ANALYSIS OF CHENNAI WEATHER DATA SETS USING DATA MINING TECHNIQUES

[1]G. KOTEESWARI, [2]Dr. M. PUSHPA,
[1] M. Phil., Research Scholar, [2] Assistant Professor
[1] PG & Research Department of Computer Science,
[1] Quaid -E- Millath Government College for Women, Chennai, India

Abstract : Weather forecasting prevent from natural disasters such as tornadoes, floods, storms. A reliable weather forecast can help both the farm and non- farm sectors. Accurate weather prediction is one of the most challenging problems around the globe. This paper makes use of classification and clustering technique to predict weather for a month in a particular region with past meteorological data set, collected between 2014 and 2017 for Chennai region of Tamil Nadu state in India. We applied Multiple Linear Regression (MLR), K-Mean and K- Nearest Neighbor (KNN) data mining techniques for weather prediction on the available dataset. Based on the experiment result KNN proved to be good with higher accuracy prediction technique for temperature and humidity than other mining techniques.

***Keywords*- Meteorological data mining, K-Nearest Neighbor, Multiple Linear Regression, K- Mean Clustering.**

I. INTRODUCTION

Weather prediction is one of the most crucial aspects around the world. The prediction of weather conditions can have significant impacts on various sectors of society in different parts of the country. They are used by government and industry to protect life, property and also to improve the efficiency of operations by individuals to plan a wide range of daily activities. The notable improvement in forecast accuracy has been achieved since 1950s that is a direct outgrowth of technological developments (Allan H. Murphy 1997). The advance knowledge of weather parameters in a particular region is advantageous in effective planning. Several studies on forecasting weather variables based on time series data in reference to a particular region have been carried out at national and international level in both the farm and non- farm sectors. It has been one of the most interesting and fascinating domain. The scientists have been trying to forecast meteorological characteristics using a large set of methods, some of them more accurate than others. The weather dataset like humidity, Pressure, Wind Speed, Wind Direction, Visibility, Temperature, Dew Point, Sunshine, Rainfall, Clouds Quality, Snow depth, and so on were observed by radiosondes that are launched all over the world approximately and those information's were transmitted to the ground station. Also, the surface weather measurements are made at observing stations around the world, from ships and buoys at sea, commercial aircraft, weather radars and satellites. All these measurements are transmitted to different metrological centers. These centers have very fast supercomputers they are programmed with equations to describe the atmosphere changes at every point.

Meteorological data mining is a form of Data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved can be transformed into usable knowledge. Useful knowledge can play important role in understanding the climate variability and climate prediction. The main aim of this paper is to have an overview the Data mining Process on weather data and to analyze data mining technique like clustering, Regression and classification. To study the use of data mining techniques in forecasting maximum temperature, Minimum Temperature, Average Temperature, Maximum Humidity, Minimum Humidity Average Humidity, Pressure, Dew point and wind speed. This was carried out using K-Means, K-Nearest Neighbor and Multiple Linear Regression algorithms.

II. LITERATURE REVIEW

Guhathakurata [2006] - Weather is a continuous, data-intensive, multi-dimensional, dynamic and chaotic process, and these properties make weather forecasting a formidable challenge. It is one of the most imperative and demanding operational responsibilities carried out by meteorological services all over the world. At present, the assessment of the nature and causes of seasonal climate variability is still conception.

Sivakumar et al. [1999] - The field of meteorology all decisions are to be taken in the visage of uncertainty associated with local of and global climatic variables. Several authors have discussed the vagueness associated with the weather systems. Chaotic features associated with the atmospheric phenomena also have attracted the attention of the modern scientists.

Dilip C and Dr. K Thippeswamy [2016] - In the first step generating local clusters on individual nodes will be done and in the second step local clusters are aggregated to form a global module. From several clusters some of the clusters acts as leaders, these leaders will do merging of local clusters into global one using overlay technique. This technique is continued until a resultant cluster is obtained. Distributed dynamic clustering deals with very large scale, distributed and heterogeneous datasets. The communication overhead is minimized by reducing the size of the dataset which is going to exchange between the systems. By using K-means algorithm local clusters are generated and analyzed. During aggregation the local clusters are merged and produce an ultimate accurate output.

Pinky Saikia Dutta, Hitesh Tabilder [2014] - Data mining techniques is used to predict the monthly rainfall of Assam. This carried out using traditional statistical technique Multiple Linear Regression. Regression model which contain more than two predictor variables are called Multiple Linear Regression. The period of 2007-2012 data collected from regional meteorological centre Guwahati. The model consider maximum temperature, minimum temperature, wind speed, mean sea level as predictors 63% accuracy in validation of rainfall for proposed model. The model can predict the monthly rainfall.

A.R.W.M.M.S.C.B. Amarakoon [2010] - Proposed a system that uses the authentic weather data and applies the data-mining calculation "K-Nearest Neighbor (KNN)" for grouping of these chronicled data into a particular time traverse. The 'k' closest time ranges is then additionally taken to anticipate the weather of Sri Lanka. It creates exact outcomes inside a sensible time for a considerable length of time ahead of time. It is inferred that KNN is valuable to dynamic data, the data that progressions or updates quickly and gives better execution when contrasted with alternate procedures. Coordinating component choice strategies can even give more precise outcomes.

III. METEOROLOGICAL DATA MINING

Meteorological data mining is a type of mining which is concerned with finding hidden patterns inside massive data available. So, the information extracted can be transformed into practical knowledge (**A. Kalyankar, Prof. S. J. Alaspurkar 2013**). The knowledge plays a vital role to predict the future of the weather. Having Knowledge of meteorological data is the key for variety of application to perform analysis and prediction of weather condition and it also does good prediction of temperature, humidity and irrigation system. These databases can become valuable information for analysts, as well to perform different operations on this data. It requires higher scientific techniques like machine learning application for effective study and prediction of weather condition. Weather can be predicted with the help of various metrological parameters by using various techniques of data mining. While some of these algorithms are more exact prediction than others.

IV. PRE-PROCESSING ON METEOROLOGICAL DATA

Moxon1996- "Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. Data mining is a knowledge discovery process of extracting previously unknown, actionable information from very large databases".

A. Data

Data collection is the systematic approach of gathering and measuring information from a variety of sources. Data collection enables us to evaluate the outcomes and make predictions about the future. Data's are one of most important aspect for this analysis. The Maximum temperature, Minimum temperature, Average Temperature, Dew point, pressure, Maximum humidity, Minimum humidity and Average humidity dataset for the period of January 2014 to December 2017 were downloaded from the web address https://www.wunderground.com/history/daily/VOMM/2014/1/1/DailyHistory.html. The Water underground website maintains the historical data. The downloaded data are between Elev **52**ft **13** °N, **80.18** °E that roughly covers the Chennai city of Tamil Nadu state in India.

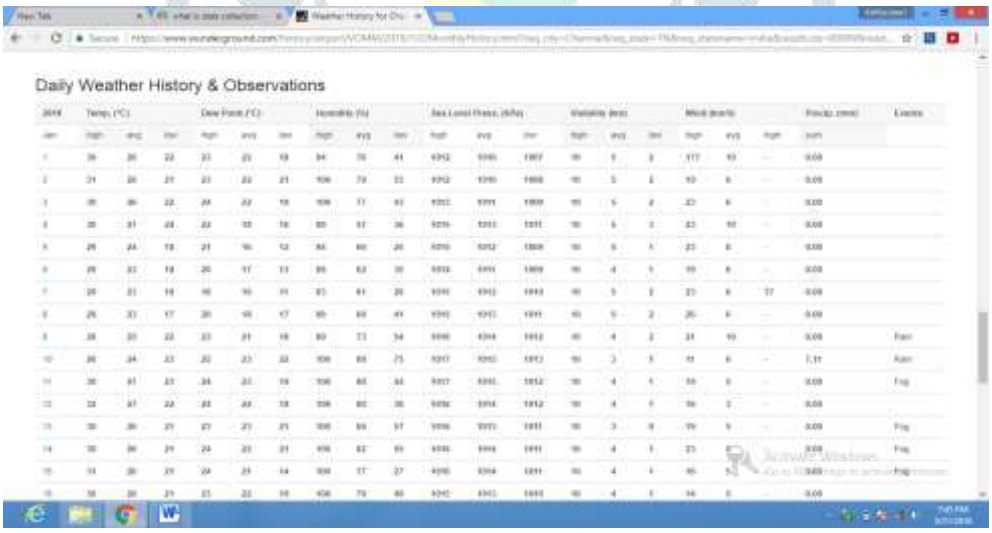

Fig. 1 sample weather dataset

B. Data Pre- Processing

The data pre-processing involves transforming raw data into an understandable format. In the metrological weather data, various parameters like Gust, Wind, Visibility, precipitation, Events, Temperature, Dew point, humidity and etc., relevant data may not be recorded due to misunderstanding or because of equipment faulty. In this analysis pre-processing means, removing unwanted parameters from the dataset and to remove noisy data. The metrological data goes through a series of steps during the preprocessing such as,

1. Data Cleaning

In this phase noisy and irrelevant data's are removed from the database. Data cleaning routines attempt to fill in missing values, smooth out noisy data while identifying outliers and correct inconsistencies in the data. It's used for to improving the data quality.

2. Data Transformation

Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system. The collected weather parameters are usually in Excel format it's converted into comma separated values (CSV) file.

As the analysis uses the weka tool with the file formats .csv, for the maximum temperature, Minimum temperature, and average temperature. The data's are represented in a monthly format like January 2014, 2015, 2016 and 2017 data were stored in one a file. The same method is repeated for humidity, temperature, pressure, and dew point also.

3. Discretization

Data Discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Concept hierarchies can be used to reduce the data by collecting and replacing low-level with high-level concepts.

C. Knowledge Discovery

Weka tool is used for knowledge extraction using various data mining techniques such as Clustering, Classification and Regression.

D. Result of analysis

The future value of temperature and humidity were predicted depending on the result of K- Mean, K- Nearest Neighbor and Multiple Linear Regression algorithms.

V. DATA MINING TECHNIQUES

Data mining is a crucial analytic process indeed to explore data, the most imperative task in data mining is to extract non-trivial nuggets from vast amount of data. The data mining techniques are set of algorithm intended to find the hidden knowledge from the data. In the following discussion we apply various data mining techniques in the metrological dataset that contains details of various parameters about the weather. Temperature and humidity parameters were used for analysis of weather condition. There are several major data mining techniques are Clustering, Classification, Outlier analysis, Association and Correlation. This work carried out the analysis using the three major data mining algorithms like K-Mean, KNN and MLR

A. Clustering

Clustering is an unsupervised learning algorithm and it deals with finding a *structure* in a collection of unlabeled data. "The process of organizing data points into groups that is similar in some way". The similar data points are one cluster and dissimilar data points are another cluster as in Fig. 2.

Fig. 2 clustering

Centroid based clustering

Centroid based clustering is also known as K- Means clustering. Clustering can uncover previously undetected relationships in a dataset. There are many applications for cluster analysis and an important issue in k- means clustering to determine the similarity between two objects, so that clusters can be formed from objects with high similarity between clusters. Commonly, to measure similarity or dissimilarity objects, a distance measured by Euclidean Distance.

Algorithm

The k-means algorithm is used for partitioning the dataset into different clusters, each cluster's center is represented by the mean value of the objects in the cluster. Steps incorporated in the k- means clustering are as follows,

STEP 1: Pick 'K' random points as cluster centers called centroids

STEP 2: Assign each x_i to nearest cluster by calculating its distance to each centroid

STEP 3: Find new cluster center by taking the average of the assigned points

STEP 4: Repeat Step 2 and 3 until none of the cluster assignments change

Input:

k: the number of clusters,

D: a data set containing n objects.

Output: A set of k clusters.

B. Classification

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical class labels as in Fig. 3.

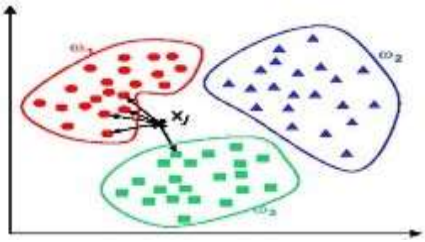

Fig. 3 Classification

K- Nearest Neighbor

Nearest-neighbor classifiers based on Lazy Learning function, a lazy learner simply stores and waits until it is given a test tuple. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. When given an unknown tuple, a k-nearest neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k "nearest neighbors" of the unknown tuple. "Closeness" is defined in terms of a distance metric, such as Euclidean distance. K- Nearest Neighbor algorithms steps are as shown below,

STEP 1: Determine parameter K= number of nearest neighbors
STEP 2: Calculate the distance between the query instance and all the training samples
STEP 3: Sort the distance and determine nearest neighbors based on the K-th minimum distance
STEP 4: Gather the category y of the nearest neighbors
STEP 5: Use simple majority of the category of nearest neighbors as the prediction value of the query instance.

C. Regression

Regression is a data mining function that predicts a number like Profit, sales, mortgage rates, house values, square footage, temperature, or distance could all be predicted using regression techniques. For example, a regression model that predicts temperature values could be developed based on observed data for temperature over a period of time.

Multiple Linear Regression

Regression model which contain more than two predictor variables are called Multiple Regression Model. Equation of the Multiple linear regression model is,

$$Y=b_0+b_1X_1+b_2X_2+b_3X_3+b_4X_4+...e \quad \text{.... (1)}$$

Where,

Y is Average temperature, Dependent variable
b_0, b_1, b_2, b_3, b_4 are Regression Coefficient
X_1, X_2, X_3, X_4 Max, Min Temperature are predictor or regressor or explanatory variables

Multiple linear regression fits a model to predict a dependent (Y) variable from two or more independent (X) variables. The predictors can be understood as independent variables and the target as a dependent variable. The error, also called the residual, is the difference between the expected and predicted value of the dependent variable. The regression parameters are also called as regression coefficients. The work is analyzing weather dataset using different data mining technique and working of the design as shown in Fig. 5.

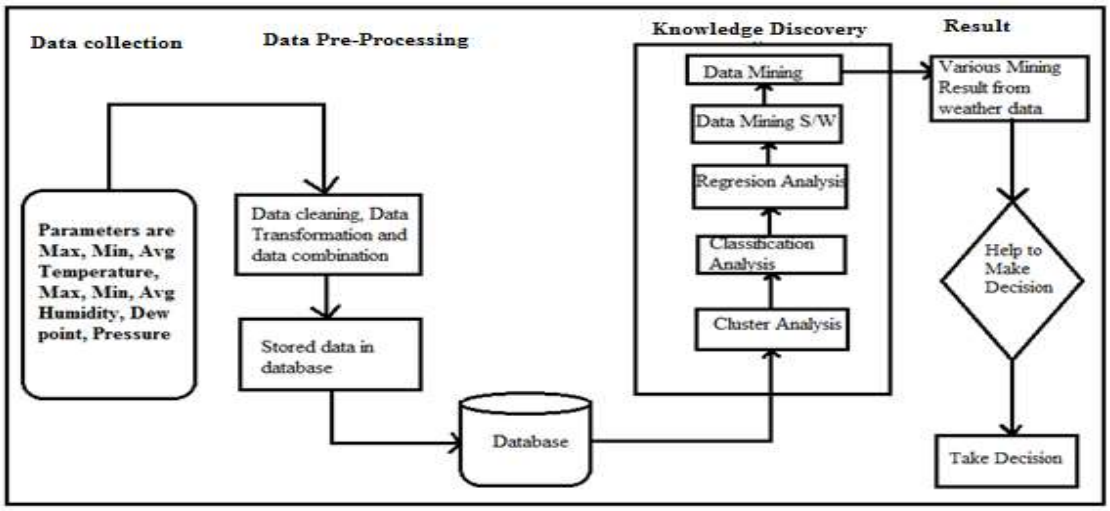

Fig. 5 Weather Data Analysis System

VI. ANALYSIS AND RESULT

The Weather parameters max, Min, Avg Humidity, Max, Min, Avg Temperature, Pressure and Dew point are fed into K-Means, K- Nearest Neighbor and Multiple Linear Regression algorithms using weka tool. The January temperature data file fed in k- Means algorithm.

Results

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     temp
Instances:    124
Attributes:   3
              High, avg, low
Test mode:    evaluate on training data

=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 10.295937100893997

Initial starting points (random):
Cluster 0: 31,26,21
Cluster 1: 30,25,20
Missing values globally replaced with mean/mode
Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1
             (124.0)    (66.0)     (58.0)
=============================================
high          29.8226    30.2424    29.3448
avg           25.3145    26.2576    24.2414
low           20.6129    21.5758    19.5172
Time taken to build model (full training data): 0.01 seconds
=== Model and evaluation on training set ===
Clustered Instances

0       66 ( 53%)
1       58 ( 47%)
```

The KNN and MLR accuracy were calculated by using the formula and temperature result for KNN and MLR as in Fig 6. The accuracy was calculated for temperature.

$$\text{Accuracy} = \frac{\text{Number of Correctly Predicted Days}}{\text{Total Number of Days}} \times 100$$

$$\text{Accuracy} = \frac{28}{31} \times 100 = 90.3\% \text{ (approxi)}$$

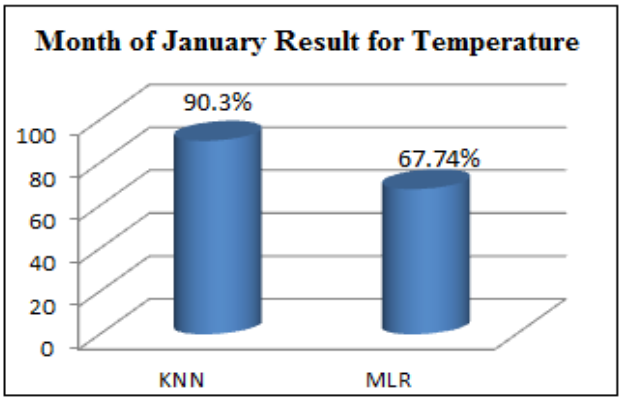

Fig. 6 Result for Temperature

VII. CONCLUSION AND FUTURE WORK

Climate affects the human society in all the possible ways. A reliable weather forecast can help many sectors like Agriculture, Aviation, Broadcast Media, Insurance, Media & Entertainment, Construction industry and so on. Analysis on weather data describes the use of data mining technique and the uses of historical data to predict the weather in a particular region or city. We use of Classification and clustering technique to predict weather for a month in a particular region with historical metrological data set. After applying clustering and classification for weather prediction KNN is found to be the most feasible technique compared to the other two data mining techniques. In future dynamic data mining methods can be used to predict nature, rapid changes and sudden events with dynamical data set. We can enlarge the database with other important attributes.

REFERENCES

1. Allan H. Murphy (1997), "The Early History of Probability Forecasts: Some Extension and Clarifications", American Meteorological Society, Vol- 13.
2. Amruta A. Taksande, P. S. Mohod (June 2015), "Applications of Data Mining in Weather Forecasting Using Frequent Pattern Growth Algorithm", International Journal of Science and Research (IJSR), Vol- 4, Issue- 6.
3. A.R.W.M.M.S.C.B. Amarakoon (2010), "Effectiveness of Using Data Mining for Predicting Climate Change in Sri Lanka".
4. Dilip c, Dr. K Thippeswamy (May 2016), "A Data Mining Concept for Weather Forecasting Using Clustering Algorithm", International Journal of Computer Engineering and Applications, Vol- X, ISSN 2321-3469.
5. Elia Georgiana Petre (2009), "A Decision Tree for Weather Prediction", Bluetinul, Vol. LXI, No.1, pp: 77-82.
6. Folorunsho Olaiya, Adesesan Barnabas Adeymo (February 2012), "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", International Journal Information Engineering and Electronic Business, Vol-1, pp : 51-59.
7. Guhathakurta P (2006), "Long range monsoon rainfall prediction of 2005 for the districts and sub-division kerala with artificial neural network". Current science, Vol. 90, pp. 773- 779.
8. Kavita Pabreja (2012), "Clustering technique to interpret Numerical Weather Prediction output products for forecast of cloudburst", International Journal of Computer Science and Information Technology (IJCSIT), Vol-3, pp: 2996-2999.
9. Meghali A. Kalyankar, Prof. S. J. Alaspurkar (February 2013), "Data Mining Techniques to Analyse the Metrological Data", International Journal of Advanced Research in Computer Science and Software Engineering, Vol-3, ISSN: 2277-128X, Pg. No: 114-117.
10. M. Kannan , S. Prabhakaran, P. Ramachandran (2010), "Rainfall Forecasting Using Data Mining Technique", International Journal of Engineering and Technology(IJET), Vol-2, pp: 397-401.
11. M Ramzan Talib, Toseef Ullah, et al. (June 2017), "Application of Data Mining Techniques in Weather Data Analysis", International Journal of Computer Science and Network Security (IJCSNS), Vol-17, No- 6.
12. Pinky Saikia Dutta, Hitesh Tahbilder (May 2014), "Prediction of Rainfall Using DataMining Technique over Assam", Indian Journal of Computer Science and Engineering (IJCSE), Vol-5, No. 2, ISSN: 0976-5166, Pg.no:85-90.
13. P.Kalaiselvi (August 2016), "Weather Forecasting A Survey", International Journal of Modern Computer Science (IJMCS), Vol-4, Issue- 4, ISSN: 2320-7686, pp: 113- 116.
14. R.Samya, R.Rathipriya (September 2016), "Predictive Analysis for Weather Prediction Using Data Mining with ANN: A Study", International Journal of Computational Intelligence and Informatics, Vol-6, No-2, ISSN: 2349-6363, pp: 150-154.
15. Sagar S. Badhiye, Nilesh U. Sambhe, P. N Chatur (January 2013), "KNN Technique for Analysis and Prediction of Temperature and Humidity Data", International Journal of Computer Applications (IJCA), Vol- 61, No: 14, ISSN: 0975 -8887.
16. Sivakumar, B., Liong, S.Y., Liow, C. Y. and Phoon, K.K. (1999), "Singapore rainfall behavior Chaotic?", Journal of Hydrology Engineering, ASCE, Vol- 4, Pg. No: 38-48.
17. T V Rajini kanth, V V SSS Balaram , N Rajasekhar (2014), "Analysis of Indian Weather Data Sets Using Data Mining Techniques", Dhinaharan Nagamalai et al.(Eds): Computer Science & Information Technology (CS & IT), pp: 89 -94.
18. U. Fayyad, G. Piatetsky-Shapiro, and P. Smith (1996), "Data Mining to Knowledge Discovery: An Overview", In U. Fayyad, G. Piatetsky-Shapiro, P. Smith and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, pages 1-34. MIT Press, Cambridge, MA.