

Short-Term Rainfall using Logistic Regression Algorithm

Nishanthi P.^{1*} and Sudha Rajesh²

¹Master of Computer Application, Department of Computer Application, B. S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Chennai, Tamil Nadu, India. Email: nisha1999pk@gmail.com

²Assistant Professor, Department of Computer Application, B. S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Chennai, Tamil Nadu, India. Email: sudharajesh@crestcent.education

*Corresponding Author

Abstract: Farmers and all peoples are expecting for the good rainfall, for farming and for usage of daily water. But for the past decades, we didn't get much rainfall and all peoples suffered a lot. So, a work is proposed to predict the rainfall. Rainfall data is collected from the hydrological data to predict the storm warning. This is study as an idea as it consumes the large amount of records from the different various distributed system. Spatial temporal characteristics using a Map Reduce Framework to manage the database. The workload is classified using Support Vector Machine (SVM) is classified for workload. It uses the reduction algorithm for dataset and use for feature selection. Different rainfall concept prediction is executed using the big rainfall data. The impact of the dataset parameters are classified into locally, hourly and overall rainfall storms. The proposed system is deals with hadoop tool for predicting a rainfall data from a large amount of data it has no limitation data in a structure manner. The result will be improve the performance of the proposed system it has more accuracy and efficiency.

Keywords: Map reduce framework, Reduction algorithm, Support Vector Machine (SVM).

I. INTRODUCTION

Rainfall is the amount of precipitation, in the form of rain (water from clouds), that descends onto the surface of Earth, whether it is on land or water. It develops when air masses travel over warm water bodies or over wet land surfaces. Atmospheric turbulence and convection carry the moisture, or water vapor, upward into air masses where they form clouds. The clouds eventually release this water vapor, which is dropped as rainfall.

Falling rain is usually composed of droplets with diameters of about 0.02 in (0.5 mm) or greater. A group of smaller raindrops is called drizzle. It is not considered rainfall when Precipitation from clouds evaporates on the way down and fails to reach Earth's surface. The meteorological processes that control rainfall are

studied intensely in order to learn more about how rainfall is distributed locally and globally.

Rainfall is important in the hydrological cycle, which is the continuing movement of water below, on, and above the surface earth. The water are stored commonly in oceans, also in lakes, rivers, etc. our peoples nowadays, all are storing the water in our homes we are drinking, bathing, cooking, and some other uses.

II. PROBLEM DEFINITION

In existing system they are using SVR and Multilayer preception, MSE, MAE algorithm and also they predict for Monsoon season only. In existing system backend dealing with MYSQL, the data will be lost and we cannot recover the data if having limitations of data in existing system.

In proposed system we are used logistic regression Algorithm and we predict for monsoon and non monsoon season. Our proposed system backend deals with hadoop. If the data are loss we can easily recover the data. No limitation of data in Proposed system by using hadoop.

III. EXISTING SYSTEM

The existing idea supervises giving back end by make use of MYSQL which contains lots of drawbacks that is the information precondition is the executing time is high when the information is large and once information is lost, we cannot recover the data. Very difficult to recovery the data, the data are having limitations. The result will take more time to execute. The cost is very high. So accordingly proposing thought by utilizing Hadoop structure.

Disadvantages of Existing System

- Limitations of data's we can easily processed.
- Getting a results taking a more time.
- Very high cost and maintenance.

IV. RELATED WORK

The scope of the project is to predict the short-term rainfall we are getting the rainfall data for annual prediction for Odisha state they are using multiple regression analysis support vector machine model show the higher performance than the MLP. Multiple regression analysis for non monsoon rainfall prediction for 5 months, January, February, April, October and November these are all non monsoon seasons. SVR is more accurate than the MLP and SVM, it is very useful for future rainfall prediction. For rest of 3 non monsoon months, March, May and December Multilayer perceptron is used for future rainfall prediction it is more accuracy. Xiaobo Zhan, Sachin Mohanty, Ajaya Kumar Parida, Subhendu Kumar Pani, Bin Dong, and Xiachun Cheng (2019).

Tirupan Madal, Vinayakam Jothiprakash (2012) [1] In this approach they are using model has a ANN and MT model it is very promising techniques for daily rainfall prediction. They just using only beyond rainfall and without any other rainfall data.

Jae-Hyun Seo, Yong Hee Lee, and Yong-Hyuk Kim (2014) [2] In this approach DE better than a result of GA, in this approach we can also use a SVM with polynomial kernel GA experiments showed the best results on average. Select the important feature for set a validation, the features are selected they are used to predict the short-term rainfall forecasting.

Jyothis Joseph, Ratheesh T. K. (2013) [3] short-term rainfall prediction is one of the most popular task for nowadays, we are predicting the climate of the day, and also we are predicting the climate dynamics around the world. In this paper they are applying the neural network for short-term rainfall prediction. And also they implement two methods clustering and classifications are used.

Pundru Chandra Shaker Reddy, Alladi Sureshbabu (2019) [4] In this paper they are executing the experiment they are using a Hadoop Map reduce it makes quick and very fast. Nowadays the whether the data estimate increase day by day like terabytes or petabytes. Rainfall anticipating holds the same structure. The proposed strategy (EMLRM) is better performance for existing approaches. All things considered, the proposed strategy accomplished lower RMSE & MAE values as 0.274 & 0.0745 respectively. More comparison statistical of the proposed of the result the method shows that it is the optimal approach for predicting the rainfall. The cutting-edge model with contarsed has been designed for proposed system. They recommendation that our proposed design regarding models the MSE, RMSE and in terms of different rainfall forecasting techniques, our project outcomes creating a new model it will be executed the outstanding for monthly rainfall prediction based on different areas. As future work, it is more accuracy for rainfall prediction using ANN for upcoming month prediction.

S. Poornima and M. Pushpalatha (2019) [5] In this paper they are comparing a methods with Holt-winter, RNN and LSTM, used for improving a accuracy of the rainfall prediction in

the proposed system. LSTM network could holding a large of amount of dataset in a own memory and missing gradient could be avoided, it shows that the better accuracy for rainfall prediction data and it can provide the better decision making applications. In this paper they are using a deep learning approach for rainfall prediction.

M. N. M. Reba, N. Roslan, A. Syafiuddin, M. Hashim [6] (2016) Estimating rainfall accurate from reflective radar, hydrological model and precipitation, quantitative estimation is crucial. Intensities for different rainfall for various rainfall size drop contribution distribute to reflected and relationship of rain rate (Z-R) radar rainfall estimated in variability. This concept focus on Z-R model evaluated. Accordingly large flood at December 2014 for various rainfall. Reflective hourly and gauge rain is used. Least square linear and non-linear regression, were apply to distinguish the accuracy of coefficients Z-R. Non linear least square study found that best solution of a and b estimation in intensity of high rainfall.

A. A. Alonge; T. J. Afullo [7] (2011) Durban Rainfall Rate is (29°97'S, 30°95'E) four different classification for South Africa: thunderstorm, drizzler, shower, widespread. In 2009 obtained data in twelve month therefore JOSS-Waldvogel distrometer. It gives a full of 729 event rain independent 5,542,147 rain dropping. The rain rate measures the functions probabilities comparing with model of Moupfouma using integration time is 1-minute. The likelihood maximum estimation in parameters for the adopted of statistical model of two others.

Lei Zhou; Yuanhang Chen; Nian Liang; Yong Ni (2016) [8] Rainfall estimate accuracy based on rainfall based-ground and rainfall based-satellite, essential measurements for environment model and hydrological data in the china of Qaidam Basin. The Accuracy will be evaluated by using TRMM. A Gwr can be used as a distribution of the spatial of product of TRMM using a longitude and latitude as variable independent. It developed rainfall model by compare based-ground and based-satellite measurement rainfall, precision model was validating by cross-validation method. In TRMM contains many errors with measured ground precipitate. The error for daily datas. The based- satellite and based-ground rainfall geographical weighted regression model improving.

A. R. Holt, G. G. Kuznetsov, A. R. Rahimi (2003) [9] In this paper shows that the certain combination for polarization and frequency. Microwave attenuation is difference in middle of frequency two terrestrial link along with giving a better performance of rain rate path-averaged can be single frequency it obtained. Different attenuation can provide successful rainfall application.

Haonan Chen; V. Chandrasekar (2015) [10] Radar Dual-polarization has a very important for precipitation for quantitative. it is very large area to cover and sample the void. This paper mainly focused on the ku-band, dual-polarization Radar for rainfall light estimator. Algorithm they are using ku-band radar rainfall the structure is error they described.

V. PROPOSED SYSTEM

The proposed system it deals with the user provide the database by using Hadoop interface we will be able to analyzing no limitation of data, we can easily recover the data if the data is loss and we can easily adding the number of machines to the clusters and we will get accurate results with short time. The cost of the maintenance is very less and high capacity, the data processing is very efficient and we can also using division and bucket technique in Hadoop.

Advantages of Proposed System

- We cannot have a data losing problem in Hadoop.
- We cannot have limitations in dataset.
- Data processing is efficiently.

The Fig. 1 shows the complete system architecture the short-term rainfall prediction data is preprocessed in order to make it suitable for the Hadoop pre-process. Once the pre-processing gets over, logistic regression algorithm is applied to generate interesting pattern. The algorithm concepts consist of probability. The limit of the logistic regression algorithm is tends to value between 0 and 1. And the model is applied on the test dataset. Accuracy is calculated.

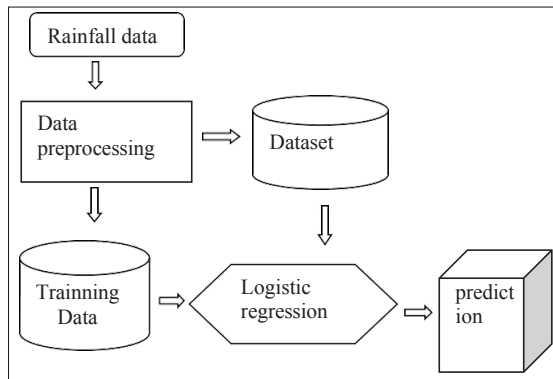


Fig. 1: Architecture Diagram

VI. REQUIREMENTS SPECIFICATION

A. Hardware Requirements

In this hardware requirement it serves the contract for a basis implementation for the system, and must be a consistent specification and complete system. They will be used system design for the starting point. It shows the how the system can be implemented.

PROCESSOR: PENTIUM IV 2.6 GHz, Intel Core RAM: 4GB

MONITOR: 15" COLOR

HARD DISK: 40GB

B. Software Requirements

In this software requirement it will specifies the system. It must be include two methods.

Specification and definition of requirement software. It provides a base for requirement software specification for creating. It can be used for cost estimating, task performing, team planning activity and team tracking etc.

FRONTEND: MYSQL

BACKEND: HADOOP

OPERATING SYSTEM: WINDOWS 10

DATA ACCESS TOOL: SQOOP

VII. MODULES

A. Pre-Processing Database

In this module, we are analyzing the various kinds of fields in Excel mysql data to csv format these data can be pre processing with the help of bigdata tool and finally got a proper rainfall dataset.

B. Pre-Processing Data

In this module we are getting historical data we will convert these historical batch processing the data from XLSC format to comma separator value format and also taking backup by using Mysql Database to avoid loss of data.

C. Data Storage

In this module we are getting a backup data from Mysql and we can also store the data and importing the data we are using the Sqoop commands through hdfs. The data are stored in HDFS it is ready to get processed.

D. Data Analysis

In this module, after analyzing the data by using Hadoop tool and convert into csv format, to import the csv file in python and processing the data.

E. Algorithm Implementaion

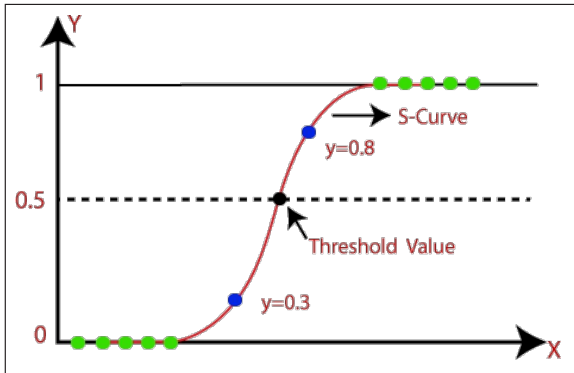
Step 1: First the user importing the Data Pre- processing.

Step 2: Next Fitting the training set to the logistic regression.

Step 3: Test the result to predict.

Step 4: Testing the correct result to the Accuracy (Confusion matrix creation).

Step 5: Last set result to visualizing the testing.



In this module, to import the data by using logistic regression algorithm and to predict the accuracy one of the most popular algorithms is logistic regression algorithm. Logistic regression algorithm provides the probability values which lie only between 0 and 1.

Linear regression is much similar to logistic regression.

Fitting a regression line instead of logistic regression, we set an “S” shaped logistic regression function maximum two values which we predict (0 or 1).

Logistic Functions (Sigmoid Functions)

The logistic function is represented as a sigmoid function; it is a mathematical function it will be used for mapping the predicted values for probability. Logistic function it mapping a any real value into another real value between the range of 0 and 1. This cannot go for off this limitation, so the curve like “S” form. Sigmoid function is also called as s-curve or logistic function. By using the threshold value tends to 1, value below the threshold value tends to 0.

VIII. RESULTS

The result that we are getting from our implemented algorithm as reflected in figure our implemented algorithm provides better results than previous algorithm term of accuracy. Also, the implemented algorithm provides better results when compared with other existing algorithm.

	STATE	UT	DISTRICT	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec	flood	Aug-Jun30Days
0	ANDAMAN	NICOBAR		107.3	57.9	65.2	117	358.5	295.5	285	271.9	354.8	326	315.2	250.9	2805.2	165.2	540.7	1207.2	892.1	0	98.5
1	ANDAMAN	SOUTH AN		43.7	26	18.6	90.5	374.4	457.2	421.3	423.1	455.6	301.2	275.8	128.3	3015.7	69.7	483.5	1757.2	705.3	0	152.4
2	ANDAMAN	N & M ANI		32.7	15.9	8.6	53.4	343.6	503.3	465.4	460.9	454.8	276.1	198.6	100	2913.3	48.6	405.6	1884.4	574.7	0	167.7667
3	ARUNACHAL	LOHIT		42.2	80.8	176.4	358.5	308.4	447	660.1	427.8	313.6	167.1	34.1	29.8	3043.8	123	841.3	1848.5	231	0	149
4	ARUNACHAL	EAST SIANG		33.3	79.5	105.9	216.5	323	738.3	990.9	711.2	508	208.9	29.5	31.7	4034.7	112.8	645.4	3008.4	268.1	1	246.1
5	ARUNACHAL	SIKCHANG		28	48.3	85.3	101.5	145.5	228.4	217.4	182.8	159.8	75.9	20.9	11.6	1200.4	76.3	327.3	788.4	108.4	0	76.3333
6	ARUNACHAL	TIAP		42.2	72.7	141	316.9	328.7	616.7	851.9	500.6	418.3	218.7	42.9	22.9	3571.5	114.9	785.6	3385.5	284.5	0	204.9
7	ARUNACHAL	ANJAM		42.2	80.8	176.4	358.5	308.4	447	660.1	427.8	313.6	167.1	34.1	29.8	3043.8	123	841.3	1848.5	231	0	149
8	ARUNACHAL	LOWER DI		83.7	153.9	303.5	383.6	288	374.2	272	160.5	266.7	167.2	64	56	2553.3	237.6	955.1	1073.4	287.2	0	124.7333
9	ARUNACHAL	CHANGLA		70.3	170.9	367.9	554.4	334.2	526.2	460.8	291.5	353.6	275	64.9	74.2	3543.9	241.2	1256.5	1632.1	414.1	0	175.4
10	ARUNACHAL	PAFUM PA		33.5	67.8	106.1	226.9	493	640.5	609.5	503.4	492.3	214.7	19.2	11.3	3378.2	101.3	786	2245.7	245.2	0	213.5
11	ARUNACHAL	LOW SUBI		97.5	109.3	92.4	204.3	264.2	284.1	248.9	270.5	192.7	78.5	49.5	27.2	1921.1	206.8	562.9	996.2	155.2	0	94.7
12	ARUNACHAL	UPPER SIA		74.3	176.7	362.6	397.5	408.7	802.9	631	417.9	686	244.9	86.9	71.7	4402.1	251	1168.8	2558.8	423.5	1	267.3
13	ARUNACHAL	WEST SIA		26	46.7	76.8	229.2	239.5	438.6	592.4	312.4	291.1	126.8	33.7	29.5	2440.7	92.7	545.5	1623.5	190	0	138.8667
14	ARUNACHAL	DIABANG VI		83.7	153.9	303.5	383.6	288	374.2	272	160.5	266.7	167.2	64	56	2553.3	237.6	955.1	1073.4	287.2	0	124.7333
15	ARUNACHAL	WEST KAN		35.2	43.5	58.9	134.3	341.1	665.3	740.9	576.1	490.9	233.9	40.3	27	3399.4	78.7	534.3	2485.2	301.2	1	221.7667
16	ARUNACHAL	EAST KAN		49	74.4	96.5	156.9	208	345.7	368.5	256.2	275.9	138.2	34.4	27.2	2030.9	123.4	461.4	1246.3	199.8	0	115.3333
17	ARUNACHAL	TARANG		35.2	43.5	58.9	134.3	341.1	665.3	740.9	576.1	490.9	233.9	40.3	27	3399.4	78.7	534.3	2485.2	301.2	1	221.7667
18	ARUNACHAL	KURUNG H		82.7	70	128.2	245.7	271.4	292.7	404	276.3	283.5	92.3	32.3	42.4	2221.5	152.7	645.3	1256.5	167	0	97.56667
19	ASSAM	CACHAR		13.3	50.2	168.3	262.5	386.4	512.1	526.2	470.8	360.8	182.4	34.8	11.4	2999.2	63.5	872.7	1989.9	238.6	0	177.3667
20	ASSAM	DARRANG		13.1	21.4	53.5	168.8	330	439.7	345.8	272.1	221.5	95.4	17.2	9.3	1857.8	34.5	542.3	1259.1	121.9	0	138.9
21	ASSAM	GOALPARU		12.7	20.4	51.1	196.6	398.8	567.8	502.8	334.6	304.9	157.7	21.7	5.2	2575.3	33.1	647.5	1710.1	184.6	0	108.2667
22	ASSAM	KAMRUPI		12	20.8	58.6	157.7	293.4	365.5	345.1	248.7	188.4	106.6	15.1	7.5	1813.4	32.8	503.7	1147.7	129.2	0	121.8333
23	ASSAM	LAKHIMPUR		27.7	48.6	76.7	165.5	331.9	528.3	605.2	467.6	424.1	140.3	23	20.4	2859.3	76.3	574.1	2025.2	183.7	0	176.1
24	ASSAM	NORTH CA		16.7	47.5	158.9	207.9	308	528.1	270.3	201.3	389.1	196.4	42.1	11.2	1977.5	64.2	674.8	988.8	249.7	0	108.3667
25	ASSAM	NAGAN		12	22.5	48.1	138.9	171.3	265.9	326.3	294.1	218.6	129	21.6	10.8	1460.1	34.5	348.1	1124.9	152.4	0	95.1
26	ASSAM	SHIVSAGAI		20.1	32.2	78.3	126.5	251.1	255.5	374.8	342.3	196.9	96.3	20.3	10.3	1811.6	53.3	481.9	1189.5	126.9	0	85.56667
27	ASSAM	BARPETA		10.3	26.9	54	175.7	395.5	694.3	757.3	527.3	462.1	142.1	20.4	12.7	3274.6	37.2	621.2	2441	175.2	1	233.4333
28	ASSAM	DHUBRI		10.3	11.7	46.6	147.5	391.6	683	554.7	418.7	340.1	155.1	19.2	4.1	2702.6	22	585.7	1595.5	178.4	0	201

Fig. 2: Dataset

Jupyter

M4 Last checkpoint 6/6/2021 (autosaved)

Logout

FileEditViewInsertCellKernelWidgetsHelp

TrusteePython 3.0

Out[12]:

	STATE_UT_NAME	DISTRICT	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	...	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
0	ANDAMAN And NICOBAR ISLANDS	NICOBAR	107.3	57.9	65.2	117.0	358.5	295.5	285.0	271.9	...	326.0	315.2	250.9	2805.2	165.2	540.7	1207.2	892.1
1	ANDAMAN And NICOBAR ISLANDS	SOUTH ANDAMAN	43.7	26.0	18.6	90.5	374.4	457.2	421.3	423.1	...	301.2	275.8	128.3	3015.7	69.7	483.5	1757.2	705.3
2	ANDAMAN And NICOBAR ISLANDS	N & M ANDAMAN	32.7	15.9	8.6	53.4	343.6	503.3	465.4	460.9	...	276.1	198.6	100.0	2913.3	48.6	405.6	1884.4	574.7
3	ARUNACHAL PRADESH	LOHIT	42.2	80.8	176.4	358.5	308.4	447.0	660.1	427.8	...	167.1	34.1	29.8	3043.8	123.0	841.3	1848.5	231.0
4	ARUNACHAL PRADESH	EAST SIANG	33.3	79.5	105.9	216.5	323.0	738.3	990.9	711.2	...	206.9	29.5	31.7	4034.7	112.8	645.4	3008.4	268.1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
636	KERALA	IDUKKI	13.4	22.1	43.6	150.4	232.6	651.6	788.9	527.3	...	343.2	172.9	48.1	3302.5	35.5	426.6	2276.2	564.2
637	KERALA	KASARODU	2.3	1.0	8.4	46.9	217.9	999.9	1108.5	636.3	...	234.9	84.6	18.4	3621.6	3.3	272.9	3007.5	337.9
638	KERALA	PATHANAMTHITTA	19.8	45.2	73.9	184.9	294.7	556.9	539.9	352.7	...	359.4	213.5	51.3	2958.4	65.0	553.5	1715.7	624.2
639	KERALA	VARANAD	4.8	8.3	17.5	83.3	174.9	688.1	1110.4	592.9	...	213.1	93.6	25.8	3253.1	13.1	275.4	2632.1	332.5
640	LAKSHADWEEP	LAKSHADWEEP	20.8	14.7	11.8	48.9	171.7	330.2	287.7	217.5	...	157.1	117.7	58.8	1600.0	35.5	232.4	998.5	333.6

641 rows x 21 columns

Fig. 3: Data Pre-Processing

