International Journal of Engineering & Technology, 7 (4) (2018) 2752-2759

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.18370
Research paper

Map reduce technique for parallel-automata analysis of large scale rainfall data

Tulasi Sunitha Manepalli [1] *, Dr. Chamakuzhi Subramanian [2]

[1] *Research Scholar, Jain University, Department of Computer Science & Engineering, India*
[2] *Jain University, Department of Computer Science & Engineering, India*
**Corresponding author E-mail: Tulasi80@gmail.com*

Abstract

Vast scale rainfall information assumes an imperative part in farming field thus early expectation of rainfall is important for the better finan-cial development of a nation. Rainfall expectation is an expert among the most troublesome issue far and wide in a year back. This data is generally secured in the unstructured course of action. Along these, tremendous measure of data has been accumulated and ar-chived. Thus, storage and handling of such tremendous information for accurate rainfall forecast are a major test. Big Data innovation like Hadoop have developed to fathom the difficulties and issues of huge information utilizing distributed computing. Till date few ex-aminations have been accounted for on the preparing of vast scale rainfall information utilizing MapReduce. In this paper, the huge scale rainfall information is anticipated by utilizing MapReduce system which plays out the capacities which are required and diminishes the task to get proficient ar-rangements through taking the information and isolating into smaller tasks. At that point, the three Regression Automata (RA) algorithms such as Linear Regression automata, Support Vector Regression Automata and Logistic Regression Automata. are utilized to forecast the future esteem of large scale rainfall data. The proposed framework serves as a tool that takes in the rainfall information from diminished information as input and predicts the future rainfall. The outcomes obviously demonstrate that the all the three RA models can anticipate the rainfall productively in different terms, such as, error rate, coefficients and mean square error.

Keywords: *Big Data; Hadoop; Map Reduce; Rainfall Forecast; Regression Automata.*

1. Introduction

Agriculture is the foundation of the world economy. The irrigation system is as yet not all that great in overall world and the vast majority of agriculture relies on the rain [1]. A decent rainfall result in the event of a dry period for quite a while or overwhelm-ing rainfall both influences the product yield and in addition the economy of the world, so because of that early expectation of rainfall is extremely pivotal. An extensive variety of rainfall con-jecture techniques are utilized in climate expectation at territorial and national levels [2-3]. Environmental change has been looking for a great deal of consideration for quite a while because of the unexpected changes that happen. There are a few impediments in better usage of climate estimating accordingly it winds up hard to forecast climate with short term productivity [4]. The forecast of atmosphere has constantly turned out to be critical and helpful. Big data gather vast volume of information and it is an incredible test for Hadoop, a piece of Big Data, which utilizes Map Reduce to keep up and process the information and concentrates helpful data in a proficient way [5]. The Big Data keeps up the immense measure of information and procedures them effectively. Enor-mous information incorporates data indexes with sizes beyond the capacity of generally utilized programming instruments to catch, oversee and process the information. We will utilize Map Reduce with a specific end goal to examine the informational collections and to perform different tasks on the information collection [6]. In the meantime, the traditional way to deal with process the infor-mation is moderate. Process the sensor information with MapRe-duce in Hadoop system, which evacuates the adaptability bottle-neck. Hadoop is an open system utilized for Big Data examination [7]. Hadoop's main processing engine is MapReduce, which is presently a standout amongst the most mainstream Big Data han-dling systems accessible.

MapReduce is a technique for executing exceedingly parallelliza-ble and distributable algorithms across huge datasets utilizing countless PCs. Utilizing MapReduce with Hadoop, the large-scale rainfall can break down without adaptability issues. The speed of handling information can increment quickly when across over multi-group disseminated system [8]. Subsequently, there is a requirement for an adaptable stage for the maintenance of this Big Data and help climate gauging utilizing that Big Data. In this way, Apache open source Hadoop is an answer for it, that gives rapid clustering processing to the examination of an expansive arrange-ment of information easily and effectively [9]. The climate can be anticipated by utilizing Time-Series Data Mining (TSDM), the time-series estimating issues have pulled in wide thoughtfulness regarding tackling the issues by giving an approach to investigate past behavior. As of late, different advancements have been pro-posed for time-series anticipating [10]. TS is a grouping of infor-mation, which is related to time, for example, day-by-day tem-perature estimation [11]. The point of the Time Series Analysis (TSA) is to plan TS information in order to learn, fit low dimen-sional models, and make gauges [12]. An analyst in assembling attempts to review any strategies that have been proposed, particu-larly to model and handling TS of worldly information [13]. The point of the proposed work is to outline a powerful portrayal sys-tem for time-series with dimensionality reduction utilizing

Copyright © 2018 Tulasi Sunitha Manepalli, Dr. Chamakuzhi Subramanian. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

MapReduce and furthermore plan proficient Automata based framework for Modeling Time series. The strategy for performing time-series mining task on climate information in view of the proposed approach and contrasting it with existing strategies to anticipate the rainfall. Python is utilized to predict the rainfall from the acquired information.

The remaining paper is discussed as: Section 2 gives the description of the models analyzed by various researchers that relate to this study. Section 3 provides a description of proposed methodology used for predicting the rainfall. Sections 4 present the results obtained by various experiments and the conclusions are made in Section 5, respectively.

2. Literature review

An earlier research on several MapReduce techniques for predicting the rainfall is described below. In this scenario, brief evaluations of some important contributions and limitations are presented.

M. Joshi, et al., [14] an endeavor was to perceive how big data arrangement was used in the field of climate expectation. Artificial Neural Network (ANN) on the Map-Reduce system was executed for rainfall expectation and yield recommendation. Here ANN was utilized to predict the next seven day's rainfall. In the method, the utilization of back proliferation strategy was executed in which there was an error remedy system that prompts high precision in the anticipated outcomes. Also, technique recognized soil and Regional investigation which were distinguished harvest contingent upon the user's area and climate condition to expand the yield of products. The usage of this arrangement on Hadoop made the strategy quicker and versatile. The ANN technique anticipated only short term rainfall for crop recommendation, and this won't much valuable for better harvest yields.

K. A. Ismail, et al., [15] in this work, MapReduce with Hadoop to break down the sensor information alias the Big Data was a successful arrangement. MapReduce is a technique for executing exceedingly parallelizable and distributable algorithms over tremendous datasets utilizing countless PCs. Utilizing MapReduce with Hadoop, the temperature was examined effectively. The adaptability bottleneck was evacuated by utilizing Hadoop with MapReduce. The expansion of more frameworks to the disseminated organize gave quicker preparing of the information. With the board work of these innovations all through the business and the interests of the open-source networks, the capacity of MapReduce and Hadoop kept on developing. The utilization of these sorts of innovations for the expansive scale of information examinations could enormously improve the climate conjecture as well. The technique didn't focus much on rainfall forecast yet the entire climate expectations were dissected.

K. Namitha, et al., [16] implemented an approach of preparing such Big volume of climate Data utilizing Hadoop. The proposed strategy included Artificial Neural Network executed on Mapreduce system for short and long term rainfall expectation. Rainfall was anticipated one day ahead by utilizing temperature and rainfall information of promptly going before days. Executing this arrangement on Hadoop made the strategy speedier and versatile. Regardless of whether the information estimate develops to terabytes or petabytes, a similar system holds useful for rainfall gauging. Temperature and Rainfall information of India over recent years (1951-2013) was utilized for this investigation. In some case, the strategy furnished poor accuracy when contrasted with investigation of other map-reducible machine learning algorithm.

C. P. Shabariram, et al., [17] exhibited a novel answer to deal with the information in view of spatial temporal qualities utilizing a Map Reduce Framework. The workload was grouped by utilizing Support Vector Machine (SVM). The technique utilized feature determination and reduction algorithm related to the dataset. The various rainstorm idea expectation was accomplished by utilizing the huge raw rainfall information. The dataset affect parameters were arranged into local, hourly, and generally tempests. The proposed framework serves as a device for forecasting rainstorm from a lot of rainfall information in an effective way. The outcome demonstrated the proposed framework enhanced the execution as far as efficiency and accuracy. The strategy could forecast the rainfall for just a single day alone.

A. Nair, et al., [18] a non-linear strategy known as ANN has been utilized on the yields of Global Climate Models (GCMs) to draw out the vagaries innate in month-to-month rainfall forecast. The ANN strategy was applied on various outfit individuals from the individual GCMs to acquire month-to-month scale forecast over India and over its spatial matrix focuses. In the present investigation, a double-cross-validation and straightforward randomization system was utilized to keep away from the over-fitting during the training procedure of the ANN framework. The execution of the ANN-anticipated rainfall from GCMs was judged by breaking down the box plots, percentile, absolute error, and contrast with linear error in likelihood space. The execution examination uncovered that the ANN demonstrate could catch the year to year varieties in rainstorm months with genuinely great accuracy in extraordinary years too. Since, the execution of the ANN relies upon the training set, the capacity of ANN system was to reproduce the vagaries of rainfall were not be enhanced if smaller datasets were acquired from GCM.

To defeat the issues, for example, speed and precision misfortune because of successive processing, the proposed strategy presented the MapRedcue technique which reduces the information dimensionality in the database. Once the MapReduce for shaping day to month and yearly from single or numerous climate stations is done, at that point the information is sent to forecast the rainfall utilizing three regression models Viz., linear regression, support vector and logistic regression model.

3. Proposed methodology

A forecast of vast scale information on rainfall is an enormous test to the climatologists. The majority of the consuming issues of our opportunity like a worldwide temperature alteration, floods, dry spell, warm waves, soil disintegration and numerous other climatic issues are specifically identified with rainfall. Agriculture is the significant wellspring of monetary exercises in the greater part of the nations of the world. Accordingly, forecasting the substantial scale rainfall effectively is vital, these days, the greater part of the linear methods and the discoveries were uncertain. In a general sense there are two ways to deal with anticipate Rainfall. They are Empirical and Dynamical Methods. Better parallelized-preprocessing strategies alongside forecast of precipitation are required for taking care of large-scale rainfall information. The Empirical approach depends on investigation of previous authentic information on climate and its relationship to an assortment of environmental factors over various parts of the world. The most broadly utilize empirical methodologies utilized for atmosphere expectation are Regression, ANN, fuzzy logic and gathering strategy for information handling. The dynamical approach, expectations are created by physical models in view of arrangement of conditions that forecast the future Rainfall. To forecast the climate by numeric means, meteorologist has created atmospheric models that set the adjustment in temperature, pressure and so forth utilizing scientific conditions. Figure 1 shows the proposed methodology of our work.

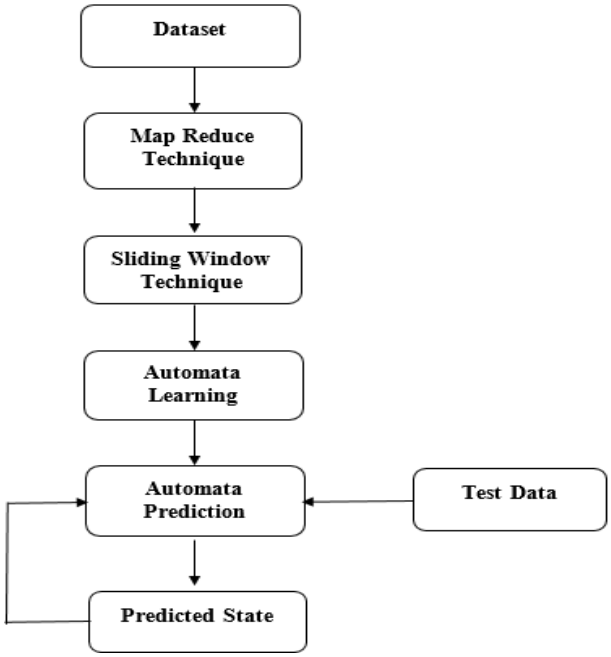

Fig. 1: Architecture of the Proposed Methodology.

In this paper, the approach utilizes the three distinctive RA procedures to anticipate precipitation. The expectation of result (e.g. absence or presence of rainfall) is acquired by utilizing the RA models depends on estimations of an accumulation of indicator factors.

3.1. Dimensionality reduction technique

Hadoop is utilized for preparing the tremendous climate information in a distributed approach. This system is to work for handling the numerous measures of information, in order to settle on its systematic decision prompting better analytical choice. Every one of the modules in Hadoop are outlined with a crucial assumption that equipment failures are regular events and ought to be naturally deal with the system. Hadoop comprises of a storage part, known as Hadoop Distributed File framework (HDFS) and a handling part called Map-Reduce. This dimensional reduction technique by using Map-Reduce will speed up the pre-processing function in parallel and helps the model to produce the output in short span.

3.1.1. Hadoop

Hadoop is a stage that offers a productive and compelling technique for storing and preparing huge measures of information. Unlike conventional contributions, Hadoop was planned and developed from the bottom to address the prerequisites and difficulties of enormous information. Hadoop is great in its capacity to enable organizations to quit worrying about building big data framework and to center around what truly matters: extricating business esteem from the information [19]. Apache Hadoop utilizes cases are many, and appear in numerous ventures, including: hazard, misrepresentation and portfolio investigation in finance related administrations; conduct examination and personalization in retail; informal organization, relationship and notion investigation for showcasing; drug communication demonstrating and genome information handling in social insurance and life sciences and so on. Furthermore, numerous organizations give Hadoop business execution and additionally support, including Cloudera, IBM, MapR, EMC, and Oracle. As indicated by the Gartner Research, Big Data Analytics is a slanting point in 2014. Hadoop is an open structure generally utilized for Big Data Analytics.
Two primary parts of Hadoop will be Hadoop Distributed File System (HDFS) and MapReduce [20]. HDFS is a conveyed document framework administration for huge datasets of sizes of gigabytes and petabytes. Also, MapReduce is a programming structure for overseeing and handling an immense measure of unstructured information in parallel based on the division of a major dataset into smaller autonomous chunks.

3.1.2. HDFS system

HDFS is Hadoop's execution of a dispersed file system. It is intended to hold a lot of information and give access to this information to numerous customers appropriated over a system [21]. HDFS has a master/slave design. The primary segments of a HDFS group are a single NameNode, a master server that deals with the document framework and control access to records by customers. Moreover, there are various DataNodes, every hub often contains one DataNode in the group, which handles storage capacity related to these hubs.

3.1.3. Map reduce technique

MapReduce is a great model for circulated registering, presented by Google in 2004. Each MapReduce work is made out of a specific number of map and reduce tasks. The MapReduce display for serving various occupations comprises of a processor sharing line for the Map Tasks and a multi-server line for the Reduce Tasks [22]. To run a MapReduce work, clients ought to outfit a map work, a reduced work, input information, and an output information area as appeared in figure 2. Whenever executed, Hadoop completes the accompanying advances: Hadoop breaks the information into various information things by new lines and runs the map work once for every data item, giving the thing as the contribution for the capacity. Whenever executed, the map work yields at least one key-value pairs. Hadoop gathers all the key-values sets created from the guide work, sorts them by the key, and gatherings together the qualities with a similar key [23]. For each particular key, Hadoop runs the reduce work once while passing the key and list of values for that key as input. In the figure, the key represents the date and the value indicates rainfall data.

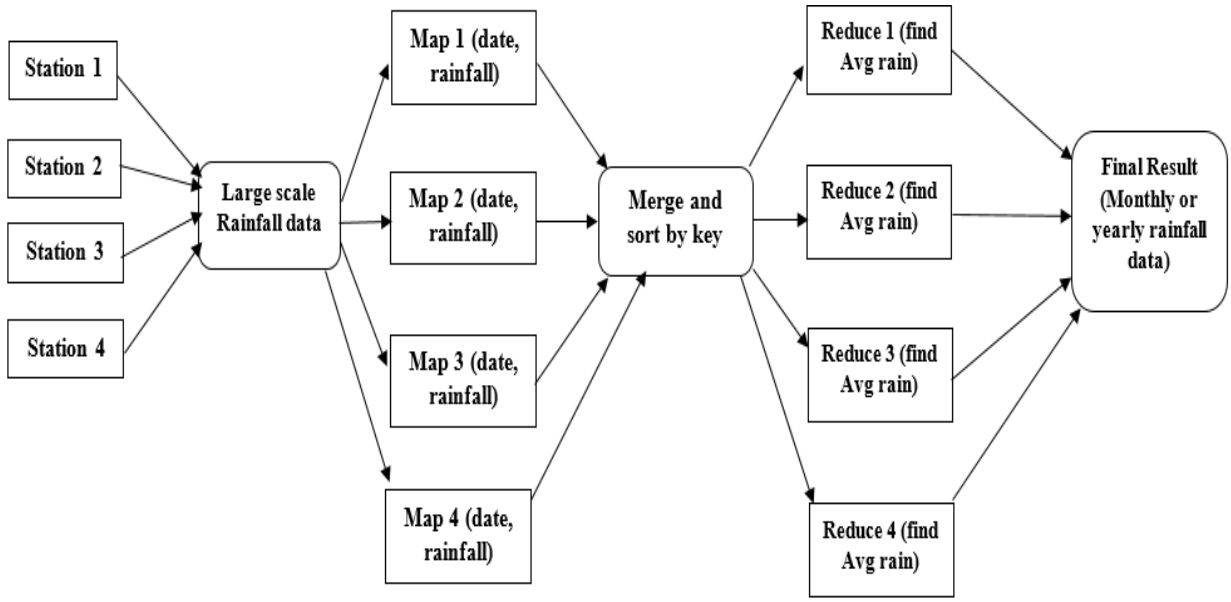

Fig. 2: Map Reduce Technique Flowchart

The reduce function may yield at least one key-value sets, and Hadoop writes them to a record as the last result. Hadoop enables the client to design the activity, submit it, control its execution, and inquiry the state. Each activity comprises of independent tasks, and every one of the tasks needs a framework opening to run. With Hadoop all task and planning decisions are made on a node space and task level for both the map and reduce stages [24].

3.2. Sliding window techniques

One of the key problems in the learning task is to determine the length of the sliding window, i.e., how many historical data points the prediction would rely on. E_{in} and E_{out} are the fitting mean square error in the training data and testing data respectively to process the data in parallel over different states. One can infer that by increasing the parallel processing complexity (sliding window length), E_{in} decreases sharply, while E_{out} becomes increasingly worse, which is typically the result of overfitting. In practice, it favors simpler models in order to minimize the risk of overfitting. The models, of which window length is less than 5, have relatively small E_{out}.

3.3. Automata learning

The subjects of CA deal with large collections (usually infinite in order to avoid boundary problems) of interconnected finite automata, each finite automaton being thought of as a cell. The approach uses the Wolfram's classification of CA for predicting the rainfall in three stations of Queensland (QLD) state. The Wolfram's algorithm [25] is anticipating the detailed properties of a specific CA, usually enough just to realize what class the cellular automata was in. The second issue is that Turing all-inclusive computation and the conceivable connection between the complexity qualities of CA are handled by the Wolfram's algorithm.

The analysis of Wolfram's includes a one-dimensional (1D) study, order $(k=2; r=2)$, where $r \in Z$ the number of neighbors and $k \in Z$ is the cardinality of the finite alphabet and find the behavior of the same classes in other CA rule spaces.

In a 1D array, a finite automaton called an Elementary Cellular Automaton (ECA) is well defined. The automaton updates two states and also the closest neighbors' state in discrete time depends on its own state and synchronously, all cells updating their states.

Wolfram's classes can be described as:

i) Class I. CA evolves chaotically.
ii) Class II. Includes all previous cases, known as a class of complex rules.
iii) Class III. CA evolves periodically
iv) Class IV. CA evolves to a homogeneous state.

Otherwise explained, in the case of a given CA,

The development is ruled by sets of cells with no characterized design for any arbitrary and long-term introductory condition, at that point it has a place with Class I.

The non-trivial structures commanded the advancement of rising and going along the development space. These spaces are intermittent, confused or uniform can coincide, at that point it has a place with Class II. This class is as often as possible labeled, for example, essentially mind boggling, complex conduct, many-sided quality or flow.

The squares of cells ruled the occasionally rehashed development for any arbitrary beginning condition, at that point it has a place with Class III.

Class IV contains the interesting condition of its letters dominated the development for any arbitrary introductory condition.

3.4. Regression automata model

The estimation of relationships among variables is evaluated by the RA which is a set of statistical processes. The focus of the RA is on the relationship between an independent variable or dependent variable and the analysis includes various techniques for evaluating and modeling several variables. The use of RA has a substantial overlap with the field of machine learning that is used for forecasting and prediction. RA explores the forms of relationship between independent and dependent variables. In this work, the regression analysis model is used to predict the daily rainfall prediction for the four stations. There are three regression models are used to predict rainfall such as a Linear Regression model (LR), Support Vector Regression model (SVR), and Logistic Regression model (LOR).

3.4.1. Linear regression model

LR is a method used for defining the relationship between one or more independent variables or explanatory variables, denoted by (X) and a dependent variable (Y). For multiple explanatory variables, the process is defined as Multiple Linear Regression (MLR). The general equation for an LR is given as in Eq. (1),

$$y_i = \beta_0 1 + \beta_i x_{i1} + ... + \beta_p x_{ip} + \mathrm{E}_i = x_i^T \beta = \mathrm{E}_i, \; i = 1, ..., n, \quad (1)$$

Copyright © 2018 Tulasi Sunitha Manepalli, Dr. Chamakuzhi Subramanian. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Where y denotes the dependent variable (rainfall) and x_i where $i=1,2,...,n$, denotes independent variables and β is called the intercept.

3.4.2. Support vector regression model

The SVR display is a learning approach, firstly utilized in the pattern acknowledgment issues. Later on, it was changed and utilized in the regression issues. The SVR is considered as a flourishing algorithm in the learning issues. In the regression problems, training procedure includes obtaining the correlation or non-linear mapping function $f(x)$ between both learners (i.e. input and output of the learner). The SVR [26] aims to provide a non-linear mapping function to map the training data $x_i, y_i; i=1,...,n$ to a high dimensional feature space. Then, the non-linear relationship between both learners can be described by a regression function as follows in the Eq. (2).

$$f(x)=W^T\varphi(x)+b \tag{2}$$

Where w and b are the coefficients to be adjusted. In fact, SVR is an optimizing problem in which objective function is given in Eq. (3).

$$Min_{w,b,\zeta^*,\zeta} R_c(W,\zeta^*,\zeta)=0.5w^Tw+c\sum_{i=1}^{n}(\zeta_i+\zeta_i^*) \tag{3}$$

Where C is the trade-off parameter between the first and second terms of the equation.
By solving the above-described optimization problem the coefficient of Eq. (2) can be found as follows in Eq. (4).

$$w=\sum_{i=n}^{n}(\beta_i-\beta_i^*)\varphi(x_i) \tag{4}$$

Where β_i is the Lagrangian coefficients. The following Eq. (5) describes the SVR regression function.

$$f(x)=\sum_{i=1}^{n}(\beta_i-\beta_i^*)k(x_i,x_j)+b \tag{5}$$

Where the Kernel function is denoted by $K(x_i,x_j)$. Among the group of Kernel works, the most normally utilized ones are the Gaussian Radial Basis Functions (RBF) and the polynomial. There are no particular rules for deciding the correct Kernel compose for particular information designs. The idea of SVR for non-direct is outwardly represented in Figure 3 [27].

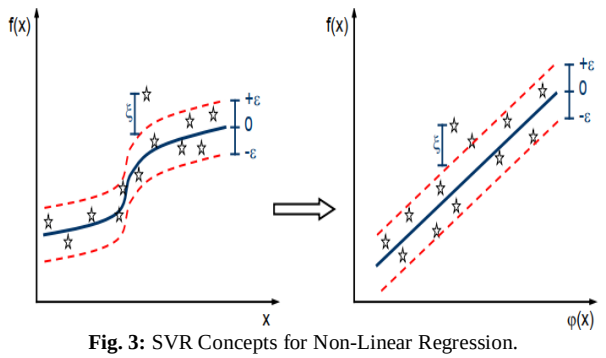

Fig. 3: SVR Concepts for Non-Linear Regression.

As mentioned, the SVR parameters affect the accuracy of the prediction. Hence, it is essential to select appropriate parameters. The parameter takes care of the trade-off between the degree of the training error and the model flatness; large values of the parameter results in only minimizing the empirical risk.

3.4.3. Logistic regression model

Logistic regression [28] enables one to forecast a discrete result, for example, regardless of whether it will rain today or not, from numerous sorts of factors that might be dichotomous, consistent, discrete, or a blend of any of these. Generally, the reaction or dependent variable is dichotomous, for example, achievement/disappointment or presence/absence, i.e., the needy variable can take the value 0 or 1 with a probability of disappointment or achievement. At that point, this sort of factor is known as a Binary (or Bernoulli) variable.
Consider a simple k variable regression model in Eq. (6).

$$E(Y \vee X)=\beta_0+\beta_1X_1+...+\beta_pX_p \tag{6}$$

Where $k=p+1$. We would logically let

$$y_i=0$$

If the ith unit does not have the characteristic.
One, if the ith unit does possess that characteristic.
Generally, where the response is twofold, the state of the response work is shown by extensive empirical confirmation that must be non-linear (in a variable). A monotonically expanding (or diminishing) S-shaped (or reversed S-formed) bend could be a superior decision. This kind of curve is obtained, if the regression chooses the specific form of the function as defined in Eq. (7).

$$\prod(X)=\frac{\exp(z)}{1+\exp(z)} \tag{7}$$

Where $Z=X\beta$. This function called as the logistic response function. Here Eq. (8) describes the linear predictor called Z.

$$Z=\ln\left(\frac{\Pi}{1-\Pi}\right) \tag{8}$$

In the Eq. (9), the term$Y$ model would be written as.

$$E(Y \vee X)=\Pi(X) \tag{9}$$

It is a notable issue that the binary response model abuses various Ordinary Least Squares (OLS) suspicions. Consequently, it is a typical practice to utilize the Maximum Likelihood (ML) technique depends on Iterative Re-Weighted Least Squares (IRLS) algorithm. The proposed regression model of LR, SVR and LOR outperforms the existing ARIMA model by its mathematical nature of predicting the future esteem robustly in shorter time span.

4. Experimental analysis

In this section, the proposed method presents an evaluation of the proposed rainfall prediction algorithm discussed. First, the experimental settings are described in this section, the series of experiments are conducted for evaluating the effectiveness of rainfall prediction algorithms and then the results are presented and discussed.

4.1. Dataset description

In worldwide, a nation's economy depends on agriculture and its items, and product yield is intensely reliant on the spring monsoon (June-September) rainfall. Hence, any reduction or increment in yearly precipitation will dependably severely affect the agriculture part in countries like India. Henceforth, the earlier information of rainstorm behavior will encourage the Government and agriculturists to take the benefit of the storm season. This learning can be extremely helpful in limiting the harvest's harm during the less rainfall in the storm season. Forecasting is a critical logical issue in the field of monsoon meteorology. In this study, day wise data

is collected from four rainfall stations in the Australian Government Bureau of Meteorology http://www.bom.gov.au/climate/data (Belmont Agforce, Glenlands, Broadmeadows, Gracemere-Lucas stations in QLD state). The changeover dates vary from State to State and year to year. More information can be available at http://www.bom.gov.au/climate/averages/tables/daysavtm.shtml.

4.2. Performance criteria

The outcome of the models developed in this study was assessed using standard statistical performance evaluation criteria which included the error ratio metric, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Nash-Sutcliffe Efficiency coefficient (NSE).

4.2.1. Root mean squared error

The predictive capabilities of the model can be provided by different types of information in RMSE. The RMSE measures the goodness-of-fit relevant to high rainfall values. RMSE is defined in Eq. (10),

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i^p - Q_i^0)^2} \tag{10}$$

Where, n is the number of input samples, Q_i^p and Q_i^0 are the observed and predicted rainfall at time t.

4.2.2. Mean absolute error

MAE given by equation 11 was used to measure the accuracy of forecasting. Small estimations of these parameters show higher model accuracy. The adjusted point of the goodness-of-fit can be yielded by MAE at direct value conveyance of the estimation errors. The MAE is predicted in Eq. (11)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Q_i^p - Q_i^0| \tag{11}$$

4.2.3. Nash-Sutcliffe efficiency coefficient

The NSE was used to evaluate the goodness of fit between the observed and the forecasted values. In addition, NSE provides higher values of this coefficient indicate better model outperformance. The following Eq. (12) represents the NSE coefficient,

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_i^0 - Q_i^p)^2}{\sum_{i=1}^{n}(Q_i^0 - Q^0)^2} \tag{12}$$

Where, $Q^{\prime 0}$ is the mean of the observed rainfall value?

4.2.4. Error ratio metric

By this way, the impact of predicted algorithm compared to the observed algorithm can be computed as follows in Eq. (13),

$$ErrorRatio = \frac{RMSD_{predicteda\,lgorithm}}{RMSD_{obxerveda\,lgorithm}} \tag{13}$$

4.3. Experimental analyses

In this section, the achievement of the regression models is evaluated by the parameters like RMSE, MAE, NSE and error ratio for the four datasets.

4.3.1. Evaluation of RMSE

The values of the RMSE are obtained from the experimental results for four databases namely ID033229, ID039043, ID039049 and ID039242. The table 1 describes the performance of the proposed method in terms of RMSE values. The LR model has the highest RMSE values (0.3325) for ID 039043 station in QLD state. Though all the three models performed well in database, the LOR model leads better performance when compared with the other two models in ID 039242.

Table 1: RMSE Evaluation for Four Databases

RMSE	ID033229	ID039043	ID039049	ID039242
LR	0.3023	0.3325	0.2841	0.2006
LOR	0.3013	0.3315	0.2831	0.1976
SVR	0.2874	0.3129	0.2722	0.2030

4.3.2. Evaluation of MAE

The rainfall is predicted by the values obtained for MAE parameters which are tabulated in Table 2. Even though the SVR model yields better performance in RMSE values when compared to LR model for all stations, the SVR model provides the poor MAE values in all the four stations. The LOR model achieved 0.039 MAE values in ID039242 station. In MAE, LOR method provides better performance in all the stations.

Table 2: MAE Evaluation for Four Databases

MAE	ID033229	ID039043	ID039049	ID039242
LR	0.0917	0.1107	0.0808	0.0403
LOR	0.0907	0.1099	0.0801	0.0390
SVR	0.1726	0.1879	0.1641	0.1312

4.3.3. Evaluation of NSE

The values of the NSE are obtained from the experimental results for four databases namely ID033229, ID039043, ID039049 and ID039242. The table 3 presents the values of NSE for all four stations. Here, compared to the other database, the LR method performed well in all the four databases. The LOR provides 2.0 as an optimal value in all the datasets, whereas next to LOR, SVR presents the values such as 1.9101, 2.0560 in ID033229 and ID039242.

Table 3: NSE Evaluation for Four Datasets

NSE	ID033229	ID039043	ID039049	ID039242
LR	2.0068	2.0060	2.0073	2.0308
LOR	2.0	2.0	2.0	2.0
SVR	1.9101	1.8909	1.9247	2.0560

4.3.4. Evaluation of error ratio

The error ratio of the proposed algorithm is calculated for predicting the performance of the three regression models. The values are given in table 4 shows the best regression model for predicting the daily rainfall for four stations in QLD state. The rainfall can be predicted by the several models, but the best appropriate models are selected by considering their error ratio. Overall, the LOR method performed well to predict the rainfall correctly.

Table 4: Error Ratio Evaluation for Four Datasets

Error Ratio	ID033229	ID039043	ID039049	ID039242
LR	-0.0917	-0.1107	-0.0808	-0.0403
LOR	-0.0907	-0.1099	-0.0801	-0.0390
SVR	0.0092	-0.0099	0.0198	0.0609

4.4. Comparative analysis

Forecast of monthly and yearly monsoon rainfall for the year 1954-2015 was done using the developed models. The gauge estimations of rainfall acquired in this examination can be utilized in planning water resources and agricultural, hydrological method study and environmental change study. Table 5 shows the performance criteria of different forecasting methods for the month and year rainfall forecasting. Note that the results were obtained using different RA techniques coupled with LOR. The proposed methodologies are implemented for reducing the large scale rainfall data and also compared with the existing method ARIMA model that were implemented by using MapReduce Technique for dimensionality reduction. The existing method ARIMA with MapReduce were provided poor performance in predicting the month and year rainfall. Therefore, the Hadoop technique using MapReduce is combined with Automata model for predicting better monthly and yearly large scale rainfall data.

Table 5: Comparison of Proposed Method with Existing Method

Authors	Metrics Used	RMSE				MAE				NSE			
	Datasets	229	043	049	242	229	043	049	242	229	043	049	242
Geetha, A. and Nasira, G.M. [29]	MapRedcue + ARIMA	84.83	70.07	48.49	37.51	72.99	61.61	40.45	33.93	0.60	0.75	1.10	0.659
	LR	0.302	0.332	0.284	0.200	0.091	0.110	0.080	0.040	2.006	2.006	2.007	2.03
Proposed Methodoliges + MapReduce	LOR	0.301	0.331	0.283	0.197	0.090	0.109	0.080	0.390	2.0	2.0	2.0	2.0
	SVR	0.287	0.312	0.272	0.203	0.172	0.187	0.164	0.131	1.910	1.890	1.924	2.056

Table.6: Gives Sample of Execution Times of Monthly Rainfall Forecast Using Innovative Methodology with Map Reduce and Without Map Reduce

Research Methodology	Dataset	ID033229	ID039043	ID039049	ID039242
Methodology without MapReduce	LR	60.52	75.3	78.56	69.3
	LOR	59.67	73.2	74.43	62.14
	SVR	57.21	71.24	75.29	61.01
Methodology with Map Reduce	LR	41.51	54.5	59.85	48.76
	LOR	40.98	57.9	51.01	44.68
	SVR	33.5	50.87	51.79	36.6

In all the cases, the execution times of monthly rainfall forecast using methodology with Map Reduce are less than those obtained with methodology without Map Reduce.

As an overall result, it is inferred that the proposed regression automata models can perform accurate prediction with parallel processing than the combination of existing method with MapReduce for effective TSA on rainfall forecasting. The existing method is affected by error rate up to 85% in all other four stations because of using ARIMA model. But, the proposed method reduced the error rate up to 0.08% by using mathematical model with the combination of MapReduce technique.

5. Conclusion

Long historical monthly and yearly rainfall depth TS (Jan. 1954–Dec. 2015) at four selected stations (ID39043, ID39049, ID39242, ID33229) have been analyzed. The total monthly rainfall depths were calculated and the climatological normal of the whole rainfall was obtained at all four stations. Analyses of these TS clearly illustrate very strong temporal and spatial variability. The information preprocessing is conveyed in the map reduce process utilizing Hadoop structure as the dataset is accessible in substantial scale and thus to enhance the execution of the cluster adaptability. More through statistical comparison of the outcome of the three RA models indicates that the second model (Logistic RA model) is the optimal model for forecasting rainfall in terms of all NSE values. This complete system serves as an application that allows big raw rainfall data to be easily analyzed and classified to obtain the both (i.e. yearly and monthly) rainfall information from the available cluster. In the future, we will enhance the robust and fast response of parallel processing time series analysis mechanism using an effective hybrid mathematical model which considers all the other attributes influencing the rainfall.

References

[1] N. Sethi, and K. Garg, "Exploiting data mining technique for rainfall prediction", *International Journal of Computer Science and Information Technologies*, Vol.5, No.3, pp.3982-3984, 2014.

[2] P.S. Dutta, and H. Tahbilder, "Prediction of rainfall using data mining technique over Assam", *Indian Journal of Computer Science and Engineering (IJCSE)*, Vol.5, No.2, pp.85-90, 2014.

[3] M. Kannan, S. Prabhakaran, and P. Ramachandran, "Rainfall forecasting using data mining technique", *International Journal of Engineering and Technology*, Vol.2, No.6, pp.397-401, 2010.

[4] A. Gautam, and P. Bedi, "MR-VSM: Map Reduce based vector Space Model for user profiling-an empirical study on News data", In *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Kochi, pp. 355-360, 2015. https://doi.org/10.1109/ICACCI.2015.7275635.

[5] A. Gautam, R. Dhingra, and P. Bedi, "Use of NoSQL database for handling semi structured data: an empirical study of news RSS feeds", *Emerging Research in Computing, Information, Communication and Applications*, New Delhi, pp.253-263, 2015.

[6] S. Navadia, P. Yadav, J. Thomas, and S. Shaikh, "Weather prediction: A novel approach for measuring and analyzing weather data," In *International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp.414-417, 2017.

[7] D. V. Sahasrabuddhe, and P. Jamsandekar. "Data structure for representation of big data of weather forecasting: a review." *International Journal of Computer Science Trends and Technology (IJCST)* vol, 3, no. 6, pp. 48-56, 2015.

[8] V. Dagade, M. Lagali, S. Avadhani, P. Kalekar, "Big Data Weather Analytics Using Hadoop", *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*, Vol.14 No.2, 2015.

[9] L. Li, F. Noorian, D.J. Moss, and P.H. Leong, "Rolling window time series prediction using MapReduce," In *15th International Conference on Information Reuse and Integration (IRI)*, pp.757-764, 2014.

[10] Y. Chen, Z. Wu, Z. Li, and Y. Zhang, "Research on time series forecasting model based on moore automata", In *Proc. of International Conf. On Advanced Data Mining and Applications*, pp.98-

105, Springer, Berlin, Heidelberg, 2010. https://doi.org/10.1007/978-3-642-17316-5_9.
[11] S. Mehrmolaei, and M.R. Keyvanpour, "Time series forecasting using improved ARIMA", In *Proc. of International Conf. On Artificial Intelligence and Robotics (IRANOPEN)*, pp.92-97 2016.
[12] R.M. Nabilah, Z. Othman, and B. A. Azuraliza, "Approaches of Handling Uncertain Time Series Data towards Prediction", *International Journal of Future Computer and Communication*, Vol.5, No.6, pp.233, 2016. https://doi.org/10.18178/ijfcc.2016.5.6.477.
[13] N.F.M. Radzuan, Z. Othman, and A.A. Bakar, "Uncertain time series in weather prediction", *Procedia Technology*, Vol.11, pp.557-564, 2013. https://doi.org/10.1016/j.protcy.2013.12.228.
[14] M. Joshi, S. Shaikh, P. Waghmode, and P. Mali, "Farmer Buddy-Weather Prediction and Crop Suggestion using Artificial Neural Network on Map-Reduce Framework", *International Journal of Computer Applications*, Vol.159, No.7, 2017.
[15] K.A. Ismail, M.A. Majid, J.M. Zain, and N.A.A. Bakar, "Big Data prediction framework for weather Temperature based on MapReduce algorithm", In *Open Systems (ICOS), 2016 IEEE Conference*, pp.13-17, 2016.
[16] K. Namitha, A. Jayapriya, and G. Santhosh Kumar, "Rainfall Prediction using Artificial Neural Network on Map-Reduce Framework," Proceedings of the Third International Symposium on Women in Computing and Informatics. ACM, 2015.
[17] C. P. Shabariram, K. E. Kannammal, and T. Manojpraphakar, "Rainfall analysis and rainstorm prediction using MapReduce Framework." Computer Communication and Informatics (ICCCI), 2016 International Conference on. IEEE, 2016.
[18] A. Nair, Gurjeet Singh, and U. C. Mohanty. "Prediction of Monthly Summer Monsoon Rainfall Using Global Climate Models Through Artificial Neural Network Technique." Pure and Applied Geophysics, vol. 175, no. 1, pp. 403-419, 2018. https://doi.org/10.1007/s00024-017-1652-5.
[19] S. Shajitha Banu, S. Manjula, S. Swathi Priya, V. Yamuna Devi, M. Thangamani, "Predictive Analysis of Rainfall Data to Help the Farmers", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol.6, No.3, 2016.
[20] A. Kaur, "Big Data: A Review of Challenges, Tools and Techniques", *International journal of scientific research in science, engineering and technology*, Vol.2, No.2, pp. 1090-1093, 2016.
[21] 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*, pp.507-518, 2010.
[22] W. Lu, Y. Shen, S. Chen, and B.C. Ooi, "Efficient processing of k nearest neighbor joins using mapreduce", *Proceedings of the VLDB Endowment*, Vol.5, No.10, pp.1016-1027, 2012. https://doi.org/10.14778/2336664.2336674.
[23] P. Riyaz, and S.M. Varghese, "Leveraging map reduce with hadoop for weather data analytics", *IOSR Journal of Computer Engineering*, Vol.17, No.3, 2015.
[24] B. Anurag, M. Prakash, V. Kanna, P. Choudhary, "Weather Forecasting using Map-Reduce", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol.5, No.9, 2017.
[25] G.J. Martínez, J.C. Seck-Tuoh-Mora, and H. Zenil, "Wolfram's classification and computation in cellular automata Classes III and IV", In *Proc. of International Conf. On Irreducibility and Computational Equivalence*, Berlin, Heidelberg, pp.237-259, 2013. https://doi.org/10.1007/978-3-642-35482-3_17.
[26] A. Kavousi-Fard, H. Samet, and F. Marzbani, "A new hybrid modified firefly algorithm and support vector regression model for accurate short term load forecasting", *Expert systems with applications*, Vol.41, No.13, pp.6047-6056, 2014. https://doi.org/10.1016/j.eswa.2014.03.053.
[27] A. Belayneh, J. Adamowski, B. Khalil, and B. Ozga-Zielinski, "Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models", *Journal of Hydrology*, Vol.508, pp.418-429, 2014. https://doi.org/10.1016/j.jhydrol.2013.10.052.
[28] A.R. Imon, M.C. Roy, and S.K. Bhattacharjee, "Prediction of rainfall using logistic regression", *Pakistan Journal of Statistics and Operation Research*, Vol.8, No.3, pp.655-667, 2012. https://doi.org/10.18187/pjsor.v8i3.535.
[29] Geetha, A. and Nasira, G.M., 2016. Time-series modelling and forecasting modelling of rainfall prediction using ARIMA model. *International Journal of Society Systems Science*, *8*(4), pp.361-372. https://doi.org/10.1504/IJSSS.2016.081411.