
02:00	4.35
03:00	3.85
04:00	3.40
05:00	3.33
06:00	2.70
07:00	1.40
08:00	0.98
09:00	1.75
10:00	1.05
11:00	0.35
12:00	0.15
13:00	0.50
14:00	0.95
15:00	1.00
16:00	0.90
17:00	0.50
18:00	0.85
19:00	1.25

20:00	2.25
21:00	2.23
22:00	3.35
23:00	4.2

Use R tool to plot above given graph, the values which we have taken are from the outer part (3.90, 4.35, 3.85, 3.40, 3.33, 2.70, 1.40, 0.98, 1.75, 1.05, 0.35, 0.15, 0.50, 0.95, 1.00, 0.90)

So the approx value of CO on 02.01.2017= (3.90+4.35+3.85+3.40+3.33+2.70+1.40+0.98+1.75+1.05+0.35+0.15+0.50+0.95+1.00+0.90)/12

=2.49

$$\text{Approx. calculated value} = \frac{\text{Sum of outer region point}}{\text{Total number of outer region point}}$$

We have stored the values of air molecules (carbon monoxide) after every one hour in the table 1 [20]. Then calculate the average value of all the other air molecules and save in the next table 2.

Table 2: (Data stored in daily basis)

Date	CO	NO2	O3	PM10	SO2	**Prob. Temp °C**
1.1.17	1.98	160.70	20.15	230.23	5.27	23-26
2.1.17	2.49	185.97	17.15	228.12	6.28	23-26
3.1.17	1.50	115.60	19.16	170.12	5.12	21-25
4.1.17	1.35	90.62	29.48	149.54	4.50	21-25
5.1.17	1.89	150.35	20.11	210.54	5.11	21-25
…..	….	…..	…..	…..	…..	…..

Table 3: Cluster wise ranges of air molecules

Cluster ID	CO2 Cluster range	NO2 Cluster range	O3 Cluster range	PM 10 Cluster range	SO2 Cluster range
Cluster 1	0.7-1.50	55-110	7-10	45-120	2.50-3.40
Cluster 2	1.50-2.5	110-175	10-17	120-170	3.40-4.6
Cluster 3	2.5-3.45	175-195	17-20	170-200	4.6.5.5
Cluster 4	3.45-4.15	195-215	20-26	200-230	5.5-6.9
…….	……	……	……	…….	…….

Table 4: Weather category after clustering (resultant table)

Date	Probable temp. range in °C	Actual temp in °C	Weather category
1.1.17	25-28	25	Haze, Smoggy, fogs, and smoke
2.1.17	25-28	26	Smog, dusty, fog and mist
3.1.17	20-24	22	Dry, smog and mist
4.1.17	20-24	23	Mist, haze and smoky

We have introduced four clusters. With the help of cluster we get to know about the description of day like hot, hazy, fogy, dusty etc. So depending upon the weather category and the on the basis of the regional database, we can predict temperature for upcoming days [16].

2. ACCURACY CALCULATION

$$\text{Accuracy} = \frac{\text{Number of matched records}}{\text{Total number of records}} \times 100$$

$$= \frac{280}{365} \times 100$$

$$= 76\% \text{ (approx)}$$

6. CONCLUSION

With the help of K-mean clustering algorithm and R tool, we have introduced a new technique to forecast the weather of upcoming days. This technique is also suitable for the dynamic environment where the weather conditions change frequently. The genetic algorithm was used by us to find out or to make guess of initial cluster center. It has given us more suitable results. This technique does not give completely accurate results; it just forecasts the probable results. In the future work, it can be extended to any immense data sets with varied attributes/ parameters for effective analysis and correct prediction. We can use this approach in dealing with some other air pollution databases of different regions for weather forecasting.

7. REFERENCES

[1] Kaur, M. (2013). Big Data and Methodology-A review. International Journal of Advanced Research in Computer Science and Software Engineering, 5.

[2] Sabia, S. K. (2014). Applications of Big Data. International Journal on Advanced Computer Theory and Engineering (IJACTE), 5.

[3] Yuvraj S. Sase, P. A. (2014). Big Data Implementation Using Hadoop and Grid Computing. International Journal of Innovative Research in Science, Engineering and Technology , 6.

[4] Harshawardhan S. Bhosale, P. D. (2014). A review paper on Big data and Hadoop. International Journal of Scientific and Research Publications , 7.

[5] C.Chandhini, MeganaL.P , P. A (2013). Grid Computing-A Next Level Challenge with Big Data. International Journal of Scientific & Engineering Research , 7.

[6] Gadekar, H. S. (2014). A Review paper on Big Data and Hadoop. International Journal of Scientific and Research Publications , 7.
[7] Singh, V. S. (2015). Big Data: Tools and Technologies in Big Data. International Journal of Computer Applications , 5.
[8] Kuchipudi Sravanthi & Tatireddy Subba Reddy. (2015). Applications of Big data in Various Fields. International Journal of Computer Science and Information Technologies , 4.
[9] Patil, S. (2016). Big Data Analytics Using R. International Research Journal of Engineering and Technology (IRJET) , 4.
[10] Hitesh Goyal, S. S. (2015). Big Data Analysis Using R (Big Data Analysis Applications, Challenges, Techniques). International Journal of Advanced Research in Computer Science and Software Engineering , 6.
[11] Melita Hadzagic, M.-O. S.-H. (2013). Maritime traffic data mining using R. 7.
[12] M. R. Bendre, R. C. (2015). Big Data in Precision Agriculture : Weather Forecasting for Future Farming. 1st International Conference on Next Generation Computing Technologies, (p.5).
[13] Basvanth Reddy, P. B. (2016). Weather Prediction Based on Big Data Using Hadoop Map Reduce Technique. international Journal of Advanced Research in Computer and Communication Engineering , 5
[14] Bolandnazar, B. K. (2015). A clustering model based on an evolutionary algorithm for better energy use in crop production. CrossMark , 15.
[15] Audireddy Gayathri, M. R. (2016). A survey on Weather forecasting by Data Mining. IJARCCE , 3.
[16] Barak, B. K. (2015). A clustering model based on an evolutionary algorithm for better energy use in crop production. CrossMark , 15.
[17] Bolandnazar, B. K. (2015). A clustering model based on an evolutionary algorithm for better energy use in crop production. CrossMark , 15.
[18] Divya Chauhan, J. T. (2014). Data Mining Techniques for Weather Prediction: A Review. International Journal on Recent and Innovation Trends in Computing and Communication , 6.
[19] P. Kalaiselvi, D. G. (2016). Weather Prediction Using J48,EM And K-Means Clustering Algorithms. IJIRCCE , 7.
[20] Ratul Dey, C. D. (2011). Weather forecasting using Convex hull & K-Means Techniques – An Apparoach. Kolkata: IJARCCE.
[21] Yadav, R. K. (2016). A Weather Forecasting Model using the Data Mining Technique. International Journal of Computer Applications , 9.

