International Journal of Computer Applications Technology and Research
Volume 6–Issue 12, 473-479, 2017, ISSN:-2319–8656

Big Data Approach and Using Data Mining Techniques in Weather Prediction

M Ramesh
IT Analyst
TCS
Hyderabad, India

S Swarajhyam
Assistant Professor
MLRIT
Hyderabad, India

B Prathyush
Sr. Enterprise application Engineer
GE India Exports PVT Limited
Hyderabad, India

Abstract: Weather analysis has been playing its vital role in meteorology and become one of the most challengeable problems both scientifically and technologically all over the world from the last century. This study carries historical weather data collected locally at Faisalabad city, Pakistan that was analyzed for useful knowledge by applying data mining techniques. Data includes ten years' period [2007-2016]. It had been tried to extract useful practical knowledge of weather data on monthly based historical analysis. Analysis and investigation was done using data mining techniques by examining changing patterns of weather parameters which includes maximum temperature, minimum temperature, wind speed and rainfall. After preprocessing of data and outlier analysis, K-means clustering algorithm and Decision Tree algorithm were applied. Two clusters were generated by using K-means Clustering algorithm with lowest and highest of mean parameters. Whereas in decision tree algorithm, a model was developed for modeling meteorological data and it was used to train an algorithm known as the classifier. 10-fold cross validation used to generate trees. The result obtained with smallest error (33%) was selected on test data set. While for the number of rules generated of the given tree was selected with minimum error of 25%. The results showed that for the given enough set data, these techniques can be used for weather analysis and climate change studies.

Keywords: Data Mining, K Mean Clustering, Decision Trees, Weather Data Analysis

1. INTRODUCTION

In present era weather forecasting and analysis has become a challenging problem around the world from the last century. The reason behind are the two main factors: Firstly, it is useful for many human activities like agriculture sector, tourism and natural disaster prevention. Secondly, due to various technological advances like the growth of computational power and ongoing improvements in measuring systems.

All over the world, major challenges faced by meteorologist are the accuracy of weather analysis and its prediction. On the other hand, researchers had tried to predict different meteorological parameters by utilizing different data mining techniques. While some of these techniques are more precise than others. Over the past few decades the availability of climate data has been increased. Such sources of climate data like observational records, understudy data, etc. makes it more important to find tools with higher accuracy rate to analyze different patterns from massive data. Therefore, meteorological data mining is a form of mining which is concerned with finding hidden patterns inside massive data available. So, the information extracted can be transformed into practical knowledge. This knowledge plays a vital role to understand the climate change and prediction. Having Knowledge of meteorological data is the key for variety of application to perform analysis and prediction of rainfall and it also does good job for prediction of temperature, humidity and irrigation system.

In this research, we have gathered useful knowledge on historical weather data that was collected locally at Faisalabad city. The data comprise ten year of period. While the records obtained include maximum temperature, minimum temperature, wind speed and rainfall observation. After data pre-processing we applied the outlier analysis, clustering algorithm and classification techniques. After utilizing these techniques and algorithm we have represented and described the importance of meteorological field by extracted knowledge.

Data mining objectives is to provide accurate knowledge in the form of useful rules, techniques, visual graphs and models for the weather parameters over the datasets. This knowledge can be used to support the decision- making for various sectors. The goals for data analysis are those which involve weather variations that affect our daily runtime changes in min and max temperature, humidity level, rainfall chances and speed of wind. This knowledge can be utilized to support many important areas which are affected by climate change includes Agriculture, Water Resources, Vegetation and Tourism. Studies show's that human society is affected in different ways by weather affects. For example, water resources are the main sources of irrigation in production of agriculture crops and the amount of rain is one of them that affect the crops abruptly due to climate change. It is also directly related to the different human activities. Moreover, poor growth and low quality is due to negative effects of weather resulting in failure of high production. Hence, changes in weather conditions are risky.

2. Literature Review

Many scholars have made efforts to implement different mining techniques in the areas of meteorological data based on weather data analysis and prediction. Meteorology data mining has been successfully employed in the field of developing important forecasting applications.

M. Viswambari, 2014;surveys the various techniques implemented in data mining to predict weather. Data mining uses various technologies to forecast weather for predict wind pressure, rainfall, humidity, etc. Classification in data mining differentiates the parameters to view the clear information. By looking at the survey provided by Divya Chauhan, 2014; it provides views of different literatures of some algorithms implemented by various researchers to utilize different data mining techniques for Predicting Weather. In this field the work done by different researchers is shown in tabular form where it has been reviewed and compared. Decision tree and k-means clustering algorithm seems to be good at predating weather with higher accuracy than the other techniques of data mining. Some researchers have tried to make the dynamic prediction. The paper of Jyotismita Goswami, 2014; discusses various models for prediction that are applied and compared with their methodologies which are available till now. Their crucial findings marked this study very valuable for a better starting point to generate a new weather prediction model with new description of methodology for predicting weather by using different models of dynamic change in climate.

Sarah N. Kohail, 2011; "tried to extract useful knowledge from daily weather historical data collected locally at Gaza city. All data mining techniques are applied and describe extracted knowledge importance in the meteorological field, used for prediction and decision making". Zahoor Jan1, 2008; developed a system for prediction weather that utilizes the historical data of an area (rainfall, temperature, wind speed etc.) and applied the algorithm of data mining i.e. "K-Nearest Neighbor (KNN)" to classify this historical data within this specific time span. The "K-Nearest Neighbor (KNN)" then uses these time spans to predict the weather accurately. These experiments demonstrate that the system is generating accurate results inside reasonable time frame for months to come. Meghali A. Kalyankar, 2013; where k-means clustering is implemented for predicting the change in climate of a regional area using historical data of the weather.

While following the same agenda, Folorunsho Olaiya, 2012;tried to use different data mining algorithms such as Decision Tree and ANN to predict different weather factors. Since meteorological data are vast and time constrained, it is not only need to modify by traditional data mining but also can be modified using some other techniques. A. R. Chaudhari, 2013;This paper is deliberating the application of various techniques of data mining that are applied in different ways to predict, associate, classify and pattern clustering of meteorological data.

Badhiye S. S., 2012;The main objective of this research was to proposed design of data analysis system regarding temperature and humidity by using an efficient data mining technique KNN to discover the unseen patterns within the huge data set for classification and prediction of climate conditions by transferring the acquired information into practical knowledge. It was able in predicting the values of climate conditions with higher accuracy rate of temperature and humidity factors. In the same situation, S. Kotsiantis, 2006, used both dynamic selection and regression fusion as a hybrid technique and combined the features of both for daily temperature forecast. In adding to this, another research was published in 2007.; to predict daily max, min and average temperatures for city of Patras in Greek by utilizing six dissimilar data mining approaches: "k-Nearest-Neighbor (KNN), Feed-Forward-Back-Propagation (FFBP), M5-rules- algorithm, Decision-tree, instance-based-learning (IB3) and linear-least-squares-regression (LR)". They had used data of four years for period [2002 -2005] of rainfall, relative humidity and, temperature. In this research the obtained results were precise regarding Correlation-Coefficient and Root-Mean-Square.

Pabreja, 2012; checked the happening of cloudburst using relative humidity and temperature and apply K- means clustering technique. It is not good for long term predictions. Agboola A.H., 2013;the present research inspects the rules of fuzzy logic for modeling the rainfall in South West of Nigeria. The created fuzzy rules based model show suppleness and capability of demonstration between input & output variables that uses an ill-defined association.

Juraj Bartoka, 2012; This work defines the prearranged involvement of the project based on DMM of the research on parameterized methods and models for detecting and predicting the significance of meteorological phenomena; particularly low covering of cloud and fogging. This venture was likely to cover the approaches for combining the scattered meteorological data that was essential for running models of prediction, training and then mining of the data in demand for predicting randomly occurring phenomena proficiently and speedily. Adeyemo, 2013;In this research the use of Self-Organizing-Maps, Co-Active-Neuro-Fuzzy- Inference-System soft computing procedures are presented for predicting weather and climate change studies utilizing historical data gathered from Nigeria's city Ibadan between year 1951- 2009. The use of soft computing procedures for knowledge discovery and analysis in weather forecast and studies of climate change can be implemented as per shown by the results. Where, the following study offered applications of ANN & learning models for predicting weather in local south of Saskatchewan in Canada. Imran; presented collaborative model for measuring performance in different circumstances with: "multi-layer-perception-network, Elman-Recurrent-Neural-Network (ERNN), Radial-Basis-Function-Network (RBFN), Hopfield-Model (HFM), Predictive-Models and Regression -Technique". In this research the training and testing of different models was made using data of relative humidity, temperature and wind speed where each model was tested for 24 hrs. ahead, while prediction was carried out for (winter, spring, summer and fall) season.

3. Materials and Methods

3.1 Sample Dataset

In this research article, daily historical weather data for ten years (2007 to 2016) was used in analysis. The data was collected from metrological station located at Faisalabad 33.4^0 North and

73.8^0 East. Following procedure was implemented includes, data cleaning, data selection, conversion of data and data mining.

3.1.1 Data Cleaning

At this phase, a reliable data model was setup for handling missing data, finding and removing duplicate data means misleading data. Finally, the procedure of cleaning takes place which successfully converts data into a suitable form for mining.

3.1.2 Data Assortment

At this phase, analysis of relevant data was decided and retrieved from the dataset. Meteorological data set with attributes, their type and description is presented in Table 1, while analysis of the numeric values is also shown in Table 2.

Table 1: Numeric-Data Values Analysis

Attributes	Types	Description
Years	Numeric	Considered Years
Months	Numeric	Considered Months
Wind-speed	Numeric	Wind as km
Max-Temp	Numeric	Maximum-Temperature
Min-Temp	Numeric	Minimum-Temperature
Rainfall	Numeric	Total-monthly-rainfall

Table 2: Dataset of Meteorological Attributes

Variable	Min.	Max.	Mean	Standard Deviation(SD)
Min Temp	4.42	33.47	20.26	8.633
Max Temp	18.44	49.64	34.24	8.055
Wind Speed	1.27	4.00	2.21	0.555
Rainfall	0	8.04	1.49	1.71
Years	2007	2016	-	-
Months	1 January	12 December		

3.1.3 Data Conversion

That is also known as the data association. This is selected form of data into a suitable data mining stage. Save the data files in comma-separated by value (CSV) format of file and data set was standardized to reduce the data scaling.

3.1.4 Data Mining Phase

This phase has divided into the three more stages. At individually stage, the algorithm for analyzing meteorological data sets is implemented. Then test methods are used in this study which is the percentage split of the data set for training, cross validation and testing of the remaining percentage. Subsequently, this recognizes the knowledge representation of interesting patterns.

4. Methodology

This article was taken different steps, using a different method in each step, with high precision temperature, wind speed and rainfall parameters values of weather data and displays the analytics power of data mining technology point of view in Figure 2.

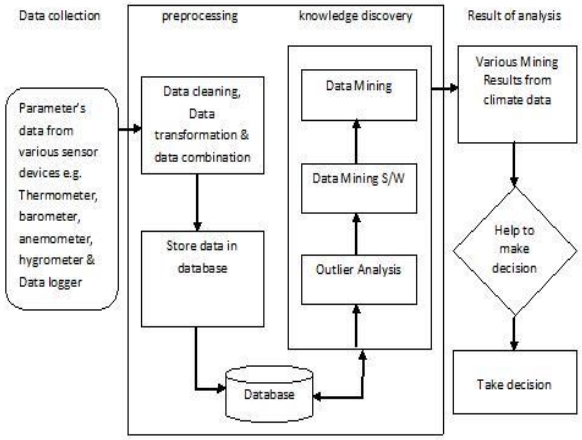

Figure2. Design System of Weather Data Analysis

4.1 Data Collection

Data collection is a main integral part of implementing mining techniques, for this challenge, a thermometer, barometer, hygrometer, anemometer and data recording systems was used. Data recording system provides weather data to excel in tabular form. "Data record based on a digital processor which is used by the built-in sensor or an external instrument and sensors associated with position data of the time of the electronic device can automatically collect and records data of 24 hours. This was the main and most important benefits of data recorder. It was used to collect weather data from local stations at Faisalabad to a devoted lab PC, then copy the transferred weather data to an Excel spreadsheet and recorded on daily basis along with monthly basis to identify data.

4.2 Data Pre-Processing

The data preprocessing is the next step of data mining after collection of data. Challenges in temperature, rainfall and wind speed data; knowledge discovery process is facing poor data quality. Thus, the data is pre-processed to remove noise and unwanted data. Pretreatment means concentrating the removal of other unwanted variables from the data, while the data preprocessing includes these steps:

4.2.1 Data scrubbing: it's the stage where noise and irrelevant data is removed. Data cleaning procedures are implemented to fill out missing values and to eliminate noise in recognizing outliers and to correct data irregularities

4.2.2 Data integration: it's recognized as the data conversion; in this stage, the suitable form of data is converted for the procedure of data mining by reduction of data and construction of attributes.

4.3 Discovery Knowledge: It's recognized as the data conversion; in this stage, the suitable form of data is converted for the procedure of data mining by reduction of data and construction of attributes.

4.4 Analysis of Result: it's recognized as the data conversion; in this stage, the suitable form of data is converted for the procedure of data mining by reduction of data and construction of attributes.

4.5 Proposed approach

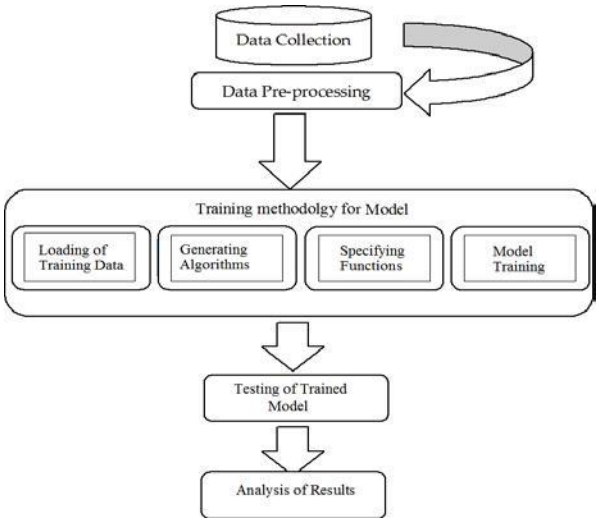

Figure 3: Training methodology of the model

In this paper, different data mining techniques are applied. Firstly, the K- means clustering algorithm was applied on the given data set which was then altered into appropriate form from unstructured data format after the stage of preprocessing. Secondly, the decision tree (J48 algorithm) was applied. Where 70% of data was taken as training data and remaining 30% was testing data. The model of training methodology was shown in figure 3.

5. Results and Discussion

There was used the Weka environment that has 4 applications which are Explorer, Knowledge Flow, Experimenter and Simple CLI. The collected data-set was changed in a file of extension ".arff" and loaded in the environment of Weka. First, the attribute reduction was used for data preprocessing then the simple k-means clustering and j48 algorithm were used.

4.1 K-means clustering

=== Run information ===

Scheme:weka.clusterers.SimpleKMeans"weka.core.Euclid

MeanDistance -R first-last"

Relation: Weather Data

Instances: 132

Attributes: 6

- Year
- Month
- TMAX
- TMIN
- RAIN
- WIND

Test mode: evaluate on training data

== Model and evaluation on training set == K-Means ======

Number of iterations: 8

Within cluster sum of squared errors: 38.6613361873891

Missing values globally replaced with mean/mode

Cluster centroids:

Cluster#

Attribute Full Data 0 1

(132) (64) (68)

==========

Year 10 2007 2016 Month 12 January December TMAX 34.2387 41.051 27.8271 TMIN 20.2652 27.754 13.2169 RAIN 1.4887 2.3028 0.7225 WIND 2.2078 2.5332 1.9016

Clustered Instances

0 64 (48%)

1 68 (52%)

Two clusters were made by applying k-Means clustering algorithm in which the month of May was recorded with extreme

temperature in Faisalabad city. The maximum temperature means getting the maximum value will be high sunshine time or The k-mean clustering algorithm molded two clusters, wherein the highest of mean-temperature and lowest of mean-temperature was estimated in the May & June's month. The representation of these clusters made by algorithm is shown in figure 4

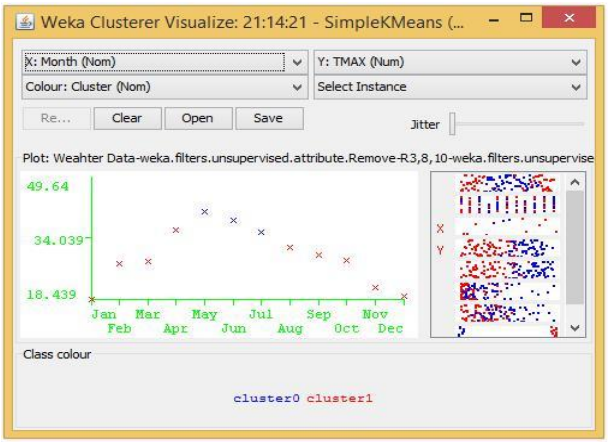

Figure 4: Visualization of simple k-means clusters

The resulting decision tree:
The obtained decision tree result by using j48 algorithm can be offered in comprehend rules which can easily understandable and useful.

The following table is shown the instantaneous runs for generating selected test by using 10fold cross validation on a dataset using j48 rules. Twelve rules are presented but the Run # 8 is selected which had minimum error:

Table 3: Summary of results j48 rule generation process

Run No	No. of Rules Generated	Error
1	16	50%
2	13	58%
3	16	50%
4	14	42%
5	13	58%
6	16	33%
7	17	33%
8	16	25%
9	20	42%
10	15	33%

J48 rule generation

daylight.

J48 Decision Tree:
J48 algorithm used in Weka environment that is newest form of two algorithms (ID3 & C4.5) . J48 standard algorithm for performing the partition has been upgraded over the time and it is totally based on the perception of information-theory. The core idea behind this is to select the appropriate variable that can provide information needed to realize suitable partitioning in individual branch to other branches for classifying training set.

Classifiers Decision Tree algorithm has one big advantage that resultant tree can be generate rules. To help the users, these rules can be written in descriptive form for understanding their data. WEKA environment generates decision tree and rules based upon the selected choices. The 10fold cross validation was used to generate Trees& rules and the results with the smallest error were selected on test dataset. Table 2 provides a summary of runs and decision tree which had the least error obtained from the Run Number 5.

Table 2: Abstract of decision tree results

Run No	No. of Trees Generated	Error
1	19	50%
2	21	58%
3	18	42%
4	15	42%
5	17	33%
6	16	33%
7	21	42%
8	15	58%
9	17	50%
10	18	58%

- Rule-1: Max-Temp< 27.4 →period January
- Rule-2: Wind > 122.66, Min-Temp>=0.3, Max-Temp< 35.8 →period February
- Rule-3: Wind < 160.45, Wind >= 131.45, Min-Temp> 7.6 Max-Temp< 42.4 →period March
- Rule-4: Wind > 141.9, Min-Temp> 13.5 & Max-Temp<=47.7, Rain< 65.6 →period April
- Rule-5: Wind < 199.98, Min-Temp>= 19.1 Max-Temp< 51.6, Rain> 53.2 →period may
- Rule-6: Wind < 252.26, Min-Temp> 27.06, Max-Temp< 49.34 & Max-Temp>= 31 →period June
- Rule-7: Wind < 103.93, Max-Temp<= 44.83 and rain<=249 →period July
- Rule-8: Wind < 89.93, Max-Temp<= 42.95 →period August
- Rule-9: Wind < 77.26, Min-Temp<= 22.86 & Max-Temp<= 42.54 →period September
- Rule-10: Wind <= 107.26, Min-Temp>=17.84 Max-Temp<= 39.12 →period October
- Rule-11: Wind <= 100.45, Min-temp>=13.69, Max-Temp<= 32.02 →period November
- Rule-12: Wind <= 78.45, Max-Temp< 27.38 and Rain <= 25 →period December

Discuss the results

Following rules can be generated by j48 algorithm:

- Rule # 2 infers that the wind-speed of between 2007 – 2016 is more than 122.6 km/day, and the temperature remains between the 0.3 ºC to 35.8 ºC during February.
- Rule # 3 infers that the wind-speed of between 2007 – 2016 varies between 131.45 km/day to 160km/day whereas the temperature remains between 7.5 ºC to 42.4 ºC during March.
- Rule # 4 infers that the wind-speed of between 2007 – 2016 is more than 141.9 km/day and the temperature remains between the 13.5 ºC to 47.7 ºC and the precipitation is less than 65.6 mm during April.
- Rule # 5 infers that the wind-speed of between 2007 – 2016 is less than 199.98 km/day and the temperature is between 19.6ºC to 51.1ºC and the rainfall is less than 53.2 mm during May.
- Rule # 6 infers that the wind-speed of between 2007 – 2016 is less than 252.26 km/day and the temperature remains between the 27.06 ºC to 49.34ºC during June.
- Rule # 7 infers that the wind-speed of between 2007 – 2016 is less than 103.93 km/day and themaximum-temperature is around 44.83ºC and rain is less than 249mm during July.
- Rule # 8 infers that the maximum temperature of between 2007 – 2016 is approximately 42.95ºC during August.
- Rule # 9 infers that the wind-speed of between 2007 – 2016 is less than 77.2 km/day and the minimum-temperature is almost 22.86 ºC and the maximum temperature remains between 42.54 ºC during September
- Rule # 10 infers that the wind-speed of between 2007 – 2016 is around 107.2 km/day and the temperature remains between the 17.84 ºC to 39.12 ºC during October.
- Rule # 11 infers that the wind-speed of between 2007 – 2016 is around 100.4 km/day and the temperature remains between 13.69 ºC to 32.02 ºC during November.
- Rule 12 infers that the wind-speed of between 2007 – 2016 is around 78.45 km/day and the maximum-temperature is less than 27.38ºC and the precipitation is around 25 mm during December.
- Rule # 1 infers that the maximum-temperature of between 2007 – 2016 is less than 27.4 ºC during January.

It was experimental that the highest value of average maximum-temperature in the months between February & April was almost 34 oC and the lowest value of average minimum-temperatures was recorded 22.2 oC in the months between June & September. The peak value of wind-speed for month of June was larger than 150.6 km/day but drop below 118 km/day for the other months. The minimum precipitation was recorded nearby 18 mm for month of December and the maximum precipitation was larger than 33.1 mm in the months of April & May.

6. Conclusions and Recommendations

In this paper, k-means clustering and decision tree building process were implementation; both are the most common data mining techniques tried to highlight the method that the stored data about past measures can be used for the future ones. Here, j48 (decision tree algorithm) was tried to create decision-trees & rules for the classification of parameters of weather such as minimum temperature, maximum temperature, precipitation and wind-speed per months and years. Experimental trends about sufficient data over-time was analyzed and the significant deviations was identified that showing the change in climate patterns. Future work can include expanded database with other important weather parameters and include using this weather information in agriculture sector reform with cutting edge technologies.

7. References

[1] A. R. Chaudhari, D. P. Rana, & R. G. Mehta. (2013). Data Mining with Meteorological Data. International Journal of Advanced Computer Research, Volume 3(Issue 11), Pages 5.Adeyemo, A. (2013). Soft Computing Techniques for Weather and Climate Change Studies. African Journal of Computing & ICT(Volume 6), Pages 14.

[2] Agboola A.H., Gabriel A. J., & Aliyu E.O., Alese B.K. (2013). Development of a Fuzzy Logic Based Rainfall Prediction Model. International Journal of Engineering and Technology, Volume 3, Pages 9.

[3] Badhiye S. S., Wakode B. V., & Chatur P. N. (2012). Analysis of Temperature and Humidity Data for Future value prediction. International Journal of Computer Science and Information Technologies, Volume 3, Pages 3.

[4] Juraj Bartoka, Ondrej Habalab, Peter Bednarc*, Martin Gazaka, & Ladislav Hluchýb. (2012). Data Mining and Integration for Predicting Significant Meteorological Phenomena. Elsevier, Pages 10.

[5] Jyotismita Goswami, & Alok Choudhury. (2014). Dynamic Modeling Technique for Weather Prediction. International Journal of Computer Science & Engineering Technology, Volume 5, Pages 8.

[6] k. somvanshi, & et al. (2006). modeling and prediction of rainfall using artificial neural network and arima techniques". j. ind. geophys. union, vol. 10(no. 2), pp. 141-151.

[7] M. Viswambari, & Dr. R. Anbu Selvi. (2014). Data Mining Techniques to Predict Weather: A Survey. International Journal of Innovative Science, Engineering & Technology, Volume 1(Issue 4), Pages 3.

[8] Meghali A, Kalyankar, & Prof. S. J. Alaspurkar. (2013). Data Mining Technique to Analyse the Metrological Data. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3(Issue 2), Pages 5.

[10] R. Nagalakshmi, M. Usha, & RM. A. N. Ramanathan. (2013). Application of Data Mining Techniques in Maximum Temperature Forecasting: A Comprehensive Literature Review. International Journal of Advance Research in Computer Science and Management Studies, Pages 9.

[11] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, & K. Menagias. (2006). A Hybrid Data Mining Technique for Estimating Mean Daily Temperature Values.

[12] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, & K. Menagias. (2007). Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values. International Journal of Mathematical, Physical and Engineering Sciences, Volume 1, Pages 5.

[13] S. S. Mesakar, & M. S. Chaudhari. (December-2012). Review Paper On Data Clustering Of Categorical Data. International Journal of Engineering Research & Technology, Vol. 1, pp.1-3.

[14] Ibrahim M. El-Hasnony, Hazem M. El-Bakry, Ahmed A. Saleh,"Classification of Breast Cancer Using Soft computing Techniques", International Journal of Electronics and Information Engineering, Vol.4, No.1, Mar 2016.

[15] Ronak Sumbaly N. Vishnusri. S. Jeyalatha ―Diagnosis of Breast Cancer using Decision Tree Data Mining Technique", , International Journal of Computer Applications (0975 – 8887) Volume 98– No.10, July 2014.

[16] Wen-Hsien Ho, King-Teh Lee, Hong-Yaw Chen, Te-Wei Ho, Herng-Chia Chiu,"Artificial Neural Network to explore effecting factors of Hepatic Cancer recurrence". Published January 3, 2012 University of Alberta Edmonton, AB, Canada.

[9] Pabreja, & Kavita. (2012). Clustering technique to interpret Numerical Weather Prediction output products for forecast of Cloudburst. International Journal of Computer Science and Information Technologies, Volume 3, Pages 4.