Mody University International Journal of Computing and Engineering Research
2 (2), 2018, 68-73
http://www.publishingindia.com/micer

An Efficient Approach for Rainfall Forecasting Using Data Mining

Satyajee Srivastava[1*] and Vivek Kumar[2]

[1]School of Computing Science and Engineering, Galgotias University, Gautam Buddh Nagar, Uttar Pradesh, India. Email: drsatyajee@gmail.com
[2]School of Computing Science and Engineering, Galgotias University, Gautam Buddh Nagar, Uttar Pradesh, India. Email: viveksaini948@gmail.com
*Corresponding Author

Abstract: Farming as the major occupation of India requires a scientific and analytic focus for its productivity. Agriculture and allied sectors like forestry and fisheries accounted for 15-19% of the GDP (Gross Domestic Product) [1]. India is a country having various seasons and different geographical conditions. India's climate can be classified as a hot tropical country, except the northern states of Himachal Pradesh and Jammu and Kashmir in north and Sikkim in the north eastern hills, which have a cooler, more continental influenced climate. These variations in climate make prediction of weather difficult. Techniques like Machine Learning algorithms are required for studying and predicting the weather conditions. This paper represents an analysis and prediction of rainfall by studying the previous data accumulated in 100 of years [2]. This paper uses basic techniques of Data Mining to conduct a trend analysis on Rainfall Data. We have analyzed data of various regions, implemented suitable data mining techniques based on literature survey to achieve our goal of analyzing and predicting the Rainfall. Hierarchical clustering is used for grouping and clustering similar data together. Regression technique is used to predict the range of values, Graph Extrapolation technique is used to estimate value and to predict the future pattern.

Keywords: Analytics, Clustering, Extrapolation, Prediction.

I. Introduction

Farming is the basic and major factor for deciding the fate of a country. So, its productivity is to be insured to boost the country's economy and farmers' income. It is important give the farmer a clear picture that how the climate is going to vary for the current and upcoming years. So, that he can choose his crop wisely that it would be more profitable to him [3].

Nearly 50% of yield is attributed to the influence of climatic factors. The following major atmospheric weather variables which influences the crop production.

a. Rainfall
b. Temperature
c. Atmospheric humidity
d. Solar radiation [4]

In this paper, an attempt has been made to predict the rainfall / precipitation by studying and analysing the historical data using data mining techniques as Clustering, regression etc., advanced graph techniques as Extrapolation and mathematical polynomial equation. The data sets used in this paper have been collected for years and are mined using data mining tools, to give a meaningful information.

As the prediction only takes into account the historical data ignoring the Al-Nino and La-Nina effect that occurs occasionally and may cause the unexpected flood and drought conditions.

Monsoon is the major season for rainfall which accounts for more than 80% of the total precipitation [5].

II. Data and Methodology

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is a process where certain techniques and methods are applied to extract data patterns [6]. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Here, the data set is collected from various meteorological stations spread across all over India. The data is collected on all days of the year more than 90% of the year. The Data for across 100 of the year is collected on all days that is 365 days. Because, the data became so huge it was very difficult to study.

The data is area weighted average all over India annually in mm. To study the data, we want it to be grouped and classified. For grouping we used Hierarchical Clustering.

It is the technique which is used to cluster the data for meaningful mining that builds hierarchy of cluster. In this, all the data points assigned to a cluster of their own, then two nearest clusters are merged to form a single cluster. This terminates when only single cluster is left.

Here, data collected on daily basis is considered as the base cluster. Then, these are merged to form clusters according to the Hierarchical clustering. This is continued till we got data on monthly and yearly basis. This data is used for mining purpose. Since, Rainfall data is collective data that is formed from day wise rainfall poured, this suggests use of hierarchical clustering.

From the data collected, a graph is plotted in terms of years and total rainfall in mm. Now, to predict the rainfall from the year we need the graph to extrapolate. Extrapolation is the process of predicting information about a point outside a curve when few points on the curve are given.

Since, graph is plotted the trendline in the curve can be represented as a mathematical equation.

It has an advantage that given the equation, predicted value can be calculated out using by just substituting the input value in terms of year.

III. Proposed Approach

The data set is collected from the various sources as area weighted average rainfall in India annually. The data is then clustered using Hierarchical Clustering. The clustered data is then plotted on the graph. From the trendline in the graph which is formed by using Regression technique, the graph is extrapolated. The extrapolated graph gives us data for upcoming years. The data predicted is compared with the previous data for analysis as what will be the rainfall trend in the upcoming years.

IV. Implementation

The data is first clustered using Hierarchical clustering annually. The value that we get is the yearly area weighted rainfall for that year. Each cluster represents the year with its rainfall. The data that we have is from 1901 to 2013. The clusters are analysed and not merged because each year has certain factors and condition on which the rainfall depends and is unique. As can be seen from the data in 1972 the rainfall was minimum 947.1 and a maximum 1463.9 mm in 1917 owing to certain unpredictable condition. The data in a year is also clustered.

Now rainfall against the year is plotted to form the given graph. Since the graph is very scattered and has many ups and downs, so a trendline is also drawn to give it a static look and that also make the graph easy to plot. The given graph has a trendline which is extrapolated till 2050. The trendline is a polynomial equation, extrapolated using numerical methods. The polynomial curve can be extended after the end of the given data. The polynomial extrapolation is usually done by *Newton's process of finite difference* or with the use of *Lagrange's interpolation* formula [7].

The future values till 2050 are estimated. The Lower and upper bound on the prediction is also estimated and predicted using Regression techniques. Certain mathematical equations are derived from graph using certain computational techniques. The data is then analysed to give the conclusion.

V. Results

A. Tables

Table I: Collected Data Using Hierarchical Clustering [8]

Year Annual (in cms.)	Year Annual (in cms.)	Year Annual (in cms.)	Year Annual (in cms.)
1901	1030.8	1961	1399.2
1902	1038.4	1962	1198
1903	1195.9	1963	1220.9
1904	1025.1	1964	1244.4
1905	977.5	1965	947.4
1906	1149.2	1966	1058
1907	1034.8	1967	1154
1908	1077.4	1968	1059.3
1909	1128.5	1969	1147.8
1910	1183.9	1970	1255
1911	1028.9	1971	1216.9
1912	1070.4	1972	947.1
1913	1061.8	1973	1219.5
1914	1185.9	1974	1055.3

Year Annual (in cms.)	Year Annual (in cms.)	Year Annual (in cms.)	Year Annual (in cms.)
1915	1124.4	1975	1294.8
1916	1324.8	1976	1131.6
1917	1463.9	1977	1269.7
1918	1020.2	1978	1237.2
1919	1287.9	1979	1030.2
1920	1039.1	1980	1182.3
1921	1225	1981	1170.7
1922	1204.2	1982	1084.4
1923	1148.6	1983	1320.9
1924	1245.9	1984	1160.8
1925	1189.5	1985	1144.9
1926	1226.2	1986	1137.6
1927	1244.6	1987	1088.9
1928	1200.2	1988	1342.1
1929	1193.2	1989	1127.4
1930	1198.5	1990	1401.4
1931	1292.8	1991	1170.2
1932	1202.9	1992	1102.7
1933	1372	1993	1207.8
1934	1217.5	1994	1295.3
1935	1127.9	1995	1242.4
1936	1321.8	1996	1182.9
1937	1204.4	1997	1183.1
1938	1290.5	1998	1208.8
1939	1111.6	1999	1116.6
1940	1201.3	2000	1035.4
1941	1073.9	2001	1105.2
1942	1272.9	2002	981.9
1943	1269.2	2003	1243.6
1944	1298.5	2004	1080.5
1945	1222	2005	1208.3
1946	1337.2	2006	1161.6
1947	1236.3	2007	1179.3
1948	1342.2	2008	1118
1949	1269.6	2009	953.7
1950	1174.2	2010	1215.5
1951	1060.6	2011	1116.3
1952	1110.1	2012	1054.7
1953	1222.1	2013	1092.5
1954	1181.4		
1955	1275.4		
1956	1362.6		
1957	1131.9		

Year Annual (in cms.)	Year Annual (in cms.)	Year Annual (in cms.)	Year Annual (in cms.)
1958	1312.3		
1959	1376.9		
1960	1154.8		

(Source- https://data.gov.in/catalog/rainfall-india)

The above table has been formed by hierarchical clustering.

This table can be further minimized and clustered but since we have to forecast the rainfall yearly so we keep the generalisation to this extent only. The Table II shows the forecasted values in terms of annual, maximum and minimum rainfall for the particular year. The values are forecasted till 2049 since various geographical and climatic changes can occur in this wide time interval.

Table II: Forecasted Table Using Trend Analysis

Year	Annual	Minimum	Maximum
2014	1122.816824	919.38	1326.25
2015	1122.802091	917.74	1327.87
2016	1122.787357	916.07	1329.50
2017	1122.772623	914.40	1331.14
2018	1122.757889	912.71	1332.80
2019	1122.743155	911.02	1334.47
2020	1122.728421	909.30	1336.15
2021	1122.713687	907.58	1337.85
2022	1122.698953	905.85	1339.55
2023	1122.684219	904.10	1341.27
2024	1122.669486	902.34	1343.00
2025	1122.654752	900.57	1344.74
2026	1122.640018	898.79	1346.49
2027	1122.625284	897.00	1348.26
2028	1122.61055	895.19	1350.03
2029	1122.595816	893.37	1351.82
2030	1122.581082	891.55	1353.62
2031	1122.566348	889.71	1355.43
2032	1122.551614	887.86	1357.25
2033	1122.536881	886.00	1359.08
2034	1122.522147	884.12	1360.92
2035	1122.507413	882.24	1362.77
2036	1122.492679	880.35	1364.64
2037	1122.477945	878.44	1366.51
2038	1122.463211	876.53	1368.40
2039	1122.448477	874.60	1370.30
2040	1122.433743	872.66	1372.20
2041	1122.419009	870.72	1374.12
2042	1122.404276	868.76	1376.05
2043	1122.389542	866.79	1377.98
2044	1122.374808	864.82	1379.93
2045	1122.360074	862.83	1381.89
2046	1122.34534	860.83	1383.86
2047	1122.330606	858.82	1385.84
2048	1122.315872	856.81	1387.82
2049	1122.301138	854.78	1389.82

B. Graphs

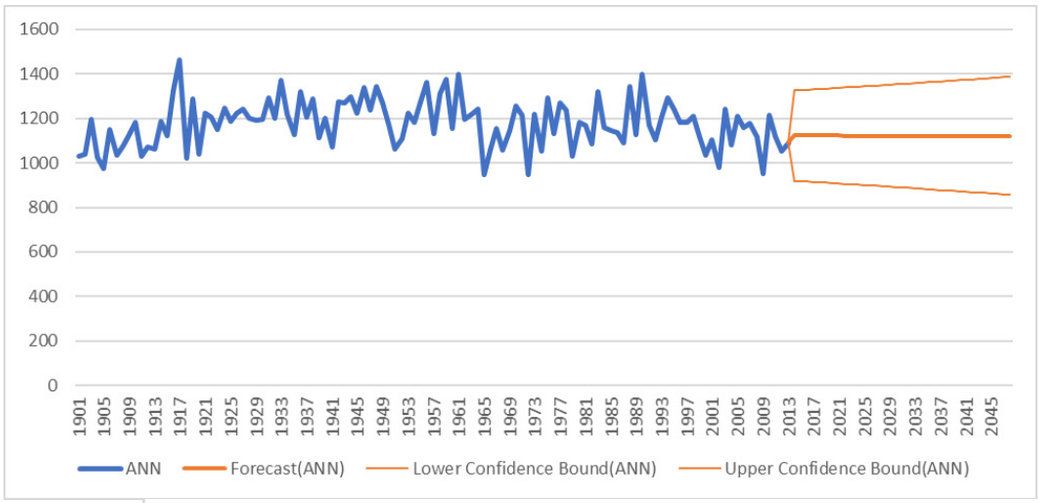

Fig. 1: Graphical Representation of the Forecasting Till 2050

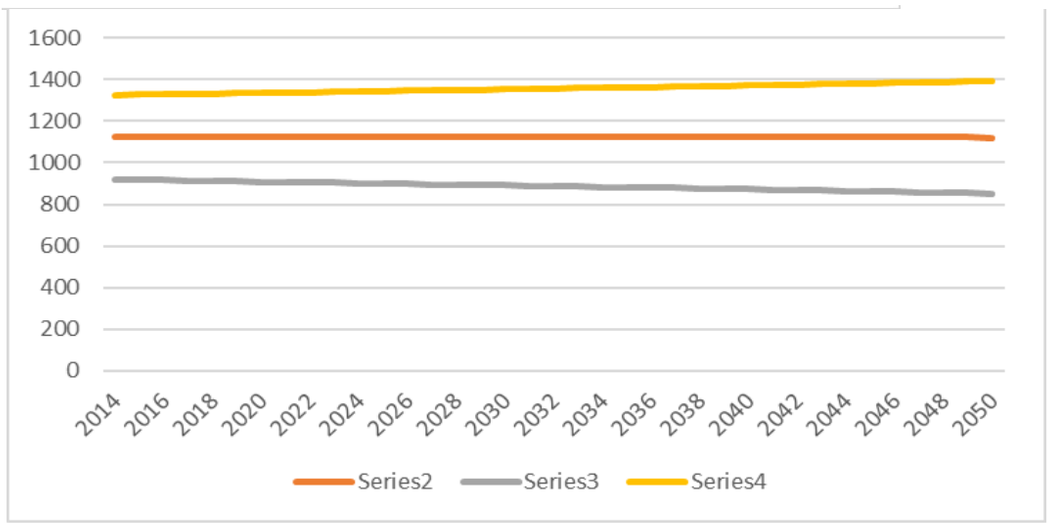

Fig. 2: Graph for the Extrapolated Trendline Till 2050

The graphs are plotted using the R Studio. The input data is processed, and a trend analysis is performed to extrapolate the graph to forecast the graph till 2045 in the Fig. 1. Various operations are also performed that also predict the maximum and minimum rainfall for a particular year in Fig. 2. We can see in Fig. 1 that the line after 2013 sees a downward slope. Also, in the Fig. 2 we can that the line series 3 representing minimum rainfall have a steeper downward slope than the series 4 have upward slope which means that the maximum is increasing but not having much increasing slope whereas the minimum is going downward with much decreasing slope as compared to series 4.

C. Equations

Polynomial Equation that we got from the trendline till 2013

$$y=-2E\text{-}05x^4+0.1491x^3-439.62x^2+576133x-3E+08 \quad (1)$$

$R^2 = 0.1912$ which when converted to a 2nd order polynomial equation gives

$$y=-0.0417x^2+163.07x-158235 \quad (2)$$

$R^2 = 0.1413$

This equation when analyzed mainly with a derivative test its 2nd derivative was found to be negative:

Y=F(x)

F''(x)=-3(139/2^3.5^4)

The second derivative for a curve shows that the curve is concave down which clearly shows that in the future following the trend the Rainfall in the India is going to decrease. The mathematical analysis of the equation also shows that the curve

has already seen a global maximum in period 1955 to 1961 which is confirmed by the data and the peaks in the graph for annual rainfall, also highlighted in the graph.

The plotted graph and predicted data of also confirms the fact that Rainfall in India will be seeing a downfall as comparison to previous year's figures.

VI. Conclusion

The Trend analysis technique shows that the rainfall in India is decreasing, though the change is not drastic but eventually and slowly it is going to decrease. Since nothing can be predicted 100% true, the paper presents a bigger mathematical picture and data-based trend study on the previous data present for rainfall. The data can't be forecasted for coming 100 of years because many geographical condition changes, also man-made situations can't be predicted. Pollution and Deforestation are some of the main factors for decreasing rainfall [9]. Either the agriculture and rainfall dependent occupation must find another source for water or change their methods which will consume less water. Since every crop is important, their water consuming capacity can't be changed but certain better methods such as drip irrigation and techniques can be made in use for better results. Also, adoption of Green Technology will produce better results [10]. Each must know that rainfall can have better figures if we focus on pollution reduction, afforestation and adopting better techniques and methods which are less water consuming so that we will make us less dependent on rainfall. Using ground water for the practices is just a temporary solution to the problem.

References

[1] https://en.wikipedia.org/wiki/Economy_of_India

[2] T. V. R. Kanth, V. V. S. S. S. Balaram, and N. Rajasekhar, "Analysis of Indian weather data sets using data mining techniques," *Computer Science & Information Technology*, pp. 89-94, 2014.

[3] H. D. Shannon, and R. P. Motha, "Managing weather and climate risks to agriculture in North America, Central America and the Caribbean," *Weather and Climate Extremes*, vol. 10, part A, pp. 50-56, December 2015.

[4] S. Neenu, A. K. Biswas, and A. S. Rao, "Impact of climatic factors on crop production - A review," *Agricultural Reviews*, vol. 34, no. 2, pp. 97-106, 2013.

[5] https://en.wikipedia.org/wiki/Climate_of_India

[6] J. H. Friedman, "Data mining and statistics: What's the connection?," in *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*, 1997.

[7] M. Milman, *Extrapolation and Optimal Decompositions with Applications to Analysis*, 1994.

[8] https://data.gov.in/catalog/all-india-area-weighted-monthly-seasonal-and-annual-rainfall-mm

[9] S. Paul, S. Ghosh, R. Oglesby, A. Pathak, A. Chandrasekharan, and R. Ramsankaran, "Weakening of Indian summer monsoon rainfall due to changes in land use land cover," *Scientific Reports,* vol. 6, article no. 32177, 2016.

[10] S. Srivastava, and R. Srivastava, "Adoption of Green Information Technology (GIT) in India - A current scenerio," *Journal of Information and Operations Management*, vol. 3, no. 1, pp. 61-63, 2012.