An Integrated Approach for Weather Forecasting based on Data Mining and Forecasting Analysis

G.Vamsi Krishna
Research scholar, Department of CSE,
GITAM University

ABSTRACT

Weather prediction is a real time challenging issue witnessed by the world in the last decade. The prediction is becoming more complex due to the ever changing weather conditions. Many models have been discussed for predicting the weather data assuming the related attributes as independent variables. For effective analysis of the weather, it is necessary to understand various influencing factors that cause the weather changes. It is therefore necessary to identify the relationship between these attributes for better understanding of the weather data. In this article, a weather prediction model based on the spatial and temporal dependencies among the climatic variables together with forecasting analysis.

Keywords

Weather data, forecasting, spatial data, dependent and independent variables.

1. INTRODUCTION

Weather forecasting is considered as the most challenging problem witnessed by the world in the last decade. This indirectly had an impact on effective prediction of the weather data. Due to the latest technological updates, the capabilities of retrieving and storing has increased; resulting in the availability of massive meteorology data in different formats. This data is generated both from the surface observation stations and aerial study stations. With the increase in the number of weather stations, huge amount of data is available on daily, weekly, monthly and yearly basis and the data is stored exponentially [1]. This data is stored and is made available for effective analysis of weather prediction, catastrophe forecasting and for the usage by other departments. In the last decade, with the advancements in science and technology, both empirical approaches and dynamical approaches were developed for the prediction of weather. In these models, the analysis of weather data is carried out using the time series analysis by considering few variables, called attributes for the evaluation of the data, neglecting its importance. Most of the meteorologists have made significant strides in forecasting the weather using models based on time series. However, to analyze the related data from this massive data, mining techniques play a vital role. To have an effective prediction; it is needed to identify the correlation between the attributes of weather, which indirectly have a role in the weather changes. Hence, in this article a model is proposed for effective weather prediction by considering various attributes together with their correlations together with data mining techniques.

2. RELATED WORK

A good amount of literature is witnessed basing on Neural Networks approach [2], [3], [4], [5], [6]. However these approaches failed to identify the abnormal patterns of the weather. Models based on, Support Vector Machine and Stochastic approaches are also projected in the literature [7], [8]. Genetic Algorithm based approaches together with Neural Networks are also projected. [9], [10], [11]. Models based on statistical approach combining Genetic Algorithms and ARIMA models are also showcased in the literature [12],[13]. In the models suggested by the researchers[2],[3],[4],[5],[6],[7],[8] effective forecasting analysis could not be achieved due to the complex data systems of the weather categorical and continuous patterns of the weather, noisy data and high dimensionality of data. Therefore effective models for predicting the weather are to be analyzed.

In this paper a weather forecasting model is built for predicting the weather effectively. In this model the weather data in the form of time series is considered and this data is converted into information where effective knowledge is derived by using the concepts of data mining techniques. Data mining techniques are used to unearth the hidden patterns and associate a linkage between the various attributes associated with the weather conditions In reality, the meteorological data exhibits a series over a period of time and hence, weather prediction can be well analyzed by using time series mining. This article is presented by using the auto regressive integrated moving average (ARIMA) model to forecast the future value. In this model, initially, a mathematical model is generated by considering ordered group of data and then, the prediction is carried out by utilizing the model using the current values and the previous data. Auto correlation models and partial auto correlation models are also considered for performing the interventional analysis. The rest of the paper is organized as follows section-3 of the paper presents an insight about weather forecasting using time series and auto regression models. Section-4 of the paper deals with a brief methodology of the proposed model. Section-5 of the paper deals with the results derived together with the conclusions.

3. TIME SERIES AND AUTO CORRELATION MODELS

The time series analysis is a methodology for building effective models by using the values of the variables which are placed at regular intervals [8]. The study of time series data helps to understand the hidden patterns of the data and helps in better analysis by fitting a model for effective forecasting. Time series Models and forecasting methods are classified into two groups, Univariate Model based and multivariate based. The way the observations are placed differentiates the models. A general supposition in the case of time series methods is the assumption of the data as stationary. Several models such as least square methodology and linear regression are some of the methods used in the time series analysis. Here the dependent and independent variables are considered as numeric data and treated as time series. An autoregressive model generates a linear regression series based on the current value and the previous available information.

The ARIMA process analyze and forecasts uniformly spaced univariate time series data, transfer function data, and intercession data using the Autoregressive Integrated Moving-Average (ARIMA) or autoregressive moving-average (ARMA) model. An ARIMA model predicts a value in a response time series as a linear combination of its own past values, past errors and current and past values of other time series. It is fitted using a random walk model, and the equation for the model is given by

$$\hat{Y}(t) - Y(t-1) = \pi \qquad (1)$$

where π is the mean of the first difference, rearranging equation(1), we have

$$\hat{Y}(t) = Y(t-1) + \pi \qquad (2)$$

The prediction process is carried out by summing the last period's value with a constant, this indirectly help to estimate the prediction changes on an average at particular intervals of time.

ARIMA (p,q,r) models, are used to identify the various seasonal changes , in general, ARIMA(0,1,0) mode, is used to estimate the non-seasonal difference and a constant term. In this paper we have used ARIMA(1,1,0) model is used, since it helps for better prediction of weather based on the autocorrelation of previous data ie lag 1 and the general equation used for fitting the model is presented in equation(3)

$$\widehat{Y}(t) = \pi + Y(t-1) + \phi(Y(t-1) - Y(t-2)) \qquad (3)$$

4. METHODOLOGY

In order to demonstrate the proposed model a data base is generated from the meteorological department of India pertaining to Visakhapatnam district. This weather data set includes several attributes such as minimum temperature, maximum temperature, wind pressure, humidity, perception, sunshine, evaporation and category. The category attribute decides the intensity and is categorized as normal, cloudy, depression, severe depression and cyclonic storm. This categorization is based on one of the attributes causing the changes in the weather and wind pressure. The relationship between various attributes of the weather data is considered and their association is generated. The relationship among these associations helps in effective analysis of the weather. If the weather changes are not understood, several impacts such as coastal erosion, agricultural and human health, damage infrastructure, agriculture and land will be at stake. Therefore in this article the hidden associations among the attributes are considered based on the Time series model. The initial estimates are identified based on the forecasting model called auto correlation and the intermediate weather changes are estimated using moving averages i.e. by using partial auto correlation. The data set available from the Indian Meteorology department is used for the analysis of the model, from http://imdtvm.gov.in

The brief procedure is presented below:

1. Preprocess the data to remove missing values.
2. Calculate the regression values and auto regression values using ARIMA model.
3. Consider different time lags to model the data.
4. Using the correlation analysis, correlate the data and rank the data according to highest correlation.
5. The data with highest correlation is considered to be most likely weather change and it is assumed to be producing destructive effects.

5. RESULTS AND CONCLUSIONS

In this paper the weather data is considered with attributes, such as wind pressure, humidity, Minimum and Maximum Temperature, Forecast and Type, of Visakhapatnam city for a period of 97days. The forecasting experiment is carried out to evaluate, the weather condition for the next 15 days by enabling the ARIMA model prediction algorithm model to predict the forecasts. Initially the ARIMA (1, 1,0), model is considered .These two models are used to predict wind pressure and humidity for the next 15days, the comparison between predicted results and real data is shown in Figure 1 and Figure 2

Table-1 Sample Input Data with Weather parameter

Day	Tmin	Tmax	W	F	Wind	H	T	C
1	14	31	3	28	10	33	1	td
2	15	31	1	28	12	38	1	td
3	16	30	1	28	11	43	1	td
4	15	31	1	28	11	42	1	td
5	15	31	1	28	10	40	1	td
6	16	31	1	28	8	44	1	td
7	17	31	1	28	8	48	1	td
8	22	31	1	34	8	53	1	td
9	22	32	1	33	4	47	1	td
10	22	32	1	34	9	51	1	td
11	21	32	1	33	8	45	1	td
12	21	31	1	33	10	51	1	td
13	21	32	1	32	13	40	1	td

14	21	32	1	32	6	42	1	td
15	21	33	1	34	7	43	1	td
16	1	10	2	5	240	5	2	five
17	1	10	2	9	50	15	2	zero
18	1	10	2	9	55	15	2	zero
19	1	10	2	8	60	10	2	one
20	1	10	2	8	80	10	3	one
21	1	10	2	8	90	9	3	two
22	1	10	2	8	90	9	3	two
23	1	10	2	8	100	8	4	three
24	1	10	2	8	110	8	4	three
25	1	10	2	6	125	8	5	four
26	1	10	2	6	135	7	5	four
27	1	10	2	6	145	7	5	four
28	1	10	2	6	155	7	5	four
29	31	70	1	62	12.7	33	1	td
30	32	70	1	46	13.8	38	1	td
31	33	70	1	72	10.8	43	1	td

Figure 1. Time series Data of the predicted weather

Instinctive investigations show that with the increase of the prediction step of length two sequences the calculated effect is getting inferior. The mean absolute percentage error (MAPE) and mean absolute error (MAE) are used for the analysis and the results derived are presented in table -3

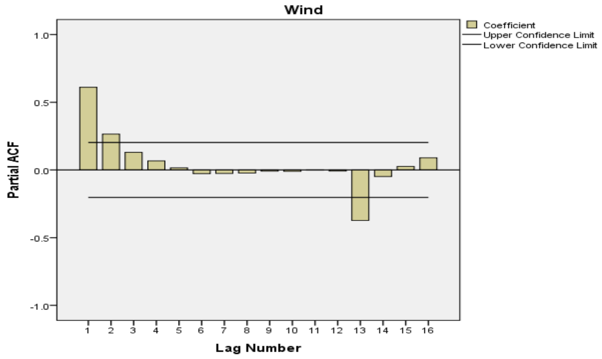

Figure- 2. Autoregressive Lag

Table-2 Paritial Autocorrelations for Series

Lag	Partial Autocorrelation	Std. Error
1	.610	.102
2	.264	.102
3	.130	.102
4	.067	.102
5	.016	.102
6	-.027	.102
7	-.026	.102
8	-.023	.102
9	-.009	.102
10	-.010	.102
11	-.002	.102
12	-.008	.102
13	-.373	.102
14	-.049	.102
15	.026	.102
16	.090	.102

Table 3. Error Analysis Table of Wind Pressure

STEP	MAE	MAPE
1	0.21	0.062
2	0.39	0.067
3	0.65	0.107
4	0.87	0.119
5	0.88	0.130
6	1.00	0.145
7	1.09	0.176
8	1.03	0.177
9	1.12	0.181
10	1.29	0.173
11	1.16	0.165
12	1.21	0.208
13	1.30	0.213
14	1.23	0.209
15	1.21	0.212
16	1.20	0.021

As shown in the MAE column and MAPE column, as the prediction step increases, the prediction error of humidity and the prediction error of wind increases.

In this paper, a methodology for weather forecasting is presented using the data mining prediction algorithm-ARIMA The proposal has the capability analyzing and weather forecasting

6. REFERENCES

[1] Y. W. Dou, L. Lu, X. Liu and Daiping Zhang, “Meteorological Data Storage and Management System”, Computer Systems & Applications, vol. 20, no. 7, (2011) July, pp. 116-120.

[2] C. Zhang, W.-B. Chen, X. Chen, R. Tiwari, L. Yang and G. Warner, “A Multimodal Data Mining Framework for Revealing Common Sources of Spam Images”, Journal of multimedia, vol. 4, no. 5, (2009) October, pp. 313-320.

[3] C. Li, M. Zhang, C. Xing and J. Hu, “Survey and Review on Key Technologies of Column Oriented Database Systems”, Computer Science, vol. 37, no. 12, (2011) February, pp. 1-8.

[4] M. Zhang, “Application of Data Mining Technology in Digital Library”, Journal of Computers, vol. 6, no. 4, (2011) April, pp. 761-768.

[5] C.-W. Shen, H.-C. Lee, C.-C. Chou and C.-C. Cheng, “Data Mining the Data Processing Technologies for Inventory Management”, Journal of Computers, vol. 6, no. 4, April (2011), pp. 784-791.

[6] Z. Danping and D. Jin, “The Data Mining of the Human Resources Data Warehouse in University Based on Association Rule”, Journal of Computers, vol. 6, no. 1, (2011) January, pp. 139-146.

[7] J. Jiang, B. Guo, W. Mo and K. Fan, “Block-Based Parallel Intra Prediction Scheme for HEVC”, Journal of Multimedia, vol. 7, no. 4, (2012) August, pp. 289-294.

[8] S.-Y. Yang, C.-M. Chao, P.-Z. Chen and C.-Hao, “SunIncremental Mining of Closed Sequential Patterns in Multiple Data Streams”, Journal of Networks, vol. 6, no. 5, (2011) May, pp. 728-735.

[9] Z. Fu, J. Bai and Q. Wang, “A Novel Dynamic Bandwidth Allocation Algorithm with Correction-based the Multiple Traffic Prediction in EPON”, Journal of Networks, vol. 7, no. 10, (2012) October, pp. 1554-1560.

[10] Z. Qiu, Z.-W. Lin and Y. Ma, “Research of Hadoop-based data flow management system”, The Journal of China Universities of Posts and Telecommunications, vol. 18, (2011) February, pp. 164-168.

[11] J. Cui, T. S. Li and H. X. Lan, “Design and Development of the Mass Data Storage Platform Based on Hadoop”, Journal of Computer Research and Development, vol. 49, no. 12, (2012) May, pp. 12-18.

[12] P. Sethia and K. Karlapalem, “A multi-agent simulation framework on small Hadoop cluster”, Engineering Applications of Artificial Intelligence, vol. 24, no. 7, (2011) May, pp. 1120-1127.

[13] H. Yu, J. Wen, H. Wang and L. Jun, “An Improved Apriori Algorithm Based on the Boolean Matrix and Hadoop”, Procedia Engineering, vol. 15, (2011) July, pp. 1827-1831.

[14] B. Dong, Q. Zheng and F. Tian, “Optimized approach for storing and accessing small files on cloud storage”, Journal of Network and Computer Applications, vol. 35, no. 6, (2012) May, pp. 1847-1862.

[15] G. Mao, “Theory and Algorithm of Data Mining”, Beijing: Tsinghua University Press, (2007), pp. 121-142.

[16] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh and D. A. Wallach, “Bigtable: A distributed storage system for structured data. Proc.of the 7th USENIX Symp.on Operating Systems Design and Implementation, (2006), pp. 205-218.

[17] S. Ghemawat, H. Gobioff and S.-T. Leung, “The Google File System”, Proc. of the 19th ACM Symp on Operating Systems Principles, (2003), pp. 29-43.