International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Applications of Data Mining in Weather Forecasting Using Frequent Pattern Growth Algorithm

Amruta A. Taksande[1], P. S. Mohod[2]

[1]Department of Computer Science and Engineering, RTMNU

[2]Professor, Department of Computer Science and Engineering, RTMNU

Abstract: ***Rainfall forecasting or Weather forecasting has been one of the most challenging problems around the world because it consists of multidimensional and nonlinear data such as in the field of agriculture to determine initial growing season. Recently, climate change causes much trouble in rainfall forecasting. Our Project describes five data mining algorithms namely neural network (NN), random forest, classification and regression tree (CRT), support vector machine (SVM) and k-nearest neighbour. Generally these algorithms are used for the prediction. Generally these algorithms used for prediction. In this project we use five years previous data from Jan 2010-Jan 2014 for Nagpur station. On available datasets we apply the Frequent Pattern Growth Algorithm for deleting the inappropriate data. Generally temperature, humidity, wind speed are mainly responsible for the rainfall prediction .On the percentage of these parameters we predict there is a rainfall or not. Based on experiment result, it can be concluded that the combination of GA and FP growth algorithm weather data can gives prediction with higher than 90% accuracy with several population size and crossover probability.***

Keywords: Data Mining Algorithms, Prediction, Neural Network, Frequent Pattern Growth Algorithm and Weather Forecasting

1. Introduction

Rainfall prediction is nothing but weather forecasting. Weather forecasting is the application of science and technology to predict the state of atmosphere for a given location. Weather forecasting has been one of the most scientifically and technologically challenging problems around the world in the last century. This is due mainly to two factors: first, it's used for many human activities and secondly, due to the opportunism created by the various technological advances that are directly related to this concrete research field, like the evolution of computation and the improvement in measurement systems. To make an accurate prediction is one of the major challenges facing meteorologist all over the world. Since ancient times, weather prediction has been one of the most interesting and fascinating domain. Rainfall is one of several important factors affecting watershed water quality. The downstream flux of nitrogen and phosphorus originating in the watershed basin depends on the amount of rainfall. Once an all-human endeavour based mainly upon changes in barometric pressure, current weather conditions, and sky condition, weather forecasting now relies on computer-based models that take many atmospheric factors into account. Human input is still required to pick the best possible forecast model to base the forecast upon, which involves pattern recognition skills, tele connections, knowledge of model performance, and knowledge of model biases. The chaotic nature of the atmosphere, the massive computational power required to solve the equations that describe the atmosphere, error involved in measuring the initial conditions, and an incomplete understanding of atmospheric processes mean that forecasts become less accurate as the difference in current time and the time for which the forecast is being made (the *range* of the forecast) increases. There are a variety of end uses to weather forecasts.

Weather forecasting is a vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last century. We investigate the use of data mining techniques in forecasting maximum temperature, rainfall, evaporation and wind speed. This was carried out using Artificial Neural Network and Decision Tree algorithms and meteorological data collected between 2010 and 2014 from the city of Nagpur, Maharashtra. The performances of these algorithms were compared using standard performance metrics, and the algorithm which gave the best results used to generate classification rules for the mean weather variables. A predictive Neural Network model was also developed for the weather prediction program and the results compared with actual weather data for the predicted periods.

In this paper we collect the five years data from Jan 2010-Jan2014 from weather department for Nagpur station. Responsible parameters for rainfall prediction are generally temperature, pressure and humidity. On this available datasets we analyse Frequent Pattern Growth Algorithm and calculate Mean Absolute Error (MAE) , Mean Square Error (MSE) and Standard Deviation (SD) factors. Climate is the long-term effect of the sun's radiation on the rotating earth's varied surface and atmosphere. The Day-by-day variations in a given area constitute the weather, whereas climate is the long-term synthesis of such variations. Weather is measured by thermometers, rain gauges, barometers, and other instruments, but the study of climate relies on statistics. Nowadays, such statistics are handled efficiently by computers. A simple, long-term summary of weather changes, however, is still not a true picture of climate. To obtain this requires the analysis of daily, monthly, and yearly patterns. Following are data mining algorithms generally used for rainfall prediction.

2. A Brief Literature Survey

In paper [1], five data-mining algorithms, neural network, random forest, classification and regression tree, support vector machine, and *k*-nearest neighbour were used to build the prediction models. In paper [2] ,rainfall forecasting system using fuzzy system based on genetic algorithm (GA) is made. This paper [3] investigates the use of weather ensemble predictions in the application of ANNs to load forecasting for lead times from one to ten days ahead. This article [4] , presents a comparison of two sub sampling nonparametric methods for designing algorithms to forecast time series from the cumulative monthly rainfall. Both approaches are based on artificial feed-forward neural networks (ANNs).

3. Data Mining Algorithms

Five data-mining algorithms, neural network (NN), random forest, classification and regression tree (C&RT), support vector machine (SVM), and *k*-nearest neighbour (*k*-NN) were used to build the prediction models. NN consists of a group of interconnected neurons, making it an adaptive system that can change its structure based on external or internal information flowing through the network during the learning phase. NNs are usually used to model complex relationships between input and output variables.. Random forest combines decision tree predictors in a way that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. It integrates a bagging idea and a random selection of features in constructing a collection of decision trees. C&RT, popularized by Breiman, is a nonparametric technique producing logical if–then rules that are easy to interpret. An SVM is a supervised learning method used for classification and regression analysis. SVM constructs one or a set of hyper planes in a high or infinite dimensional space. The key advantage of SVM is the use of kernel functions making SVM suitable for modelling in complex nonlinear domains. *k*-NN is an instance-based learning method accounting for contributions of the neighbours.. It offers good performance for some classes of applications.

4. Data and Methodology

4.1 Data Collection

The data used for this work was collected from weather department through the Meteorological Agency from Nagpur station, Maharashtra State. The case data covered for previous five years, that is, January 2010 to January 2014. Nagpur has tropical wet and dry climate with dry conditions prevailing for most of the year. It receives an annual rainfall of 1,205 mm (47.44 inches) from monsoon rains during June to September. The highest recorded daily rainfall was 304 mm on 14 July 1994.The highest recorded temperature in the city was 47.9 °C on May 22, 2013, while the lowest was 3.9 °C. The following procedures were adopted at this stage of the research: Data Cleaning, Data Selection, Data Transformation and Data Mining.

.1.1 Data Cleaning

In this stage, a consistent format for the data model was developed which took care of missing data, finding duplicated data, and weeding out of bad data. Finally, the cleaned data were transformed into a format suitable for data mining.

4.1.2 Data Selection

At this stage, data relevant to the analysis was decided on and retrieved from the dataset. The meteorological dataset had eight (8) attributes, their type and description is presented in Table 1, while an analysis of the numeric values are presented in Table 2. Due to the nature of the Cloud Form data where all the values are the same and the high percentage of missing values in the sunshine data both were not used in the analysis.

Table 4.1.2: Attributes of Meteorological Dataset

Attribute	*Type*	*Description*
Month	Numerical	Month considered
Year	Numerical	Year considered
Temperature	Numerical	Monthly min temperature
Humidity	Numerical	Monthly min humidity
Wind Speed	Numerical	Wind run in km
Sea Level Pressure	Numerical	Pressure in Pascal
Rainfall	Numerical	Total monthly rainfall

4.1.3 Data Transformation

This is also known as data consolidation. It is the stage in which the selected data is transformed into forms appropriate for data mining. The data file was saved in Commas Separated Value (CVS) file format and the datasets were normalized to reduce the effect of scaling on the data. Analyses of numeric values are shown in following table:

Table 4.1.3: Analysis of numeric data values

No	Variable	Min	Max	Mean	MAE	MSE	SD
1	Month	June(1)	June(30)	-	-	-	-
2	Year	2015	2015	-	-	-	-
3	Temperature	27	32	29.5	0.5957	1.0638	0.6981
4	Humidity	73	82	77.5	0.5957	1.0638	0.6981
5	Wind Speed	6	23	14.5	0.5957	1.0638	0.6981
6	Sea Level Pressure	999	1007	1003	0.5957	1.0638	0.6981
7	Rainfall	Rain	Rain	Rain	Rain	Rain	Rain

4.1.4 Data Mining Stage

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the meteorological datasets. The testing method adopted for this research was percentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Thereafter interesting patterns representing knowledge were identified.

4.2 Evaluation Metrics

1) **Mean Absolute Error (MAE):** It is a commonly used quantity in time-series analysis that measures how close the predictions are to the observations.
2) **Mean Square Error:** Measures the difference between values implied by the prediction model and true observations. It incorporates both the variance of the prediction model and its bias. Mean-squared error is one

of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value.

3) **Standard Deviation:** Measures the difference between values implied by the prediction model and true observations. It incorporates both the variance of the prediction model and its bias. It indicates how much variation exists in the values from the average. It is commonly used to measure confidence in statistical conclusions. Mathematical equations are as follows:

$MAE = \sum_{i=1}^{n} : fi - yi :$
$MSE = \sum_{i=1}^{n} (fi - yi)2$
$SD = \sqrt{\sum_{i=1}^{n} (: fi - yi : -MAE)2/(n-1)}$

Where *fi* is the predicted value produced by the model, *yi* is the actual collected value, and *n* represents the number of test data points.

4.3 Project Flow Diagram

The overall project is done with the help of Frequent Pattern Growth Algorithm in the stages of following data flow diagram:

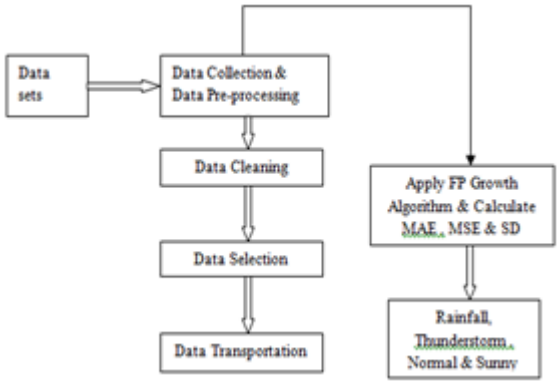

Figure 4.3.1: Work Flow

In this approach project is completed in four stages as shown in above figure i.e data collection and data pre processing, data cleaning, data selection and finally data transportation. Generally responsible parameters for the rainfall prediction are temperature, pressure, humidity and wind speed. These are collected from weather department from Nagpur station and then perform the FP tree algorithm on available datasets and calculate MAE,MSE and SD with the help of basic equations and predict there is a rainfall or not.

5. Result

I calculate the MAE, MSE and SD with the help of basic equations. On available input data sets we apply FP Growth Algorithm with the correct measurement than existing algorithm and evaluate the next year prediction of rainfall in Nagpur region, Maharashtra.

5.1 GUI for the system

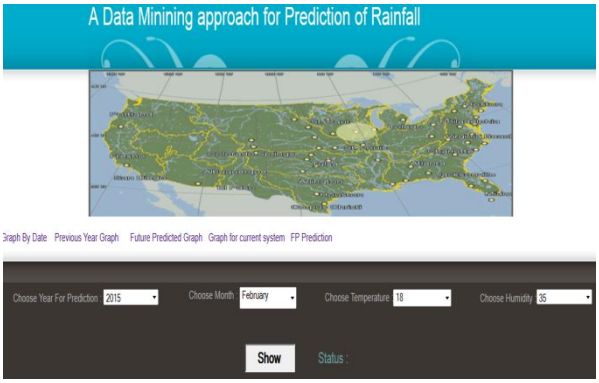

Figure 6.1.1: GUI for the system

Following is the execution of the system:
We are collected input data sets from weather department for previous five years data for Nagpur region, Maharashtra.

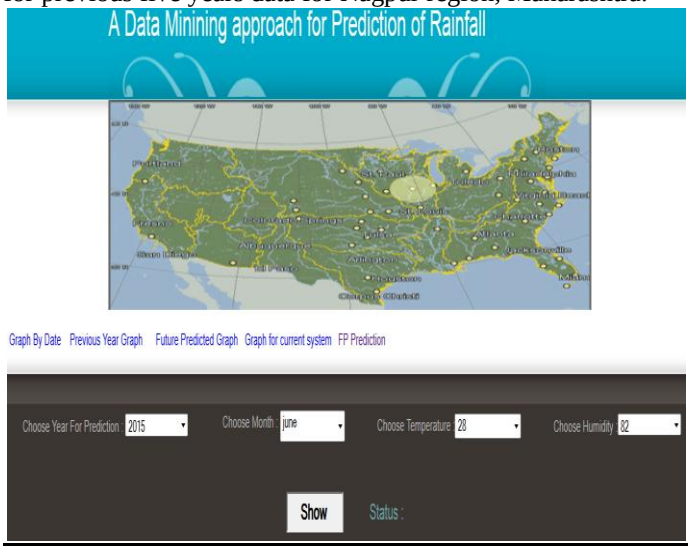

Figure 6.1.2: Completion of field of the data

- First we choose the year for the prediction. It includes future year of the data 2015, 2016 and 2017.
- Second select the month of year of 2015. Consider the select of June month of 2015 for the region of Nagpur.
- Then after select the respective temperature and humidity that are mainly responsible for rainfall prediction.
- After completion of every field of the data then Calculate MAE, MSE and SD for accuracy of the result.

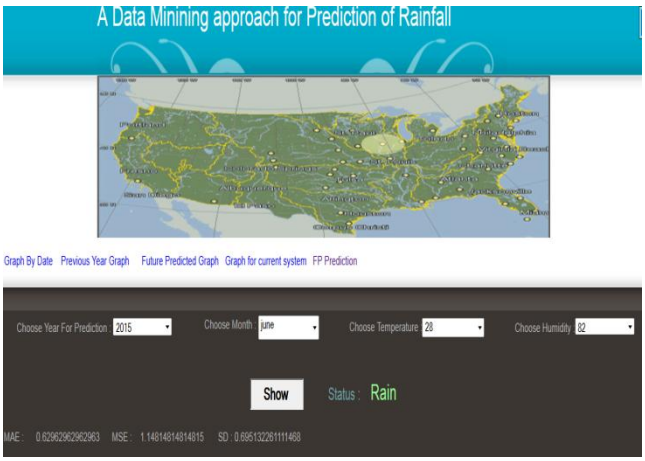

Figure 6.1.3: Output of the system

- And finally shows the status or output of the whole system.

6. Conclusion

In this work the FP Growth Algorithm was used to generate decision trees and rules for classifying weather parameters such as maximum temperature, minimum temperature, rainfall, humidity and wind speed in terms of the month and year. The input data used in this project ate collected from weather department in between 2010 to 2014 for Nagpur station. On the available data sets apply FP Growth algorithm with the evaluation of MAE, MSE and SD. These calculations are accurate more than the existing model Neural Network (NN). It requires the output sensor s like radar, tipping bucket and etc. With FP Growth Algorithm shows correct monthly rainfall prediction than Neural Network. This work is important to climatic change studies because the variation in weather conditions in term of temperature, rainfall and wind speed can be studied using these data mining techniques.

References

[1] Carlos Domenech and Tobias Wehr," Use of Artificial Neural Networks to Retrieve TOA SW Radiative Fluxes for the Earth CARE Mission" IEEE trans on geo science and remote sensing,VOL.49,NO.6,pp.1841-1843 ,JUNE 2011.
[2] Andrew Kusiak," Modelling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach" " IEEE trans on geo science and remote sensing,VOL.51,NO.4,pp.2337-2339,APRIL 2013.
[3] Stephen Dunne and Bidisha Ghosh," Weather Adaptive Traffic Prediction Using Neuro wavelet Models" , IEEE trans on intelligent transportation system, VOL. 14, NO. 1,pp.370, MARCH 2013.
[4] Kit Yan Chan,'' Neural-Network-Based Models for Short-Term Traffic Flow Forecasting Using a Hybrid Exponential Smoothing and Leven berg–Marquardt Algorithm" , IEEE trans on intelligent transportation system, VOL. 13, NO. 2, pp.644-646,JUNE 2012.
[5] Rongrui Xiao, Member, IEEE, and V. Chandrashekar, Member, IEEE," Development of a Neural Network Based Algorithm for Rainfall Estimation from Radar Observations" " IEEE trans on geo science and remote sensing,VOL.35,NO.1,pp.160,JANUARY 2007
[6] Lorenzo Luini and Carlo Capsoni," A Unified Model for the Prediction of Spatial and Temporal Rainfall Rate Statistics" IEEE trans on 2013.
[7] Jarno Mielikainen, Bormin Huang, Hung-Lung Allen Huang, and Mitchell D. Goldberg," GPU Acceleration of the Updated Goddard Shortwave Radiation Scheme in the Weather Research and Forecasting (WRF) Model" IEEE journal, VOL. 5,NO. 2,APRIL 2012.
[8] G. Peter Zhang and Douglas M. Kline," Quarterly Time-Series Forecasting With Neural Networks" IEEE transaction on neural networks VOL.18,NO.18,NOVEMBER 2007.
[9] Yang Hong, Robert F. Adler, and George Huffman," An Experimental Global Prediction System for Rainfall-Triggered Landslides Using Satellite Remote Sensing and Geospatial Datasets" IEEE trans on intelligent transportation system, VOL. 45, NO. 6, JUNE 2007.
[10] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," IEEE Trans. Geosci. Remote Sens., vol. 46, no. 11,pp. 3804–3814, Nov. 2008.