Survey on Analysis of Meteorological Condition Based on Data Mining Techniques

M. Manikandan[1], Dr. R. Mala[2]

[1]PG and Research Department of Computer Science, Marudupandiyar College of Arts and Science, Vallam, Thanjavur, Tamil Nadu

[2]Research Advisor, Department of Computer Science, Marudupandiyar College of Arts and Science, Vallam, Thanjavur, Tamil Nadu

Abstract— *An application of data mining is a rich focus to Classification algorithm, Association algorithm, Clustering algorithm which can be applied to the field of various resources it concerns with developing methods that discover the knowledge from data origination. In this paper, focuses on meteorological data analysis in form of data mining is concerned to predict the knowledge of weather condition. Rainfall analysis, temperature analysis, based on climatic condition, cyclone form data analysis is vital application role for meteorological analysis in data mining techniques. Prediction, association and forecasting are the several method in data mining used for meteorological analysis. Many countries have already experienced deadly droughts and floods also climate-induced natural disasters have displaced hundreds of thousands of people across the world. Mainly due to over ambitious strategies and actions of human beings on the eco-system, data mining play a significant role in determining the climate trends in crucial manner. In this research work is discussing the application of different data mining techniques applied in several ways to predict or to associate or to classify or to cluster the pattern of meteorological data. It can be provided for future direction for research.*

Keywords— Meterological analysis, Correlation analysis, Forcasting, Clustering Techniques, Classification methods, Association Rule.

I. INTRODUCTION

In this paper represents the analyses of weather and climate events, comes into proper historical perspective, understanding their strangeness, and, increasingly, comparing recent events to expectations of future climate conditions. The climate conditions can monitoring yearly, seasonal, and monthly in regional and sectorial specific climate information, information on disasters and extreme events. Our earth is surrounded by a layer of air called atmosphere. Sometimes air becomes hot and sometimes it becomes cool this change in air is known as weather. The commonly seasons are winter, spring, summer, autumn. The most common data mining technique is used to identify and extract the weather condition based on Regression analysis, Correlation analysis, Artificial Neural Networks, Fuzzy Logic Techniques, Association rule, k-Nearest Neighbor and other classification techniques, multi linear regression analysis and clustering algorithm.

In this paper to focuses an overview of data mining techniques used to analyzing the meteorological data for identifying the weather condition in terms of results show that given enough case data.

II. RELATED WORKS

Data mining techniques are now the important techniques utilized in all application area related to meteorological data for the prediction and decision making by discovering interesting rules or patterns or groups that indicate the relation between variables.

The weather data used for the research include daily temperature, daily pressure and monthly rainfall. Akash D Dubey [20], presented different artificial neural networks have been created for the rainfall prediction of Pondicherry, a coastal region in India. These ANN [2] [14] models were created using three different training algorithms namely, feed-forward back propagation algorithm, layer recurrent algorithm and feed-forward distributed time delay algorithm. The number of neurons for all the models was kept at 20. The mean squared error was measured for each model and the best accuracy was obtained by feed-forward distributed time delay algorithm with MSE value as low as .0083.

Pritpal Singh, et.al.[19], discussed about increased the accuracy of forecasting of Indian summer monsoon rainfall. In this study, they used feed-forward back-propagation neural network algorithm for ISMR forecasting. Based on this algorithm, five neural network architectures designated as BP1, BP2,…BP5 using three layers of neurons (one input layer, one hidden layer and one output layer). The data set is trained and tested separately for each of the neural network architecture, viz., BP1–BP5. The forecasted results obtained for the

training and testing data are then compared with existing model.

Ria Faulina, Suhartono [21], discussed hybrid and ensemble model of forecasting method for ten-daily rainfall prediction based on ARIMA (Autoregressive Integrated Moving Average) and ANFIS (Adaptive Neuro Fuzzy Inference System) at six certain area in Indonesia. In this study, Triangular, Gaussian, and Gbell function are used as membership function in ANFIS [16].

Dhinaharan Nagamalai et al discussed about the typical weather conditions consisting of various seasons and geographical conditions in India. Extreme high temperatures at Rajasthan desert, cold climate at Himalayas and heavy rainfall at Chirapunji. These extreme variations in temperatures make us to feel difficult in inferring predictions of weather effectively. It requires higher scientific techniques methods like machine learning algorithms applications for effective study and predictions [17] of weather conditions. In this paper, they applied K-means cluster algorithm for grouping similar data sets together and also applied J48 classification technique along with linear regression analysis.

This is known as the deterministic model

$$Y = A + BX \qquad (1)$$

Here Y =Dependent variable X=independent variable A, B= Regression parameter which is reported in the form of constant parameter a, which is a reflection of the trend of the parameter lies between two variable.

Olaiya Folorunsho [3] [10] proposed to investigate the use of data mining techniques in forecasting maximum temperature, rainfall, evaporation and wind speed [6] [11]. This was carried out using Artificial Neural Network and Decision Tree algorithms and meteorological
data collected between 2000 and 2009 from the city of Ibadan, Nigeria. Classifier algorithms were analysed meteorological data and developed the weather condition. 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.

A.S.Cofiño et.al [13] represented meteorology analysis involved into three subtasks of Cross Grid Project, including air pollution issues, mesoscale weather forecasting for maritime applications and development of data mining systems. The challenging points are: seamless access to large distributed databases in the Grid environment, development of distributed data mining techniques suited to the meteorological and hydrological applications and integration in a user-friendly interactive way, including specific portal tools. These research works involved Spanish-Polish team are focused on implementation of data mining techniques. This application meets the requirements to be suitable for GRID processing, since databases are distributed among different weather services and research laboratories, and data mining algorithms are computing intensive, thus requiring parallel processing.

A. Artificial Neural Network

Prince Gupta, et.al., discussed an artificial neural network (ANN) [8] based model was developed for rainfall time series forecasting. Different topologies of Neural Networks were created with change in hidden layer, number of processing element and activation function. Mean Absolute Error (MAE), Mean Squared Error (MSE) and correlation coefficient (CC) are used to evaluate the model performance. On the basis of these evaluation parameter results, it is found that multilayer perceptron (MLP) network predict more accurate than other traditional models.

B. K- Nearest neighbor

The classification algorithm [7] [9] k-Nearest Neighbor is used based on Euclidean distance between two points, used to find out the closeness between unknown samples with the known classes by the domain value of temperature and humidity prediction of rain fall data has to be predicted depending on the classification algorithm.

C. Multi linear Regression Analysis

In Multiple regressions [1] [4] [5] there are more than two variables among which one is dependent variable and all others are independent variable and the equation look like this:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \ldots\ldots \beta_p x_{ip} \qquad (2)$$

The predict rainfall in any one of the future's year by using climatic factors. Now for moving towards this approach first they select 4 climate factors with rain dataset of Udaipur city, Rajasthan, India and multiple regression approach on that data set and find out predictable equation between rain and climate factors by knowing climate factors which is very useful for farmers for their agriculture purpose.

III. DATA MINING TASKS INVOLVES IN WEATHER RESEARCH

The Weather Research and Forecasting (WRF) Model is a next-generation mesoscale numerical weather prediction system designed for both atmospheric research and operational forecasting needs [17]. These data includes surface weather observations, upper air soundings [15], radar imagery, satellite imagery, three-dimensional analysis from National Weather Service meteorological models, lightning data, snow and ice cover data, and

climatological data and summaries. If needed, high-resolution numerical modeling can be coupled with available data to reconstruct an event. Forensic meteorological services include evaluating the following meteorological conditions:

- Wind, temperature, humidity, visibility, and precipitation
- Ice or snow cover
- Snow melt and refreezing
- Heat waves and cold waves
- Fogging and visibility
- Air pollution and odors

Extraction of information is not only process need to perform, data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation.

User interface is the module of data mining system that helps the communication between users and the data mining system [23]. User Interface allows the following functionalities:-

- Interact with the system by specifying a data mining query task.
- Providing information to help focus the search.
- Mining based on the intermediate data mining results.
- Browse database and data warehouse schemas or data structures.
- Evaluate mined patterns.
- Visualize the patterns in different forms.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge from Fig.1.

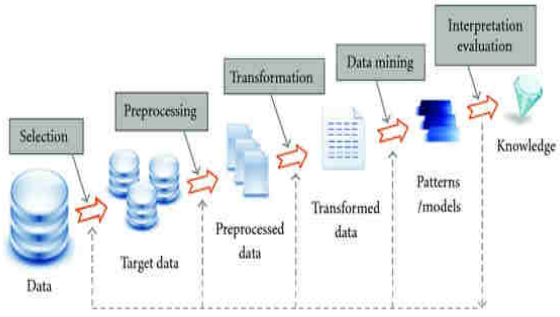

Fig.1: Data mining Task

The iterative process consists of the following steps [8]:

1. Data cleaning**:** Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistencies in data. Data cleaning involves transformations to correct the wrong data.
2. Data integration**:** It is a data preprocessing technique that merges the data from multiple heterogeneous data sources into a coherent data store.
3. Data selection**:** At this step, data Selection is the process where data relevant to the analysis task are retrieved from the database. Sometimes data transformation and consolidation are performed before the data selection process.
4. Data Transformation: It is the crucial step in which clever techniques are applied to extract data patterns potentially useful.
5. Data Mining: In this step, intelligent methods are applied in order to extract data patterns.
6. Pattern evaluation**:** In this step, data patterns representing knowledge are evaluated based on given measures.
7. Knowledge representation**:** It is the final phase in which the discovered knowledge is visually represented to the user. Techniques involves in data mining.

A. Clustering Technique

Clustering methods, depending of approach in cluster formation, are Partitioning methods, Hierarchical methods, Density-based methods, Grid-based methods, Models-based methods [22].

Partition-based methods construct the clusters by creating various partitions of the dataset. So, partition gives for each data object the cluster index pi. The user provides the desired number of clusters M, and some criterion function is used in order to evaluate the proposed partition or the solution. This measure of quality could be the average distance between clusters.

Hierarchical clustering methods build a cluster hierarchy, i.e. a tree of clusters also known as dendogram. A dendrogram is a tree diagram often used to represent the results of a cluster analysis. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down).

The key idea of density-based methods is that for each object of a cluster the neighborhood of a given radius has to contain a certain number of objects.

The grid-based clustering algorithms are quantizing the space into a finite number of cells that form a grid structure and then these algorithms do all the operations on the quantized space. Model-based methods hypothesize a model for each of the clusters and find the best fit of that model to each other.

The outlier detection problem in some cases is similar to the classification problem. For example, the main concern of clustering-based outlier detection algorithms is to find clusters and outliers, which are often regarded as noise that should be removed in order to make more reliable clustering.

B. Classification

The data analysis task is classification [7], where model or classifier is constructed to predict categorical labels. Data analysis task is an example of numeric prediction where the model constructed predicts a continuous – valued function or ordered value, as opposed to a categorical label. There are several classification methods such as,

- Decision tree
- Rule-based induction
- Neural networks
- Memory(Case) based reasoning
- Genetic algorithms
- Bayesian networks
- Rough set approach
- Fuzzy set Approach

C. Decision Tree

A Decision Tree is a flow-chart-like tree structure. Each internal node denotes a test on an attribute. Each branch represents an outcome of the test. Leaf nodes represent class distribution. The decision tree structure provides an explicit set of "if-then" rules, making the results easy to interpret [8]. In the tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. In decision analysis, a decision tree can be used visually and explicitly to represent decisions and decision making. The concept of information gain is used to decide the splitting value at an internal node. The splitting value that would provide the most information gain is chosen. Formally, information gain is defined by entropy. In other to improve the accuracy and generalization of classification and regression trees, various techniques were introduced like boosting and pruning.

D. Association rules mining

Finding frequent patterns, associations, correlations or casual structures among sets of items or objects in transaction databases, relational databases and other information repositories[23]. Support and Confidence involve to measure statistical analysis of the data.

Support

The rule X→Y holds with support s if s% of transactions in D contain X ∪Y. Rules that have a 's' greater than a user-specified support is said to have minimum support.

Confidence

The rule X→Y holds with confidence c if c% of transactions in D that contain X also contain Y. Rules that have a 'c' greater than a user-specified confidence is said to have a minimum confidence.

An itemset (or a pattern) is frequent if its support is equal to or more than a user specified minimum support (a statement of generality of the discovered association rules). Association rule mining is to identify all rules meeting user-specified constraints such as minimum support and minimum confidence (a statement of predictive ability of the discovered rules). One key step of association mining is frequent itemset (pattern) mining, which is to mine all itemsets satisfying user specified minimum support.

Complexity

Choice of minimum support threshold

- lowering support threshold results in more frequent itemsets
- This may increase number of candidates and max length of frequent itemsets.

Dimensionality (number of items) of the data set

- more space is needed to store support count of each item
- if number of frequent items also increases, both computation and I/O costs may also increase

Size of database

- Apriori makes multiple passes, run time of algorithm may increase with number of transactions

Average transaction width

- Transaction width increases with denser data set.
- This may increase max length of frequent itemsets and traversals of hash tree (number of subsets in a transaction increases with its width).

E. Similarity Functions

Similarity measure [8] is defined as the distance between various data points. The performance of many algorithms depends upon selecting a good distance function over input data set. While, similarity is an amount that reflects the strength of relationship between two data items, dissimilarity deals with the measurement of divergence between two data items.

An alternative concept to that of the distance is the similarity function s (x_i, x_j) that compares the two vectors x_i and x_j. A similarity function where the target range is [0, 1] is called a dichotomous similarity function. In fact, the methods described in the previous sections for calculating the "distances" in the case of binary and nominal attributes may be considered as similarity functions, rather than distances. This function should be symmetrical (namely s (x_i, x_j) = s (x_j, x_i)) and have a large value when x_i and x_j are somehow "similar" and constitute the largest value for identical vectors.

Here, it present a brief overview of similarity measure functions such as:

1. **Euclidean distance:** Euclidean distance determines the root of square differences between the

coordinates of a pair of objects. For vectors x and y distance d (x, y) is given by:

$$\textbf{Sim (x, y) = d} = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i \right)^2}$$

where x and y is n-dimensional vectors.

2. **Cosine distance:** Cosine distance measure for text clustering determines the cosine of the angle between two vectors given by the following formula [2]:

$$\textbf{Sim(xi, xj)= cos}\theta = \frac{\left(x_i \cdot x_j \right)}{\left(\left| x_i \right| \times \left| x_j \right| \right)}$$ Where, θ

refers to the angle between two vectors and xi, xj are n-dimensional vectors.

3. **Jaccard distance:** The Jaccard distance, involves the measurement of similarity as the intersection divided by the union of the data items [3]. The formulae could be stated as:

$$\textbf{Sim(xi, xj)} = \frac{\left(x_i \cdot x_j \right)}{\left(\left| x_i \right|^2 + \left| x_j \right|^2 - x_i \cdot x_j \right)}$$

4. **Pearson Correlation distance:** Pearson‟s correlation distance is another measure of the extent to which two vectors are related [3]. The distance measure could be mathematically stated as:

$$\textbf{Sim(x, y)} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n} \right)\left(\sum y^2 - \frac{(\sum y)^2}{n} \right)}}$$

IV. CONCLUSION

In this paper, it can be concluded that the particular survey represents meteorological data mining with finding the hidden data to discover the results generate different region of weather condition based on data mining tasks provide a very useful and accurate knowledge to predict the climatic condition of the region. It can also recognize that to follow the framework of data mining task has to be used to obtain useful prediction and support the decision making for different sectors. In future scope, it is found that to compare correlation coefficient technique and Fuzzy technique for identify the climatic condition in other direction.

REFERENCES

[1] Nikhil Sethi, Dr.Kanwal Garg “Exploiting Data Mining Technique for Rainfall prediction” , International Journal of Computer Science and information Technologies, Vol. 5 (3) , 2014, 3982-3984.

[2] Fallah-Ghalhary G.A., Mousavi-Baygi, M. and Habibi-Nokhandan, M.(2009). Seasonal Rainfall Forecasting Using Artificial Neural Network.Journal of Applied Sciences,9:1098-1105.

[3] Olaiya Folorunsho (2012):Application of Data mining Techniques in Weather Prediction and Climate change Studies.

[4] Paras ,et.al,” A Simple Weather Forecasting Model Using Mathematical Regression” in Indian Research Journal of Extension Education Special Issue (Volume I), January, 2012.

[5] Z.ismail,et.al,” Forecasting Gold Prices Using Multiple Linear Regression Method” in American Journal of Applied Sciences 6 (8): 1509-1514, 2009 ISSN 1546-9239.

[6] S. Kotsiantis and et. al., “Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values”, World Academy of Science, Engineering and Technology 2007 pp. 450-454

[7] Hair Nu Phyu, “Survey of classification techniques in Data Mining”, IMECS 2009 Volume 1 Hong Kong pp. 1-5

[8] Han J., Kamber M.: *Data Mining concepts and Techniques*, Elsevier Science and Technology, Amsterdam 2006.

[9] Cover T, Hart P (1967) “Nearest neighbor pattern classification”. IEEE Trans Inform Theory Volume 13(1) pp. 21–27.

[10] Olaiya, Folorunsho, and Adesesan Barnabas Adeyemo. "Application of data mining techniques in weather prediction and climate change studies."International Journal of Information Engineering and Electronic Business (IJIEEB) 4.1 (2012): 51.

[11] Lawrence, Mark G. "The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications." Bulletin of the American Meteorological Society 86.2 (2005): 225-233.

[12] Zhang, Guang Jun, and Michael J. Mcphaden. "The relationship between sea surface temperature and latent heat flux in the equatorial Pacific." Journal of climate 8.3 (1995): 589-605.

[13] A.S. Cofiño and J.M. Gutiérrez and B. Jakubiak and M. Melonek, Implementation of Data mining Techniques For Meteorological Applications”, World Scientific, 2003 (215-240).

[14] G. Li, and J. Shi, "On comparing three artificial neural networks for wind speed forecasting," Applied Energy, vol. 87, no. 7, pp. 2313-2320, Jul. 2010.

[15] Bilgin T., and Çamurcu Y., "A Data Mining Application on Air Temperature Database," Advances in Information Systems, Springer Berlin, Heidelberg, pp.68-76 .2004.

[16] Bartok J., Habala O., Bednar P., Gazak M., and Hluch L., "Data mining and integration for predicting significant meteorological phenomena," Procedia Computer Science, pp.37 – 46. 2010.
[17] Baboo S., and Shereef K., "Applicability of Data Mining Techniques for Climate Prediction – A Survey Approach," International Journal of Computer Science and Information Security, Vol. 8, No. 1, April 2010.
[18] Siddiqui, K.J. and Nugen, S.M., "Knowledge based system for weather information processing and forecasting," Geoscience and Remote Sensing Symposium, pp.1099-1101. 27-31 May 1996.
[19] Pritpal Singh, Bhogeswar Borah, monsoon rainfall prediction using artificial neural network, Springer 2013.
[20] Akash D Dubey, "Artificial Neural Network Models for Rainfall Prediction in Pondicherry ", International Journal of Computer Applications (0975 – 8887) Volume 120 – No.3, June 2015.
[21] RiaFaulina, Suhartono2, Hybrid ARIMA-ANFIS for Rainfall Prediction in Indonesia, International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064 Volume 2 Issue 2, February 2013.
[22] Berkhin, Pavel. 2006. "A survey of clustering data mining techniques."Grouping multidimensional data. Springer Berlin Heidelberg, 25-71.
[23] Kantardzic, 2011. Mehmed. Data mining: concepts, models, methods, and algorithms. John Wiley and Sons.