Big Data Analytics for Rainfall Prediction using MapReduce-Based Regression Model

Kyi Lai Lai Khine, Thi Thi Soe Nyunt
University of Computer Studies, Yangon
kyilailai67@gmail.com, thithisn@gmail.com

Abstract

The most significant climatic element which impacts on agriculture sector is rainfall and rainfall prediction becomes an important issue in agriculture country like Myanmar. Collecting, storing and processing of huge amount of climatic data (Big Data) require high-performance analytical systems running on distributed environments for accurate prediction of weather. Traditional standard data analytics algorithms need to be adapted to take advantage of cloud computing models which provide scalability and flexibility. In this paper, Multiple Linear Regression which is an empirical, statistical and mathematically mature method in data analysis is applied in Rainfall Prediction. To prove conventional Multiple Linear Regression work efficiently in distributed environments, we propose a parallel processing of Regression Model called MapReduce-based Multiple Linear Regression (MR-MLR). Weekly Rainfall Prediction with the proposed regression model using large scale weather data will base on the QR Decomposition and Ordinary Least Squares method adapted to MapReduce Framework. Correlation-based Filter Approach by using Symmetrical Uncertainty (SU) will be applied in selecting correlated and relevant features for improving the proposed regression model's prediction accuracy.

Keywords*: Big Data, Rainfall Prediction, Multiple Linear Regression, QR Decomposition, Ordinary Least Squares, Symmetrical Uncertainty*

1. Introduction

The term big data is derived from the fact that the datasets are so large that typical database systems are not able to store and analyze the datasets. Big data moves around 7 Vs- volume, velocity, variety, value and veracity, variability and visibility. Storing huge volume of data available in various formats which is increasing with high velocity to gain values out it is itself a big deal [5]. As data is tremendously increasing in meteorological domain, storing and analyzing them becomes a great challenge for environmentalist and most of the burning issues of our time like global warming, floods, draught, heat waves, soil erosion and many other climatic issues are directly related with rainfall.

Prediction of rainfall is still a huge challenge to the climatologists and agriculture which is greatly dependent on rainfall requires analysis of weather data related with rainfall. Accurate rainfall prediction results in better control of water availability, more refined operation of reservoirs and improved hydropower generation and also increment of crop productivity rate. Fundamentally, there are two approaches to predict rainfall. They are empirical and dynamical methods. The empirical approach is based on analysis of historical data of the weather and its relationship to a variety of atmospheric and oceanic variables over different parts of the world. The most widely use empirical approaches used for climate prediction are regression, artificial neural network, stochastic, fuzzy logic and group method of data handling. In dynamical approach, predictions are generated by physical models based on systems of equations that predict the evolution of the global climate system in response to initial atmospheric conditions.

Regression is an empirical statistical technique and is widely used in business, the social and behavioral sciences, the biological sciences, climate prediction, and many other areas. In this paper, predictive analysis of rainfall with the use of empirical statistical regression approach is implemented on parallel and distributed environment to meet scalability and flexibility needs for big weather data. The structure of the paper is organized as follows. The next section 2 presents the related works to the proposed system. Theoretical background of the paper such as big data analytics, regression analysis and multiple linear regression are described in section 3. Then, the explanation for the proposed system including system design, dropping

irrelevant independent variables and MapReduce-Based Multiple Linear Regression (MR-MLR) with QR Decomposition are discussed in section 4. Some experimentation is presented in section 5 and the paper is concluded and future works are described in section 6.

2. Related Work

S. Shajitha Banu, et al. [4] presented an analysis system to analyze rainfall and water inflow patterns in a dam based on MapReduce of Hadoop. Building an analytical engine to perform analysis over huge data and perform computation over the data (calculating average) using MapReduce and predicting or forecasting future pattern using R-programming. A.P.Dhananjay and K.Deepak [2] discussed about the statistical relationship between rainfall amount and other climate data each search with the use of second order MLR equation and Multiple Linear Regression technique gives more efficient result than other technique. They computed the rainfall prediction day wise, week wise, month wise using seven years data of five cities in India i.e. Nagpur, Pune, Mumbai, Chennai and Delhi. B.R.Austin, et al. [3] described how to compute a stable tall-and-skinny QR factorization on MapReduce architecture in only slightly more than 2 passes over the data. They also showed the performance comparison between the direct QR factorization and indirect QR factorization on MapReduce Architecture. A.Moufida Rehab and B. Faouzi [1] proposed the parallel version of the multiple linear regression can efficiently handle very large datasets on commodity hardware with a good performance on different evaluation criterions, including number, size and structure of machines in the cluster of Hadoop MapReduce. They presented that with a distributed algorithm based on the MapReduce paradigm, we can manage to increase the processing performance by avoiding the memory limits. W.T.Zaw, et al. [14] predicted the rainfall over Myanmar. For rainfall prediction over Myanmar used second order Multi variables polynomial regression (MPR). The MPR is a way to describe the complex nonlinear input output relationships that why outcome variable can be predicted from the other or others. I.Lily, et al. [8] indicated that too many or possibly redundant features can cause the rainfall forecasting to be inefficient and lower the accuracy. Therefore, the selection of relevant features and elimination of irrelevant and redundant ones are primarily need to increase in prediction accuracy and avoid over fitting of the training data. A.H.M. Rahmatullah Imon, et al. [12] showed a logistic regression model can successfully predict rainfall provided that all its important predictors are in place. They also proved that rainfall can be successfully predicted by the climatic variables such as maximum temperature, minimum temperature, evaporation, morning and afternoon humidity.

3. Big Data Analytics

Big data analytics can be defined as the combination of traditional analytics and data mining techniques along with large volumes of data to create a fundamental platform to analyze, model and predict the behavior of customers, markets, products, services and the competition, thereby enabling an outcome-based strategy precisely tailored to meet the needs of the enterprise for the market and customer segment.

There are three types of big data analytics:

Descriptive Analytics, which use data aggregation and data mining techniques to provide insight into the past and answer: “What has happened?”

Prescriptive Analytics, which use optimization and simulation algorithms to advice on possible outcomes and answer: “What should we do to happen in future?”

Predictive Analytics, which use statistical models and forecasts techniques to understand the future and answer: “What could happen in future?”

Predictive Big Data Analytics comprises a variety of techniques that predict future outcomes based on historical and current data [6]. Statistics is a good analytical tool for big data analysis because big data is the data considered in statistics.

3.1. Regression Analysis

Regression analysis is a statistical method for big data analysis because regression model is popular for data analysis including big data analysis. It is not only a statistical process for estimating the relationships between variables, but it is also widely used for prediction and forecasting. Moreover, regression is a technique that utilizes the relation between two or more quantitative variables on observational database so that an outcome variable can be predicted from the others. In regression technique for single variable use simple regression and multiple variables use multiple regression [9].

The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

3.2. Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical model used to describe a linear relationship between a dependent variable called "explain" or "endogenous", and a set of independent or predictor variables called "explanatory" or "exogenous" reflecting observable phenomena. It is possible to estimate this relationship statistically, from a series of observations. The simplest form of regression, linear regression, uses the formula of a straight line ($y_i = \beta_i x_i + \xi$) and determines the appropriate value for β and ξ to predict the value of y based on the inputs parameters, x. Therefore, multiple linear regression model is represented by the following:

$$\mathbf{Y}=\boldsymbol{\beta_0}+\boldsymbol{\beta_1}\boldsymbol{X_1}+\boldsymbol{\beta_2}\boldsymbol{X_2}+\cdots+\boldsymbol{\beta_n}\boldsymbol{X_n}+\boldsymbol{\varepsilon} \qquad (1)$$

where Y is the dependent or explain variable; $X_1, X_2,\ldots,X_n$ are the independent or explanatory variables measured without error (not random); $\beta_0,\beta_1,\ldots..,\beta_n$ are the parameters of the model. This equation specifies how the dependent variable Y is connected to the independent variables X. Objective is to find $\beta_0,\beta_{1,\ldots\ldots..},\beta_n$ so that the sum of squared errors is the smallest (minimum). A primary goal of a regression analysis is to estimate the relationship between the predictor Y and the target variables X or equivalently, to estimate the unknown parameter β to find the influence degree β_0, $\beta_{1,\ldots\ldots..,n}$ of these factors (X) on the variable (Y).

4. The Proposed System

4.1. Problem Statement

The occurrence of prolonged dry period or heavy rain at the critical stages of the crop growth and development may lead to significant reduce in crop yield. Myanmar is an agricultural country and its economy is largely based upon crop productivity. Thus, rainfall prediction becomes an important factor in agricultural countries like Myanmar. In this paper, we would like to build the regression model for weekly rainfall prediction using large scale weather data related with rainfall. The resulted rainfall amounts (in mm) are intended to help farmers in making decision concerning with their crop. It is possible to predict weekly rainfall amount with one weekly ahead with acceptable accuracy.

In construction of regression model, Multiple Linear Regression with massive weather data processing will apply for rainfall prediction. In this work, our contribution is to show that the adaptation of classical learning algorithms of data is possible to provide a response to the phenomenon of big data. We particularly focused on the adaptation of the Multiple Linear Regression with large scale data processing. Multiple Linear Regression proves unsuited to the scalability of the data processed and its principle treatment is focused on a central approach, where the computation is done on a set of data stored in a single machine. The use of parallel and distributed computing with MapReduce paradigm seems like a natural solution to this problem.

4.2. Proposed System Design

Our proposed system consists of the following major steps:

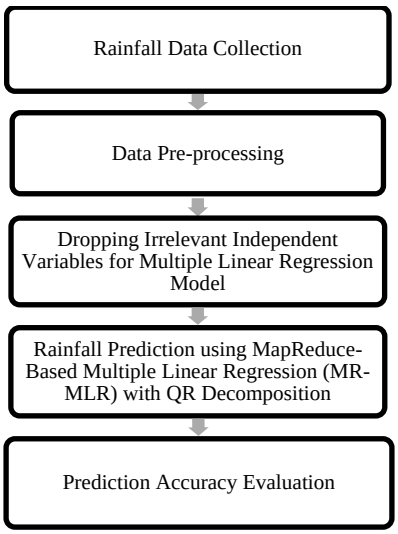

Figure 1. Proposed System Design

Rainfall data are collected as the first step. Data pre-processing procedures for rainfall data are then applied. Choosing the best predictors for regression model is important in defining the optimal regression equation for better prediction accuracy of the system. Therefore, the next step is removing the irrelevant independent variables or predictors for the regression model by applying the proposed algorithm in figure 2. After that, the statistical relationship between weekly rainfall amount and other climatic data or predictors for rainfall is computed through our proposed model MR-MLR, MapReduce-Based Multiple Linear Regression with QR Decomposition.

For estimating the proposed model's parameter, β is calculated the least squares method by using equation (8). The least square prediction Y for weekly rainfall amount is applied with equation (1). The sum of squared errors from the proposed regression model should be the smallest (minimum) to get the better prediction accuracy for the rainfall. Finally, prediction model accuracy will be evaluated by determining the prediction error comparing with the model's predicted rainfall amount and observed rainfall amount.

4.3. Dropping Irrelevant Independent Variables

It is always better to make predictions with models that do not include irrelevant variables. It can be assumed that the independent variables in training data as the features for dependent variable or predicted output to make predictions. Thus, we would like to select the most important features for the Multiple Linear Regression Model. In this way, we could also drop irrelevant features or independent variables for the model. In this section, we will discuss how to evaluate the goodness of features for prediction. In general, a feature is *good* if it is *relevant* to the class concept or target output variable but is not *redundant* to any of the other relevant features. Moreover, it is adopted that a feature is good if it is highly correlated to the target class but not highly correlated to any of the other features [15].

In this system, Correlation-based Filter Approach by using Symmetrical Uncertainty (*SU*) will be applied as the goodness measure for selecting important features for the proposed regression model. We would like to contribute Symmetrical Uncertainty (*SU*) for dropping irrelevant independent variables or selecting relevant and important features by developing a procedure based on correlation analysis of features (including the class or target variable) described in figure 2.

Algorithm 1: Dropping Irrelevant Independent Variables

Input: S (F_1, F_2, ... F_N, C) // Features and target class from training data set

Output: S_{List} // Ranked List for Selected Features

1. Begin
2. for i=1 to N do
3. begin
   a. Calculate $SU_{Fi,C}$ for F_i
   b. if ($SU_{Fi,C} > 0$) Append F_i to S_{List};
4. end;
5. Order S_{List} in descending $SU_{Fi,C}$ value;
6. End

Figure 2. Proposed Algorithm for Dropping Irrelevant Independent Variable

4.3.1. Calculation steps in Symmetrical Uncertainty (SU)

Step 1: The entropy of a variable *X and Y* is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \quad (2)$$

$$H(Y) = -\sum_i P(y_i)\log_2(P(y_i)) \quad (3)$$

Step 2: The entropy of *X* after observing values of another variable *Y* is defined as

$$H(X|Y) = -\sum_i P(y_i)\sum_i P(x_i|y_i)log_2(P(x_i|y_i)) \quad (4)$$

Step 3: The amount by which the entropy of *X* decreases reflects additional information about *X* provided by *Y* and is called *Information Gain* (IG) given by

$$IG(X|Y) = H(X) - H(X|Y) \quad (5)$$

Step4: The *Symmetrical Uncertainty* (SU) between two variables is defined as

$$SU(X|Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right] \quad (6)$$

4.4. MapReduce-Based Multiple Linear Regression (MR-MLR) with QR Decomposition

Multiple Linear Regression is among the most powerful and mathematically mature method in data analysis. Its principle treatment is focused on a central approach, where the computation is done on a set of data stored in a single machine. With an increasing volume of data, the transition to the scalability of the algorithm is indispensable. As a

result, training Multiple Linear Regression on a single machine is usually very time-consuming to finish or even cannot be done. Hadoop is an open framework used for big data analytics and its main processing engine is MapReduce, which is currently one of the most popular big data processing frameworks available. MapReduce is a framework for executing highly parallelizable and distributable algorithms across huge data sets using a large number of commodity computers [13].

To overcome the limitations of Multiple Linear Regression and adaptable for huge amount of data, we would like to present a new computational approach, MapReduce with QR Decomposition. QR Decomposition is one of the steps for the resolution of Multiple Linear Regression. We will subsequently describe our scalable approach based on the QR decomposition and the ordinary least squares method in MapReduce paradigm. MapReduce-Based Multiple Linear Regression with QR Decomposition is proposed in this paper to make conventional Multiple Linear Regression work efficiently in parallel and distributed environment like cloud computing platform and MapReduce Paradigm that can meet the challenges of big data.

QR Decomposition (also called QR Factorization) is one of the most common decomposition matrices in scientific computing to solve the problems of ordinary least squares. The QR decomposition of a matrix X is a decomposition of the latter into an orthogonal matrix Q and an upper triangular matrix R. QR decomposition is a decomposition of X such that:

$$\mathbf{X = QR} \qquad (7)$$

The matrix formulation of the Multiple Linear Regression is

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

The QR Decomposition or Factorization of X looks like

$$\mathbf{X} = \begin{pmatrix} \mathbf{X_1} \\ \hline \mathbf{X_2} \\ \hline \mathbf{X_3} \\ \hline \mathbf{X_4} \end{pmatrix} = \begin{pmatrix} Q_1R_1 \\ \hline Q_2R_2 \\ \hline Q_3R_3 \\ \hline Q_4R_4 \end{pmatrix} \text{ Where: } X_i = Q_i R_i$$

To determine the Multiple Linear Regression Model's coefficient, β the method is to simplify the calculation by decomposing the data matrix X into two matrices Q and R obtained with the QR decomposition, and thus:

$$\beta = (X^T X)^{-1} X^T Y$$

By Replacing X=QR,

$$\beta = (Q^T R^T Q R)^{-1} Q^T R^T Y$$

$$= (R^T R)^{-1} Q^T R^T Y$$

$$= (R^T)^{-1} (R)^{-1} Q^T R^T Y$$

$$= (R^T)^{-1} (R)^{-1} Q^T R^T Y$$

$$= (R)^{-1} Q^T Y$$

Finally, we obtain:

$$\boldsymbol{\beta = (R)^{-1} Q^T Y} \qquad (8)$$

4.4.1. Two-Stage Process for MR-MLR with QR Decomposition

In this section, we would like to present how to implement Multiple Linear Regression with QR Decomposition on Mapreduce illustration in figure 3 and 4. In figure 3, we show the first stage of the process. It takes input as the matrix of observations X (m,n) for training data and decomposes and distributes it as several matrices X_i (Blocsize, n) on several tasks "map". Every map task has the same QR factorization function. The block number will be generated as:

$$nbBloc = \frac{m}{\text{Bloc size}} \qquad (9)$$

The results matrices Q_i (Bloc size, n) and R_i (n, n) are associated with the key «Key_i » (i: equal to nbBloc) and sent them to "reduce". Each R_i will be used to construct the matrix R_{temp} (n * nbBloc, n) by superposing the matrices R_i. At the end of this stage, the QR decomposition is applied to the matrix R_{temp}. R_{final} with the key "R" and each Q'_i is associated with the key «Key_i» (i: equal to nbBloc) will be saved in the output files of the first stage.

In figure 4, we illustrate the second stage of the process. It takes input as the result of the first stage and the vector y. In the step of "map", the vector y is decomposed into several vector y_i (Bloc

size) and sent to “reduce” with "Key_i ". The task "reduce" will execute the input data according to the associated key. If the key is "R" then R_{final} is saved and will be used in the calculation of β at the end of process. Otherwise, we will need necessarily both matrices Q_i, Q'_i and y_i vector to be used in the calculation of the vector V_i as follows:

$$Q := \text{Multiply}\ (Q_i, Q'_i)$$
$$Q^T := \text{Transpose}\ (Q)$$
$$V_i := \text{Multiply}\ (Q^T, y_i)$$

At the end of process, the V_i vectors are added to have the final vector V.

$$\beta := \text{Solve}(R_{final}, \sum Vi[\,])$$

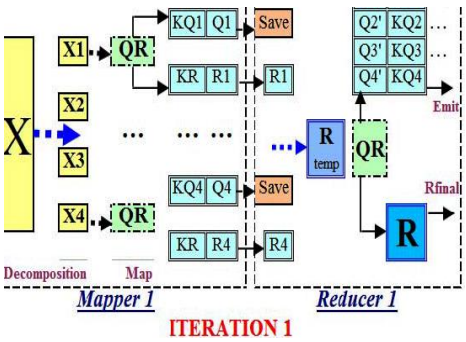

Figure 3. First Stage of MapReduce-Based Multiple Linear Regression

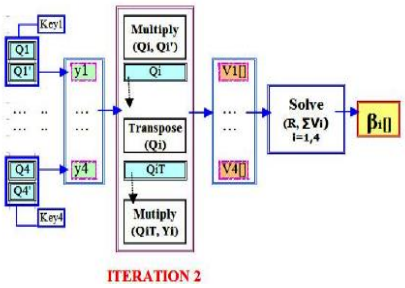

Figure 4. Second Stage of MapReduce-Based Multiple Linear Regression

5. Experimentation

To test our approach, a Hadoop Cluster in version 2.7.1 is created. Then, the java implementation for the Multiple Linear Regression on Mapreduce is tested with sample weather dataset related with rainfall. The evaluation result is still ongoing and we will demonstrate later that our parallel version of the Multiple Linear Regression based on MapReduce Framework can efficiently handle large scale weather data for rainfall prediction on commodity hardware with a good performance on different evaluation criterions, including number, size and structure of machines in the cluster.

6. Conclusion and Future Work

In this system, our contribution is to show that the adaptation of classical learning algorithms of data generally and predictive algorithms practically is possible to provide a response to the phenomenon of big data. We focused particularly on the adaptation of the Multiple Linear Regression with QR Decomposition for massive data processing. With a parallel and distributed version based on the MapReduce paradigm, we will intend to increase processing performance by avoiding the memory limits on massive data providing scalability and flexibility. The predicted rainfall amounts from the proposed MR-MLR model are intended to help farmers in making decision concerning with their crop productivity rate. We also contributed an algorithm in dropping independent variables for proposed MR-MLR model’s prediction accuracy purpose. In future work, we will prove that our proposed algorithm for dropping irrelevant independent variables could efficiently improve the model’s prediction accuracy.

References

[1] A.Moufida Rehab and B. Faouzi, “A massively parallel processing for the Multiple Linear Regression”, Tenth International Conference on Signal-Image Technology & Internet-Based System, IEEE, 2014

[2] A.P.Dhananjay and K.Deepak, “Statistical Modeling for Rainfall Prediction using Data Mining Technique”, An International Journal of Engineering & Technology, Vol. 2, No. 1, January, 2015

[3] B.R.Austin, et al., “Direct QR factorizations for tall-and-skinny matrices in MapReduce architectures”, 2013.

[4] B. Shajitha, et al., “Predictive Analysis of Rainfall Data to Help the Farmers”, Volume 6, Issue 3, March 2016

[5] C. Florina and G. Elena, “Perspectives on Big Data and Big Data Analytics”, 2013

[6] G. Amir and H. Murtaza, “Beyound the hype: Big data concepts, methods and analytics”, International Journal of Management, 2014

[7] H.Urban Nchimunya, “Predictions of Future Aspects of the Rainy Season Using Simple and Multiple Linear Regression Analysis- A Case Study of Chingóme Mission Daily Rainfall Data in Zambia”, International Journal of Applied Science and Technology, 2013

[8] I.Lily, et al., “Machine Learning Techniques for Short-Term Rain Forecasting System in the Northeastern”, International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol.2, No.5, 2008

[9] J.Sunghae, et al., “A Divided Regression Analysis for Big Data”, International Journal of Software Engineering and Its Applications, Vol. 9, No. 5, 2015

[10] L.James, “Selection Process for Multiple Regression”, 2010

[11] M.Sanjay, et al., “A Simple Weather Forecasting Model Using Mathematical Regression”, Indian Research Journal of Extension Education Special Issue, 2012

[12] R.C. Manos, et al., “Prediction of Rainfall Using Logistic Regression”, July, 2012

[13] V.Surekha Mariam and A.P.Riyaz, “Leveraging MapReduce with Hadoop for Weather Data Analytics”, IOSR Journal of Computer Engineering (IOSR-JCE), Volume 17, Issue 3, Ver. II (May – Jun), 2015

[14] W.T.Zaw, et al., “Empirical Statistical Modeling of Rainfall Prediction over Myanmar”, World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol.2, No.10, 2008

[15] Y.Lei and L.Huan, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution”, Proceedings of the Twentieth International Conference on Machine Learning, Washington DC, 2003