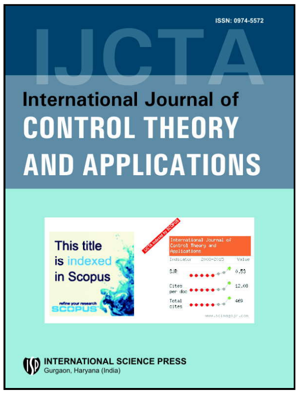

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 14 • 2017

Analysis of Weather Datasets Using Data Mining Techniques

M. Sri Saranya[1] and S. Vigneshwari[1]

[1] School of Computing Sathyabama University, Chennai, TamilNadu, India, Email: srisaranya.m@gmail.com
[2] School of Computing Sathyabama University, Chennai, TamilNadu, India, Email: vikiraju@gmail.com

Abstract: Weather forecasting is an important application in meteorology and also a challenging problem in science and technology around the world in the last century. Prediction of weather by analyzing meteorological data is very useful for people in various fields. Here, we use data mining techniques on weather data collected between year 2010 and 2014 in the Centre for Earth and Atmospheric Sciences department in Sathyabama University from the city of Chennai, India. CRISP DM framework is used in this analysis. After initial data understanding, data preparation included data cleanup and transformations are performed. Descriptive and exploratory data analysis is conducted to identify the maximum and minimum temperature per day and identification of the existence/nonexistence of correlation between the attributes that determine the weather. Clustering techniques are used to group/classify the attributes of weather. Time series analysis is also conducted to forecast the temperature for future dates by building neural network modes. The model is evaluated using a portion of the data and compared based on different evaluation metrics.

Index Terms/ keywords: chennai weather, Data mining, weather forecasting, clustering

1. INTRODUCTION

Weather prediction is an important application to predict the state of the atmosphere of a given location. Weather prediction has a wide usage in many fields like disaster management, military, agriculture, construction etc. As farming depends on various factors like type of soil, water and climate, farmers can be helped to choose their crops based on the seasons by monitoring the seasonal changes in weather [13]. Since weather condition in a day changes continuously, predicting the weather accurately is a cumbersome task for meteorologists. Climate is the long term effect but weather is a short term effect. Weather can be simply defined as a day-by-day variations, whereas climate is the long term mixture of those variations. Instruments like thermometers, barometers can be used to measure weather but the study of climate depends on statistics [6].

Weather prediction describes the present state of the atmosphere that changes. Atmospheric conditions are obtained by Satellite, Observation from Ships, Aircrafts and Ground. This information are collected and sent to meteorological centers for further process. Analyzing the meteorological data is essential for applications like rainfall, temperature, energy application studies and cloud conditions. Several weather prediction methods are used for analyzing the weather data [12].

Data mining is a technique used to find the hidden information from large amount of structured and unstructured data. Data mining is also defined as the process of extracting useful information from databases. Data mining is seen as an important tool for searching interesting patterns and transforms data into business intelligence. Because users may not have any idea regarding their patterns in data and they search for several kinds of patterns. Data mining techniques are used to specify the useful patterns from data and also it focuses the search for interesting patterns. Data mining techniques are also very popular because they are more flexible and efficient for analysis than the traditional methods. Data mining tasks can be categorized into types namely descriptive, exploratory, inferential and predictive. First, describes about the general properties of the existing data and the last, describes based on inference on available data. Data mining is considered as a step by step process of "Knowledge Discovery from Databases". There are several steps like Data collection, Preprocessing known as Data cleaning, Data Selection, Data Transformation.

1.1. Clustering

Clustering can be simply defined as similar items are grouped together into a single unit. Clustering comes under the unsupervised classification of data. The items or attributes that has same characteristics are belong to one cluster and items with different characteristics belong to another cluster. Clustering is useful for analyzing a large dataset if we do not have any prior knowledge about the partition of data.

1.2. Time Series Modelling

Time Series Modelling is a statistical technique which is used to predict the future values of an attribute. Here the independent variable is the time and the dependent variable is the attribute under question.

2. STATE OF ART

In 2015 the statement for weather analysis and forecasting was published by American Meteorological Society (AMS). This statement describes the current state and importance of weather forecasting. Forecasting plays a vital role in providing products and services to community. And also save life and property in disaster management. Based on weather about 90% of the emergencies are declared by the Federal Emergency Management Agency. Also more than 7000 road death happens is attributed to weather [2].

There are several data mining techniques used to predict the weather condition in an effective way. In 2012, Folorunsho Olaiya used Artificial Neural Network and Decision Tree Algorithm to investigate rainfall, maximum temperature and wind speed. The results are compared with actual weather data and rules are formed for classifying weather parameters [6]. In 2014, Divya Chauhan and Jawahar Thakur provides a study of decision tree, k- means clustering, k-Nearest Neighbour which describes that for higher accuracy in weather prediction, decision tree and k-means clustering techniques yields good results than traditional approaches [5]. Data mining techniques has wide applications in different fields. In 2014, Rajini kanth applied data mining techniques in agriculture field. In agriculture, farming depends mainly on weather conditions. Weather condition consists of various seasons around the world and so it is necessary to predict the weather very often. For this it requires machine learning algorithms like J48 classification along with linear regression analysis to provide a better result for predicting weather [13].

3. MATERIALS AND METHOD

3.1. Data Collection

The data used for this work was collected from the Centre for Earth and Atmospheric Sciences department, Sathyabama University. The data covered the period of 56 Months from the year 01-2010 to 08-2014. The data includes the measurements of temperature (°C), relative humidity (%) and wind speed (km/hr) collected at a height /altitude of 2m, 8m, 16m and 50m above the ground level. The data contains the above weather attributes collected at an interval of 10 minutes.

3.2. Data Cleaning

Data cleaning (Pre-processing) is an important and critical step in Data mining process. Raw data that includes noise, incomplete and inconsistent values which affect the accuracy of data in analysis is need to be pre-processed to improve the quality of data and also to achieve better results. The Null values are checked and removed from the attributes. Outliers are imputed and duplicate records are also removed. Finally, the cleaned data is transformed for further data mining process.

3.3. Data Selection

In this step, relevant data that are suitable for analysis was decided/finalized/obtained and retrieved from the dataset.The weather dataset records/contains the temperature, wind speed and humidity at four different levels of altitude recorded for every ten minutes. The weather dataset has 15 attributes such as date, time, temperature (TEMP) measured in degree Celsius, relative humidity (RH) measured in percentage and wind speed (WS) measured in km/hr at different altitudes 2, 8, 16 & 50 meters. Rainfall attribute in the data set has high percentage of missing values. So it is not used in the analysis.

3.4. Data Transformation

Data is aggregated by day and month in order to obtain daily and monthly readings. The attributes are scaled for normalization whose range is between 0.0 and 1.0.

3.5. Preliminary Data Analysis

Table 1
Preliminary Exploratory Data Analysis

	Daily Maximum	*Daily Minimum*	*Correlation Temperature*	*Correlation Relative Humidity*	*Correlation Wind Speed*
Temperature	36.65	21.05	X	-0.6877891	0.3382945
Relative Humidity	115.2	34.4	-0.6877891	X	-0.3792466
Wind Speed	3.9	0.3	0.3382945	-0.3792466	X

The Table 1 shows the preliminary results of exploratory data analysis. The table contains the results of daily maximum and minimum of temperature, relative humidity and wind speed. The relationship between the attributes is determined by correlation coefficient. As expected, temperature and relative humidity are negatively correlated and temperature and wind speed are positively correlated.

The trend of the weather attributes over the years 2010 through 2014 is also presented here in the Figure 1.

4. EVALUATION METRICS

1. Mean Squared Error: Mean Squared Error is the most commonly used measure for numeric prediction. It measures the average of the squares of the differences between the predicted and the actual values. MSE is a good measure of performance.

2. Error Percentage: Error percentage is defined as the measure of the difference between the approximate values and exact values, as a percentage of the exact value

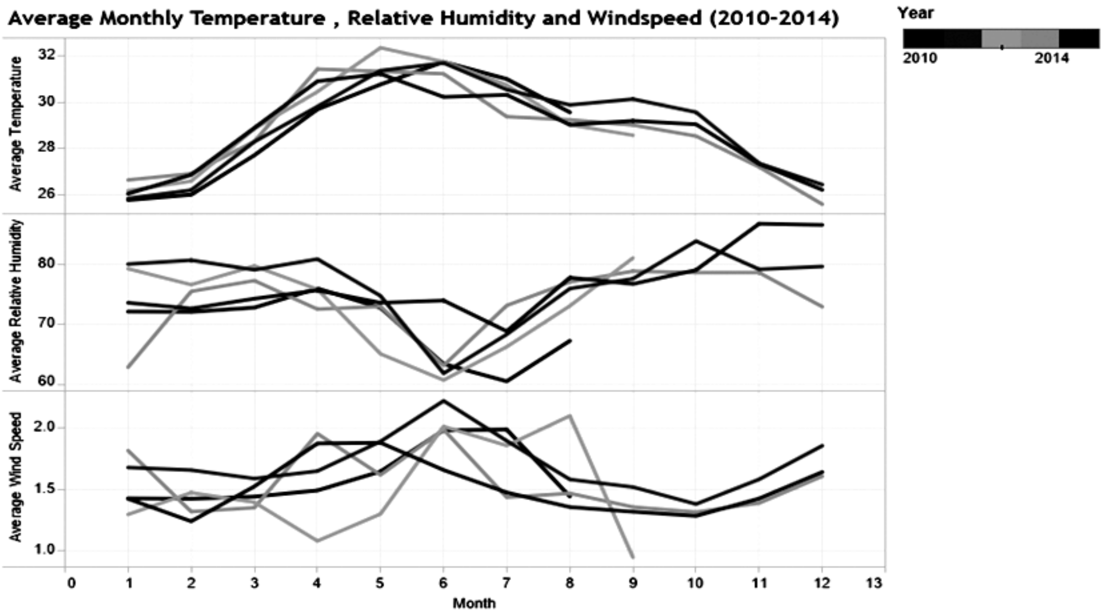

Figure 1: Trend of weather attributes

3. Mean Absolute prediction Error (MAPE): It is a measure of prediction accuracy in statistics for forecasting model that measures the size of the error in percentage.

4. R^2: R Square measures the fraction of the total sum of the squared distance between the data points that can be explained by the sum of squared distance between the clusters. The higher this fraction results the better clustering.

5. EXPERIMENTAL DESIGN

R is a language and environment for statistical and graphical techniques, and is highly extensible. R studio v 3.3.1 is used to analyze the meteorological data.

Algorithm for classifying the weather data

1. Read the dataset
2. Preprocessing: clean , remove outliers , scale data
3. Cluster the data with average temperature, average humidity and average wind speed
4. Find the accuracy
5. Visualize the results

Algorithm for Forecast using Neural Networks

1. Split data into training and test set
2. Convert the training data into time series
3. Smoothen the time series using Sliding Window/ Holt winters function
4. Create time series model using nnetar
5. Forecast / Predict the future values using the model
6. Report the accuracy of the model by comparing with test data.

6. RESULTS AND ANALYSIS

Results of clustering on average temperature, humidity and wind speed show that the weather attributes can be grouped into three groups. Comparing against the original data shows that all months of the year has days falling into these 3 clusters. The results are represented in Figure 2. The low inter cluster distance between the clusters can be attributed to the presence of outliers in the data.

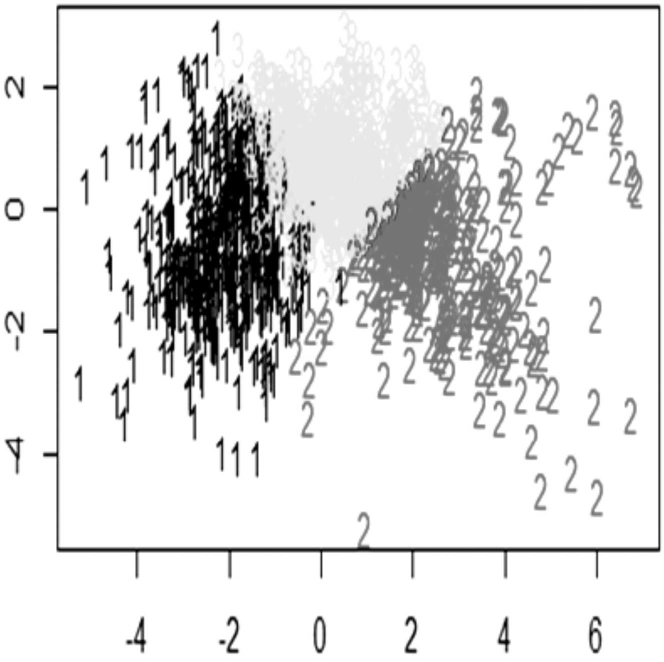

Figure 2: THW cluster

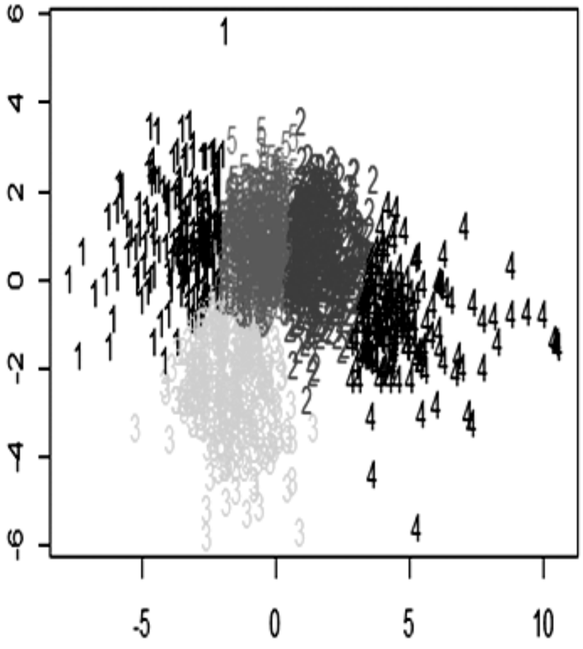

Figure 3: TH cluster

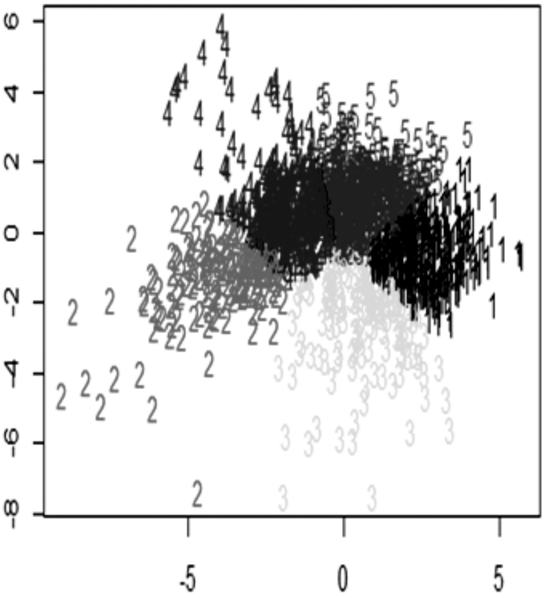

Figure 4: TW cluster

Results of clustering on average temperature & humidity show that the weather attributes can be grouped into five groups. Comparing against the original data shows that all months of the year has days falling into these 5 clusters. The results are represented in Figure 3. The ratio of the within cluster sum of square to the total sum of squares is 80%

Results of clustering on average temperature & wind speed show that the weather attributes can be grouped into five groups. Comparing against the original data shows that all months of the year has days falling into these 5 clusters. The results are represented in Figure 4. The ratio of the within cluster sum of square to the total sum of squares is 74%.

The Figure 5(a) shows the original time series followed by the forecasted values of maximum temperature using neural network model.

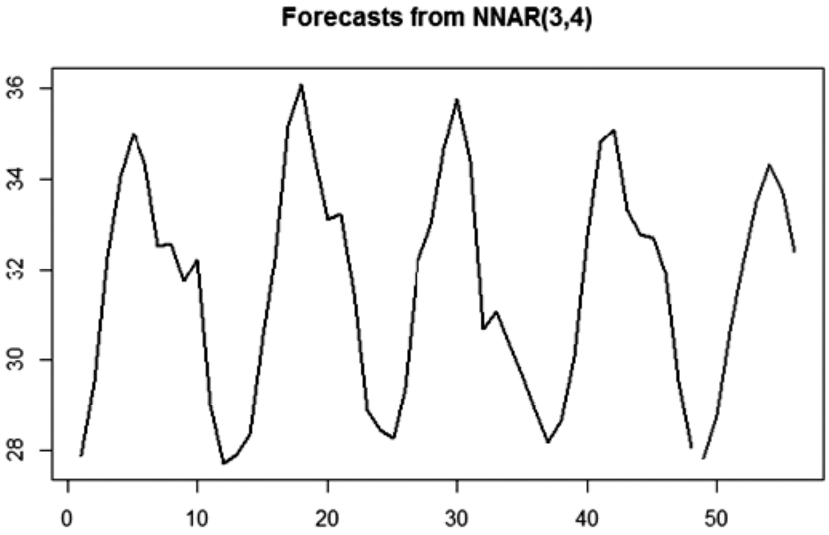

Figure 5 (a): NNETAR: Original Vs Predicted

The Figure 6(a) shows the original time series followed by the forecasted values of minimum temperature using neural network model.

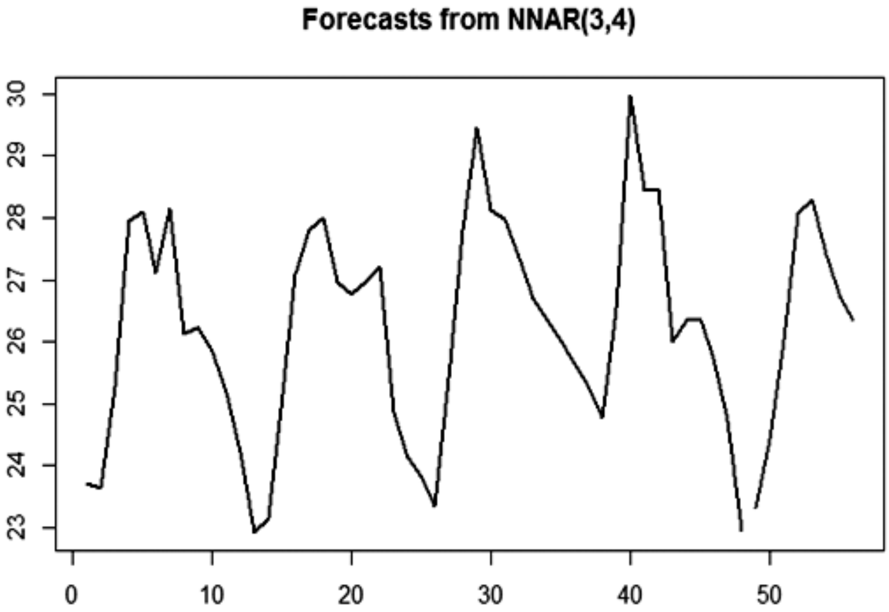

Figure 6: (a) NNETAR: Original Vs Predicted

The figure 7(a) shows the original time series followed by the forecasted values of maximum relative humidity using neural network model.

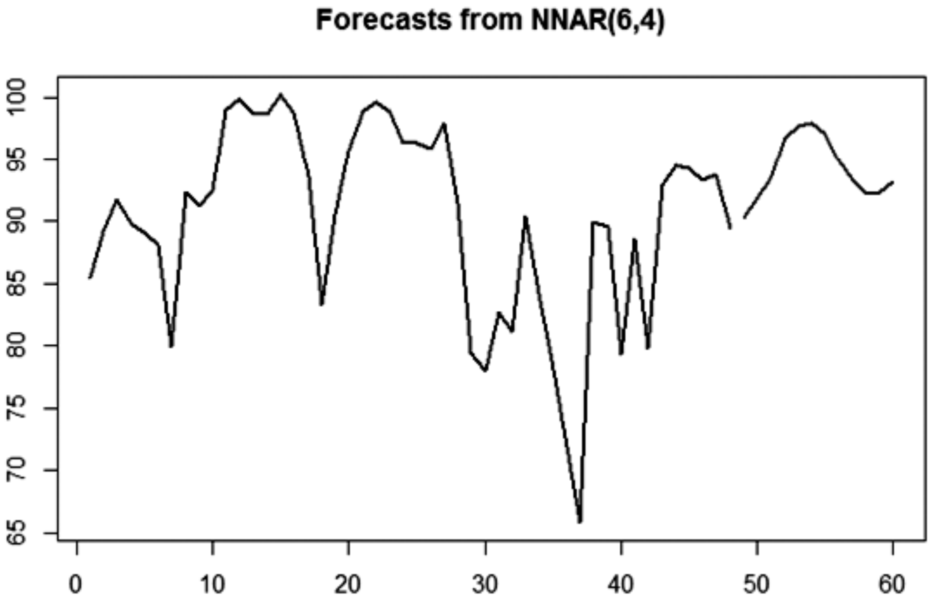

Figure 7: (a) NNETAR: Original Vs Predicted

The figure 8(a) shows the original time series followed by the forecasted values of minimum relative humidity using neural network model.

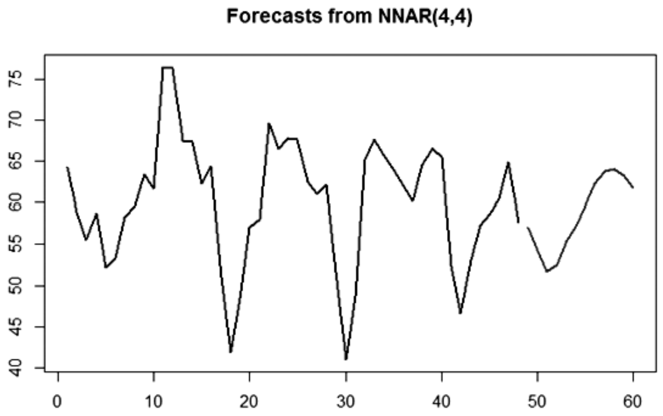

Figure 8: (a) NNETAR: Original Vs Predicted

The figure 9(a) shows the original time series followed by the forecasted values of maximum wind speed using neural network model.

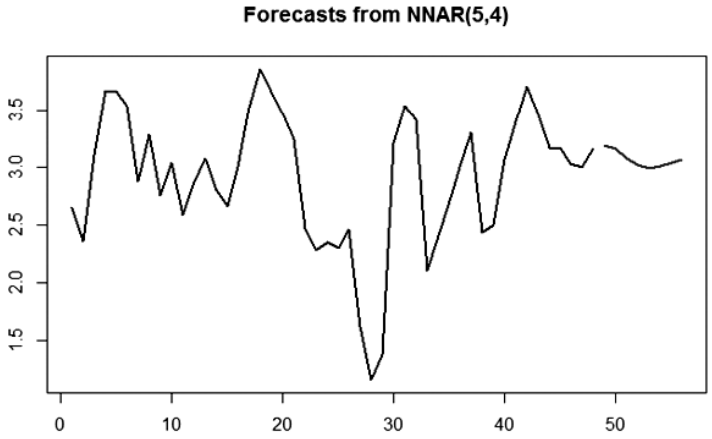

Figure 9 (a) NNETAR: Original Vs Predicted

The figure 10(a) shows the original time series followed by the forecasted values of minimum wind speed using neural network model.

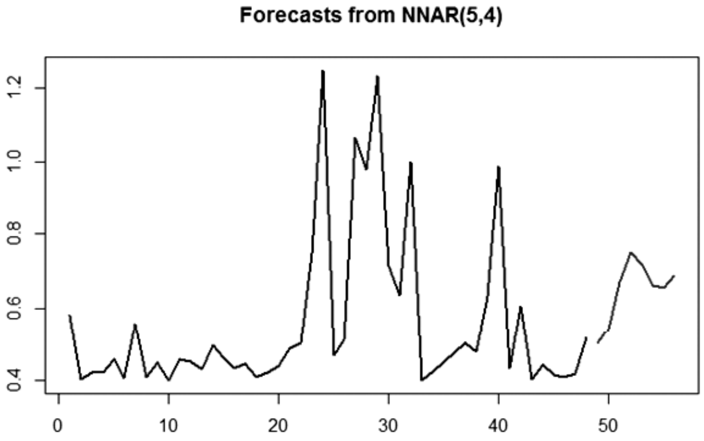

Figure 10 (a) NNETAR: Original Vs Predicted

The below table 2 shows the evaluation metrics for maximum and minimum values of temperature, relative humidity and wind speed modelled by using neural networks. Among the above, the model for forecasting the maximum and minimum temperature is acceptable and the model for forecasting maximum and minimum humidity is relatively acceptable. The model for wind speed is rejected due to very high MAPE value.

Table 2
NNETAR Training and Test data Statistics

	Temperature				*Relative Humidity*				*Wind Speed*			
	Maximum		*Minimum*		*Maximum*		*Minimum*		*Maximum*		*Minimum*	
	Training Data	*Test Data*	*Training Data*	*Test Data*	*Training Data*	*Test Data*	*Training Data*	*Test Data*	*Training Data*	*Test Data*	*Training Data*	*Test Data*
ME	0.0350	0.7644	-0.0078	-0.674	0.1975	-7.930	0.4024	-6.103	0.0258	0.0885	0.0199	-0.232
RMSE	0.8794	0.9945	1.1438	1.1065	4.8211	8.5922	4.8550	11.969	0.3320	0.4572	0.1620	0.232
MAPE	1.9875	2.3044	3.8830	2.8164	4.2796	9.2022	6.0590	21.139	10.2681	12.687	17.464	56.620

7. CONCLUSION AND FUTURE WORKS

This paper presents the results of data mining conducted on the weather data collected in Centre for Earth and Atmospheric Sciences in chennai city following the CRISP Data mining framework. The initial exploratory data analysis brings out the daily maximum, minimum temperature, the correlation between the weather attributes, the average monthly temperature and the trend of the monthly temperature is identified. K means clustering algorithm is used to cluster all the weather attributes into 3 clusters and the months having weather conditions falling into these clusters are identified. The temperature & humidity is clustered into 5 clusters and the months having weather conditions falling into these clusters are identified. Similarly, the temperature & wind speed is clustered into 5 clusters and the months having weather conditions falling into these clusters are identified. Neural Network model is used to forecast the weather attributes for future months. The evaluation criteria are also compared based on the model. This paper deals with weather attributes measured at 2m from the ground. One potential future work can be analysis of the apparent weather experienced at different altitudes above the ground, using the weather attributes collected at 8m, 16m and 50m above the ground level. Forecasting can be conducted using other forecasting models and accuracy can be improved.

ACKNOWLEDGMENT

The authors would like to acknowledge that this work has been carried out at DST-FIST sponsored Cloud Computing Lab (order Saction No. : **SR/FST/ETI-364/2014** Dated: **21 November, 2014**), School of Computing, Sathyabama University.

REFERENCES

[1] Ahrens, C.D., 2007, “Meteorology” Microsoft Student 2008, Redmond, WA; Microsoft Corporation, 2007.

[2] American Meteorological Society, “Weather Analysis and Forecasting” https://www.ametsoc.org

[3] Auroop R Ganguly, and Karsten Teinhaeuser, “Data Mining for Climate Change and Impacts”, IEEE International Conference on Data Mining, 2008.

[4] Chaudhari A.R, Rana D.P, Mehta R.G, “Data Mining with Meteorological Data” International Journal of Advanced Computer Research, Volume-3, Number-3, Issue-11, September-2013

[5] Divya Chauhan, “Data Mining Techniques for Weather Prediction: A Review”, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 8

[6] Folorunsho Olaiya, "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", I.J. Information Engineering and Electronic Business, 2012, 1, 51-59

[7] Indian Meteorological Department, http://www.gov.in

[8] Kavita Pabreja, "Clustering technique to interpret Numerical Weather Prediction output products for forecast of Cloudburst" International Journal of Computer Science and Information Technologies, Vol. 3 (1) , 2012, 2996 – 2999

[9] Mark G Lawrence, "The Relationship between Relative Humidity and the Dew point Temperature in Moist Air", American Meteorological Society, 2005, 225-233.

[10] S. Liao, P. Chu, P. Hsiao, "Data Mining Techniques and Applications- A Decade Review from 2000 to 2011", Expert Systems with Applications, Vol. 39 (12), pp 11303-11311, 2012

[11] Sarah N. Kohail, Alaa M. El-Hales, "Implementation of Data Mining Techniques for Meteorological Data Analysis", IJCSIT Journal Volume 1 No. 3, July 2011

[12] S. Kotsiantis "Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values", World Science, Engineering and Technology, 450-454, 2007

[13] Rajini kanth T.V, Balaram V.V SSS and Rajasekhar N, "Analysis Of Indian Weather Data Sets Using Data Mining Techniques" Computer Science & Information Technology (CS & IT), pp. 89–94, 2014. © CS & IT-CSCP 2014.

[14] University of Alberta, Osmar R. Zaiane, "ChapterI: Introduction to Data Mining ", CMPUT690 Principles of Knowledge Discovery in Databases, 1990.

[15] Vamsi Krishna G, "A Review of Weather Forecasting Models-Based on Data Mining and Artificial Neural Networks", IJCSC, Vol 6 Number 2 April - Sep 2015 pp. 214-222.