IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

A Survey Paper on Climate Changes Prediction Using Data mining

E. Sreehari[1], J. Velmurugan[2], Dr. M. Venkatesan[3]

M.Tech Scholar, Department of Computer Science and Engineering,

Sri Venkateswara College of Engineering and Technology, Chittoor, India[1]

Associate Professor, Sri Venkateswara College of Engineering and Technology, Chittoor[2]

Research Scholar in VIT University, Vellore, India[2]

Associate Professor, School of Computing Science & Engineering, VIT University, Vellore, India[3]

Abstract: The purpose of data mining effort is generally to create a descriptive model or a predictive model. In this paper the concepts of regression was summarized to accomplish the task of prediction and various methodologies of regression and its significance was illustrated. The methodologies include such as multiple regressions, covariance matrix regression method and other methods can be defined here. We provide all the concepts of regression as a framework to achieve prediction in different paths. This paper can also define other best and consistent approaches or methods for prediction in which that can be concluded by other research scholars and scientists.

Keywords: Data Mining, Regression, Prediction, Regression framework, Coefficients.

I. INTRODUCTION

Regression is the popular technique used for prediction in areas like climate prediction and other areas. The climate can be defined as the average state of atmosphere over a longer period similarly weather can be changes strongly day by day. The natural disasters which was occurred during the year 2015 in the regions such as Chennai, Vishakhapatnam, Jammu & Kashmir and Uttarakhand which may results in the loss of property, reducing natural resources and death of human, natural species also. If we use the concept of prediction there may be a scope to mitigate losses and including property and damage of the public. Weather prediction may be at regional or national levels. Generally, two approaches used for predicting rainfall. One is Empirical approach and other is Dynamical approach. Empirical based on historical data to be collected and its relationship to various atmospheric variables. Dynamical approach, defines physical models based on systems of equation for prediction and can be implemented by using numerical rainfall forecasting method [6, 26].

The most widely used empirical approaches which are used for climate prediction; they are regression, artificial neural network, fuzzy logic and group method of data handling. Support vector machines [18] are a set of supervised learning methods that create a decision maker system which tries to predict new values. A simple climate forecasting [7] can be done by regression techniques. In the recent years, the use of data mining process in the field of hydrology is increasing. The studies have been performed using data mining process in many areas [14-16]. Integrated evaporation model [17], using DM process for three lakes in Turkey by Keskin. Now-a-days Artificial Intelligent methods used in the estimation of rainfall [19-25].

The rest of the paper is organized as follows. In Section 2, we give a brief overview of the regression evaluation in which it can represent regression's Origen and structure. In Section 3, we explain our methodologies, models and solution to the climate prediction using regression. In Section 4, we define the conclusion and represented other several baselines for regression, acknowledgement and references concluded at the end.

II. THE WHOLLY REGRESSION EVALUATION

Regression [10] is a statically empirical technique widely used for prediction and forecasting and it can estimate relationship among variables. Regression analysis infers the casual relationships between the independent and dependent variables. A regression analysis method depends on the form of the data generating process.

a. History of Regression

The earliest form of regression was the method of least square which was published by Legendre in 1805 and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the sun. Gauss published a further development of the theory of least squares in 1821, including a version of the Gauss-Markov theorem. The term regression was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. In the 1950s and 1960s, economists used electromechanical desk calculators to calculate regressions. Before 1970 it sometimes took up to 24 hours to receive the result from one regression.

Regression methods continue to be an area of active research. In recent decades, new methods have developed for robust regression, regression involving correlated

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

responses such as time series and growth curves, images, graphs and other complex data objects. Regression is the one in which the predictor variables are measured with error.

b. Regression Structure

A Simple Regression [13] equation in the form

$$Y=A+BX \quad (1)$$

Here,
Y=Dependent variable.
X=Independent variable.
A, B are Regression parameters that provide information the dependent variable:

- If N data points N<K, Regression Assume now that the vector of unknown parameters B is of length k, In order to perform a regression analysis the user must analysis cannot be performed.
- If exactly N=K data points are observed, and the function f is linear, the equation Y=f(X, B) can be solved.
- The most common situation is where N>K data points are observed. In this case, there is enough information in the data to estimate a value for B that best fits the data in some sense.

III. REGRESSION METHODOLOGIES

Regression in simple terms is defined as predicting one variable from another also called simple linear regression [13]. Later statisticians coined the term multiple regression to describe the process by which several variables are used to predict another. The following are the various multiple regression methodologies are to be described here.

1. Multiple Linear Regression
2. Matrix formulation of Multiple Linear Regression.
3. Regression Equation using Covariance matrix method.
4. Multi Structure Regression equations for same dependent variable with different independent alternatives.

a. Multiple Linear Regression

In multiple regressions [3] there are more than two variables among one is dependent variable and all others are independent variables and equation look like this:

$$Y_i = B_0+B_1X_{i1}+B_2X_{i2}+B_3X_{i3}\text{--------}B_nX_{in} \quad (2)$$

To develop the multiple linear regression the parameter are obtained from the training data and variable are extracted from the dataset using correlation.

The term r, called the linear correlation coefficient measure the strength and direction of relationship between the two variables. The linear correlation coefficient sometimes called Pearson's correlation coefficient.

The mathematical formulae for r [11] is given as

$$r = \frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{n(\sum x^2)-(\sum x)^2}-\sqrt{n(\sum y^2)-(\sum y)^2}}$$

The coefficient of determination describes how well the regression line represents data, if the regression line passes through every point on the scatter plot it would be able to explain all of the variation.

b. Matrix formulation of Multiple Linear Regression

A multiple regression analysis [10] is work out to predict the values of dependent variable Y, with given a set of explanatory variables (x1, x2, x3------xn). Consider an equation for each observation:

$$Y_1 = B_0+B_1X_{i1}+B_2X_{i2}+B_3X_{i3}\text{--------}B_nX_{in}$$
$$Y_2= B_0+B_1X_{i1}+B_2X_{i2}+B_3X_{i3}\text{--------}B_nX_{in}$$
$$\vdots$$
$$Y_n = B_0+B_1X_{i1}+B_2X_{i2}+B_3X_{i3}\text{--------}B_nX_{in}$$

The MLR model in Matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} B_0+B_1X_{11}+B_2X_{12}+B_3X_{13}\text{--------}B_nX_{1n} \\ B_0+B_1X_{21}+B_2X_{22}+B_3X_{23}\text{--------}B_nX_{2n} \\ \vdots \\ B_0+B_1X_{i1}+B_2X_{i2}+B_3X_{i3}\text{---------}B_nX_{in} \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12}\ldots\ldots X_{1n} \\ 1 & X_{21} & X_{22}\ldots\ldots X_{2n} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2}\ldots\ldots X_{nn} \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_n \end{bmatrix}$$

$$Y=XB$$

- **X** is called the design matrix.
- **B** is the vector of parameters.
- **Y** is the response vector.
- **E** is the error vector.

$$Y=XB+E.$$

Parameter Estimation values can be exploited by solving this matrix equation

$$B = (X^TX)^{-1}X^TY.$$

Steps to calculate the coefficients [4] (parameters):

1. Construct matrix X and Y where Y matrix indicates nx1 and X matrix indicates nxn.
2. Find X^T
3. Multiply matrices X and X^T
4. Multiply matrices X^T and Y
5. Find the inverse for result obtained in step 3.
6. Multiply the resultant matrix with step 4.

Thus this is the process of calculating the coefficients for constructing regression equation.

c. Regression Equation using Covariance matrix method

NOTE: To evaluate covariance matrix approach [12] it is necessary to install Data Analysis tab by the following steps

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Office Button > Excel Options > Add-Ins > Go>Analysis ToolPack

Steps to achieve regression using covariance approach:

STEP 1: Declare all independent variables and dependent variable values.
STEP 2: Select Data tab and in Data Analysis choose covariance then click ok. Then define input and output ranges.
STEP 3: The lower triangular matrix was the output, with the obtained result transpose matrix and it will become upper triangular matrix.
STEP 4: Construct the covariance matrix by combining both upper and lower triangular matrices and keeping diagonal elements as constant values.
STEP 5: Define equations with covariance matrix by eliminating dependent variable column.
STEP 6: With the obtained result construct two matrices one from independent variable called "**X**" and other matrix called "**Y"** came from dependent values of covariance matrix of equations.
STEP 7: Accomplish the task in a given form $\mathbf{B=X^{-1}Y.}$

The outcomes are called coefficients or unknown parameters ($\mathbf{B_1}$**, B2, B3,…Bn**).

The main task is to calculate the intercept for equation.

The mathematical formulae for intercept term is

STEP 1: Calculate averages individually for both dependent and independent variables.
STEP 2: Solve M and multiply (Independent variables) Coefficients values and store in **Z.** Finally substitutes with the obtained values for getting intercept

$\mathbf{B_0}$=Avg (dependent Variable)

Construct the regression equation with calculated coefficient values and Intercept then compute the values as usual.

d. Multi Structure Regression Equation
This method predicts the data [9] by regression analysis through dependent and independent variables. It includes the same dependent variable and different independent variables based on the various instances.

Schema 1: Defining regression equation by considering one dependent and one independent variable, using regression term in Excel.

EX: Regression analysis: Rating versus Moisture

Schema 2: Similarly construct regression equation by considering same dependent variable and other single independent variable, which is not considered in schema1.

EX: Regression analysis: Rating versus Sweetness

Schema 3: This pattern defines the need to consider the same dependent variable and independent variables of both schema 1 and schema 2. Equation can be generated by using regression term in which it can produces the Intercept and coefficient values, and these terms are enough for implementing regression criteria.

EX: Regression analysis: Rating versus Moisture, Sweetness.

By three schemas in the multi structure regression methodology can produces different regression equations and they are not similar to each other, can also produces three different alternatives for prediction [5].

IV. CONCLUSION

This paper presents a survey that using Data mining Regression methodologies for prediction and can be considered as an alternative to traditional metrological approaches. The study describes the capabilities of various algorithms in predicting several weather phenomena such as temperature, thunderstorms, rainfall and so on. Data mining regression techniques like firstly apply correlation analysis then regression analysis. So that we can predict the future year by knowing climate factors which is very useful for public. This is the only for prediction regarding different areas but not accurate because of natural indices like climate factors. As we know that climate factors changes due to different reasons and here we define some best methods to achieve prediction based on empirical studies some climate scientists and research scholars Comparison is made in the other papers, which shows that decision trees and k-mean clustering [1] are best suited data mining technique for this application. With the increase in size of training set, the accuracy is first increased but then decreased after a certain limit. With all the observations made the regression is considered to be best method for prediction [2].

ACKNOWLEDGMENT

I acknowledge my sincere and profound gratitude to my guide, **J. Velmurugan**, for his valuable guidance, dedicated concentration and support throughout this work. I also acknowledge my sincere gratitude to authorities of Sri Venkateswara College Of Engineering and Technology, Chittoor and other teaching staff of Dept. of Computer Science Engineering for their help and support. I am also thankful to my friends for helping in successful completion of my Literature survey.

REFERENCES

1. Data Mining Techniques for Weather Prediction by Divya Chauahan and Jawhar Thakur published in International Journal on recent and innovation trends in computing and communication.
2. Monthly Rainfall Estimation by Data-Mining Process by Ozlem terzi, Hindwani Publishing corporation in the year 2012.
3. Exploiting Data Mining Technique for Rainfall Prediction by Nikhil Sethi and Dr. Kanwal Garg published in International Journal of Comuputer Science and Information Technologies in the year 2014.
4. Rainfall Prediction using Multiple Regression Technique by Imran Ahmed, Sruthi Menon and Nikitha in the year 2014.
5. Prediction of Rainfall by Datamining Technique in Assam by Pinky Saikia Dutta and Hitesh Tahbilder in the year 2013.
6. Z.Ismail et. al, Forecasting Gold Pieces Using Multiple Linear Regression Method in American Journal of Applied Sciences in the year 2009.
7. Paras, et.al, "A simple Weather Forecasting Model Using mathematical Regression" in Indian Research Journal of extension Education Special Issue (Vol. 1) in the year 2012.
8. http://indiawaterportal.org/met_data/
9. https://onlinecourses.science.psu.edu/stat501/node/311
10. http://www.stat.prude.edu/-jennings/stat514/stat512notes/topic3.pdf

11. http://mathbits.com/MathBits/TISection/Statistics2/correlation.html
12. http://real-statistics.com/multiple-regression/least-squares-method-multiple-regression
13. http://www.oxfordjornals.org/our_journals/tropej/online/ma_chap2.pdf
14. Flood prediction using time series data mining by C. Damle and A. Yalcin, Journal of Hydrology, vol. 333, no.2-4, pp. 305-316, 2007.
15. Data mining and multivariate statistical analysis for ecological system in coastal waters by K. W. Chau and N. Muttil, Journal of Hydroinformatics, vol. 9, no. 4, pp. 305-317,2007.
16. E. P. Roz, Water quality modeling and rainfall estimation: a data driven approach [M.S.thesis], University of Iowa, Iowa city, Iowa, USA, 2011.
17. Datamining process for integrated evoparation model by M. E. Keskin and O. Terzi in the year 2009.
18. Atmospheric Temparature prediction using support vector machines by Y. Radhika and M. Shashi in the year 2009.
19. Data mining techniques for improved WSR-88D rain fall estimation by T. B. Trafalis.M.B. Richman, A.White, and B. Santosa in 2002.
20. An Application of artificial neural networks for rainfall forecasting by J. E. Ball and A. Sharma in the year 2001.
21. Rainfall estimation using artificial neural network group by M. Zhang, A. Scofield and J. Fulcher in 1997.
22. Statistical and geostatistical analysis of rainfall in central Japan by T. Shoji and H. Kitaura in 2006.
23. M. C. V. Ram´ırez, H. F. C. Velho, and N. J. Ferreira, "Artificial neural network technique for rainfall forecasting applied to the S˜ao Paulo region in 2005.
24. R. S. V. Teegavarapu and V. Chandramouli, "Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records in 2005.
25. Y.-M. Chiang, F. J. Chang, B. J. D. Jou, and P. F. Lin, "Dynamic ANN for precipitation estimation and forecasting from radar observations 2007.
26. https://en.wikipidia.org/wiki/Numerical_weather_prediction