Weather Forecasting using Adaptive technique in Data Mining

Miss. Shraddha V. Shingne
M.Tech II year Dept. of Computer Science and engineering
Nuva College of engineering and technology
Nagpur, Maharashtra
s.shingne90@gmail.com

Prof. Anil D.Warbhe
Asst. Prof Dept. of Electronics engineering
Manoharbhai Patel institute of engg. and technology
Gondia , Maharashtra
anilwarbhe@hotmail.com

Prof. Shyam Dubey
H.O.D. Dept. of Computer Science and engineering
Nuva College of engineering and technology
Nagpur, Maharashtra
Shyam.nuva@rediffmail.com

***Abstract* -** Now a day's climate change is very important challenge to sustain in urban living, to solve this problem there is need to study meteorology. Meteorology is the interdisciplinary scientific study of atmosphere i.e. temperature, pressure, humidity, wind, etc. The increasing availability of climate data during the last decades makes it important to find effective and accurate tools to analyze and extract hidden knowledge from this huge data. After collecting that meteorological data we use the data mining technique such as Classification, Prediction, Clustering and Outlier analysis. After applying the data mining technique we can find out Weather data and rare patterns present in the large dataset so as to transfer the retrieved information into important knowledge for classification and prediction of climate condition. Useful knowledge can play important role in understanding the climate variability and climate prediction. In turn, this concept can be used to support many important sectors that are affected by climate like agriculture, tourism, water resources and vegetation.

***Index Terms* -** *Data Mining, Data Mining Techniques, weather data, meteorological data , outliers.*

******

I. INTRODUCTION

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Outliers [3] are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.

Climate change is a widely recognized global environmental challenge [1]. In the real world context in climate, sudden heavy rain fall, no rainfall, acid rain ,earthquake, sudden increase or decrease in temperature etc. can be considered as outliers if occur at unusual time period. These unusual patterns are often referred to as different terms in different application domains, such as rare patterns [2], outliers [3], [4], faults [5], peculiarities or contaminants, and etc. [6].Forecasting is very important for prediction of the future events. Computer Science and technology together has made significant advances over the past several years and using those advanced technologies and few past patterns, it grows the ability to predict the future. Weather forecasting is directly dependent with the characteristics of the particulate matters present in the climate.

The analysis step of the "Knowledge Discovery in Databases" process, or KDD is, an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and post-processing of discovered structures. Detection of Rare Patterns in Climate Change using Data Mining Techniques is used to understand and detect the patterns of climate change .The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional .

II. RELATED WORKS

Many researchers and scientists used data mining technologies in areas of meteorology and weather prediction. Zhaoxia WANG, Gary LEE, Hoong Maeng CHAN, Reuben LI, Xiuju FU and Rick GOH proposed an adaptive Markov chain pattern detection (AMCPD) method [7]for disclosing the climate change patterns of Singapore through meteorological data mining. Meteorological variables, including daily mean temperature, mean dew point temperature, mean visibility, mean wind speed, maximum sustained wind speed, maximum temperature and minimum temperature are simultaneously considered for identifying climate change patterns in this study. The results depict various weather patterns from 1962 to 2011 in Singapore, based on the records of the Changi Meteorological Station.

SANJAY CHAKRABORTY Prof. N.K.NAGWANI LOPAMUDRA DEY used a generic methodology for weather forecasting is proposed by the help of incremental K-means clustering algorithm[8]. Weather in this research is done based on the incremental air pollution database of west Bengal in the years of 2009 and 2010.

Kotsiantis et al. [9] predict daily average, maximum and minimum temperature for Patras city in Greek by using six different data mining methods: Feed-Forward Back Propagation (BP), k-Nearest Neighbor (KNN), M5rules algorithm, linear least-squares regression (LR), Decision tree and instance based learning (IB3). They use four years period data [2002-2005] of temperature, relative humidity and rainfall. The results they obtained in this study were accurate in terms of Correlation Coefficient and Root Mean Square. Data mining have been employed successfully to build a very important application in the field of meteorology like predicting abnormal events like hurricanes, storms and river flood prediction [10]. These applications can maintain public safety and welfare. Godfrey C. Onwubolu1, Petr Buryan, Sitaram Garimella, Visagaperuman Ramachandran, Viti Buadromo and Ajith Abraham, presented the data mining activity that was employed in weather data prediction or forecasting. The approach employed is the enhanced Group Method of Data Handling (e-GMDH). The weather data used for the research include daily temperature, daily pressure and monthly rainfall [11].Sarah N. Kohail, Alaa M. El-Halees, described Data Mining for meteorological Data and applied knowledge discovery process to extract knowledge from Gaza city weather dataset [12]. S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias proposed a hybrid data mining technique that can be used to predict more accurately the mean daily temperature values [9]. These are all related work about this paper which can be used to find missing meteorological data.

III. DATA MINING IN METEOROLOGY

Meteorology [13] is the interdisciplinary scientific study of the atmosphere. It observes the changes in temperature, air pressure, and moisture and wind direction. Usually, temperature, pressure, wind measurements and humidity are the variables that are measured by a thermometer, barometer, anemometer, and hygrometer, respectively. There are many methods of collecting data and Radar, Lidar, satellites are some of them. Weather forecasts are made by collecting quantitative data about the current state of the atmosphere. The main issue arise in this prediction is, it involves high-dimensional characters. To overcome this issue, it is necessary to first analyze and simplify the data before proceeding with other analysis. Some data mining techniques are appropriate in this context. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to analyze important information in data warehouses. Consequently, data mining consists of more than collecting and analyzing data, it also includes analyze and predictions. 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.

There are different steps in which this method will be implemented and various techniques are used in each step as shown in the figure below to detect and predict the values of weather data.

IV. OVERALL PROCESS OF DATA MINING

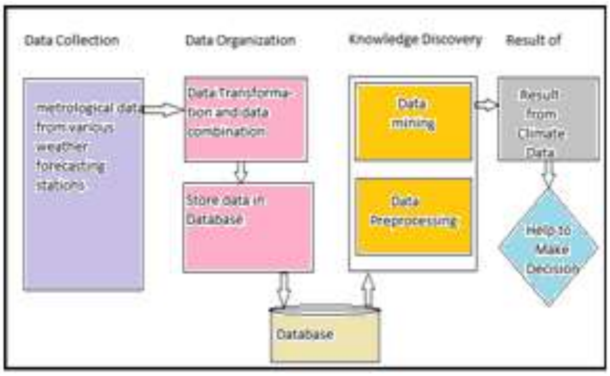

Fig.1Overall Proposed method

A. Collection of Data

The most important part for implementing any of the data mining technique is collection of data. For this purpose 10 channel midi-data logger system can be used. From this system we can get climate data in the form of excel sheets. Data Loggers are based on digital processor. It is an electronic device that keeps the record of data over the time in relation to location either with a built in instrument or sensor or via external instruments and sensors. Data Logger can automatically collect data on a 24 -hour basis; this is the primary and the most important benefit of using the data loggers [13]. It is used to capture the weather data from the local weather station to a dedicated PC located in the laboratory. The transmitted weather data was then copied to Excel spreadsheets and archived on daily basis as well as monthly basis to ease data identification.

Meteorological data in the form of daily summary data can also be extracted from National Climate Data Center (NCDC), National Oceanic and Atmospheric Administration (NOAA) [14]. Climate variables like longitude, latitude, mer wind, zone wind, humidity, air temperature and sea surface temperature etc. with most available values can be included in the weather forecasting data mining process.

B. Data Pre-Processing

An important step in data mining is data pre-processing. The challenge faced in knowledge discovery process is poor data quality. Thus, data should be pre-processed so that the noisy and unwanted data can be removed. In this field of study, metrological data is used for knowledge discovery which includes different types of metrological variables like sea surface temperature, mean temperature, humidity, rain, wind speed, etc., pre-processing means make the data set ready to apply data mining techniques to discover important knowledge. Data pre-processing includes following processes:

- *Data cleansing:* In this method the noisy and irrelevant data get removed from the data sets available [15]. Data cleansing is used to fill in missing values, make the noise present smoother while identifying outliers, and correct inconsistencies in the data

- *Data transformation:* Data transformation also called as data consolidation transforms the data available in the required format which is proper for the knowledge discovery by data mining procedure. Data transformation can involve the following [16]:

- Smoothing: This works to remove the noise from data. Such techniques include bining, regression, and clustering. Aggregation: This step is typically used in constructing a data cube for analysis of the data at multiple granularities. Generalization: Here low-level data are replaced by higher-level concepts through the use of concept hierarchies. Normalization: The attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0. Attribute construction: The new attributes are constructed and added from the given set of attributes to help the mining process. Data reduction: It includes data cube aggregation, attribute subset selection, dimensionality reduction, and discretization can be used to obtain a reduced representation of data while minimizing the loss of information content. Discretization and generating concept hierarchies: 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.

As we know huge amount of data for different metrological variables is missing or not recorded since so many years and such missing data cannot lead to appropriate results in data mining. For this reason data should be prepared carefully to obtain accurate and correct results. First we choose the most related attributes to our mining task. For this purpose we select all the data and try to fill the missing with appropriate values. Because we are working with weather data that is a form of time series, we must preserve the series smoothness and consistency. So the researchers can use different techniques like imputation method[17]. This method is effective method to fill missing values in the case of time series where the missed value is strongly related to its previous and next values
Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values. Missing values in data sets can be handled in different ways [17]:

- Ignorance of the tuple: It is usually done when label of class is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Manual replacement of missing values: In today's era time matters exclusively and the manual replacement of missing values leads so much overhead also it is very time consuming, tedious and infeasible way for the replacement of missing values.
- Global constant filled in the missing value: This also didn't prove to be a feasible solution for the replacements**.** e.g., "unknown", a new class?!
- Imputation: This uses the mean of all attributes to fill in the missing value, or use the attribute mean for all samples present in the data sets belonging to the same class to fill in the missing value which is quite smart solution for this.

From the logical point of view the missing values can be replaced easily by imputation [17].As it's a fact that sudden change in climate for each and every metrological variable in the data set is not possible, so missing values can be logically replaced. The number of missing values for each metrological variable can be one, two, three or its continuous and more than three values. If only single value is missing then it can be replace by the value of previous or next day for that particular variable as there will not much difference for every metrological value for that particular day. If continuous values are missing then each next missing value can be replaced by the mean of three consecutive, exactly previous values to the missing value. Thus the missing values can be efficiently replaced by imputation process using the attributes values belonging to that same class and we logically improve the imputation method to replace more proper values.

C. Knowledge Discovery

For knowledge discovery from data sets various following data mining techniques can be applied in Statistical Data Miner Software.

- *Classification:* Classification is the process of finding a model that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
- *Prediction:* Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute

values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.

- *Clustering:* Clustering analyses data objects without consulting a known class label. The unsupervised learning technique of clustering is a useful method for ascertaining trends and patterns in data, when there are no pre-defined classes.

D. Outlier analysis: A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.

These are techniques which are used in Data mining in the knowledge of discovery process. But here we used cluster analysis mainly.

E. Result of Analysis

The future values of metrological variables are predicted depending on the result of the classification algorithm.

VI. CLUSTER ANALYSIS

Clustering techniques [18] have a wide use and importance nowadays and this importance tends to increase as the amount of data grows. Data clustering, also called cluster analysis[13], is the discovery of semantically meaningful grouping of natural data sets or patterns or objects. It can also be defined as a given representation of a set of objects that are divided into k groups based on similarity measure so that the similarity between any two objects within a group, Kiis maximized and the similarity between any two objects within any two groups' Ki and kHz is minimized. Cluster analysis is prevalent in any discipline that involves analysis of multivariate data. It has been used for under laying structures to gain insight into data, generate hypothesis, detect anomalies and identify silent features , for a natural classification to identify the degree of similarity among engineering materials[23][24], and for compression to organize the data and summarizing it through cluster prototypes. Clustering [13] analyses data objects without consulting a known class label. The unsupervised learning technique of clustering is a useful method for ascertaining trends and patterns in data, when there are no pre -defined classes. There are two main types of clustering, hierarchical and partition. In hierarchical clustering, each data point is initially in its own cluster and then clusters are successively joined to create a clustering structure. This is known as the agglomerative method. In partition clustering, the number of clusters must be known a priori. The partitioning is done by minimizing a measure of dissimilarity within each cluster and maximizing the dissimilarity between different clusters.

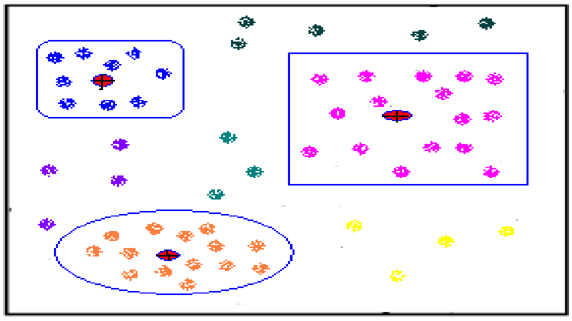

Fig. 2 Cluster of similar range data value

K-means clustering [18] is a partitioning based clustering technique of classifying/grouping items into k groups (where k is user specified number of clusters). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid. A centroid (also called mean vector) is "the center of mass of a geometric object of uniform density".

This algorithm randomly selects K number of objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and cluster mean. Then it computes new mean for each cluster. This process iterates until the criterion function converges

K-means Algorithm:

- Specify k, the number of clusters to be generated
- Choose k, points at random as cluster centers
- Assign each instance to its closest cluster center using Euclidean distance formula
- Calculate the centroid (mean) for each cluster; use it as a new cluster center
- Reassign all instances to the closest cluster center
- Iterate until the cluster centers don't change anymore.

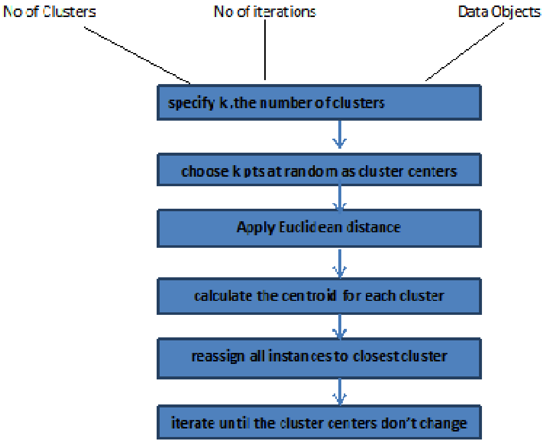
No of Clusters
No of iterations
Data Objects
specify k,the number of clusters
choose k pts at random as cluster centers
Apply Euclideen distance
calculate the centroid for each cluster
reassign all instances to closest cluster
iterate until the cluster centers don't change

Fig .3 Diagrammatic Representation k-means algorithm

VII. APPLICATION

Data mining have proved to be extremely useful for discovering useful knowledge from previous available data. The use and applications of data mining technologies in metrology and weather prediction areas are:

A. Estimating of minimum, maximum and mean temperature values at a particular time of day, from daytime and daily profiles, are needed for a number of environmental, ecological, agricultural and technical applications, ranging from natural hazards assessments, crop growth forecasting to design of solar energy systems.

B. Data Mining process applied to weather data acquired at the School of Engineering and Physics, University of the South Pacific, Fiji to demonstrate the usefulness of this emerging technology in practical real-life applications. The weather data include daily temperature and pressure observed using automated instruments and a chaotic rainfall data set observed for the city of Suva.

C. predicting abnormal events like acid rains,hurricanes, storms and river flood prediction

D. Finding a strong relation between severe conditions and the change tendencies of the measurements of the weather.

E. To detect severe events using data mining and volumetric radar data to detect storm events and classify them into four types: hail, Rainfall prediction,tornadoes, and wind.

F. The self-organizing data mining approach employed is the enhanced Group Method of Data Handling (e-GMDH). The weather data used for the DM research include daily temperature, daily pressure and monthly rainfall.

G. Gathering climate and atmospheric data, together with soil, and plant data in order to determine the inter-dependencies of variable values that both inform enhanced crop management practices and where possible, predict optimal growing conditions. The application of some novel data mining technique together with the use of computational neural networks as a means to modeling and then predicting frost.

g. identifying influence patterns and sentiment patterns in social networks.

h. Identifying weather category like hot, cold, cloudy, dusty etc.

These applications can maintain public safety and welfare.

VIII. CONCLUSION

In today's world to predict the atmosphere change, we need correct and accurate meteorological data to identify weather patterns in the long-term while consistent with global climate change on weather patterns. Large amount of Meteorological data is available in the form of daily summary data can also be extracted from National Climate Data Center(NCDC),National Oceanic and Atmospheric Administration(NOAA) In this paper we overviewed how data mining technique can be applied on the Metrological data for the knowledge discovery. With the help of this concept we can easily identify the occurrence of rare patterns in weather which helps us to predict future environmental changes in the climate.

References

[1] UNEP, *"United Nations Environment Programme, Climate Change*, "Available from: http://www.unep.org/climatechange; [cited 8 March 2013].

[2] Y. Meng, M. Dunham, F. Marchetti, and J. Huang, "*Rare event detection in a spatiotemporal environment,"* Proceedings of the Second IEEE International Conference on Granular Computing (GrC'06), pp. 10–12, 2006.

[3] Z. Yang, N. Meratnia, and P. Havinga, *"Outlier Detection Techniques for Wireless Sensor Networks: A Survey,"* IEEE Communications Surveys& Tutorials, vol. 12, no. 2, pp. 159–170, 2010.

[4] P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita, "*A Survey of Outlier Detection Methods in Network Anomaly Identification," The Computer Journal*, vol. 54, no. 4, pp. 570–588, Mar. 2011.

[5] Z. Wang, C. S. Chang, and Y. Zhang, *"A feature based frequency domain analysis algorithm for fault detection of induction motors," Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on*, pp. 27–32, 2011.

[6] V. Chandola, A. Banerjee, and V. Kumar, *"Anomaly detection: A survey," ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[7] Zhaoxia WANG, Gary LEE, Hoong Maeng CHAN,Reuben LI, Xiuju FU and Rick GOH, Pauline AWPoh Kimand Martin L. HIBBERD, Hoong Chor CHIN, *"Disclosing climate change patterns using an adaptive Markov chain pattern detection method"*. International Conference on Social Intelligence and Technology, 2013.

[8] SANJAY CHAKRABORTY Prof. N.K.NAGWANI LOPAMUDRA

[9] 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

[10] Bartok J., Habala O., Bednar P., Gazak M., and Hluch L., *"Data mining and integration for predicting significant meteorological phenomena,"* Procedia Computer Science, p.37 – 46. 2010.

[11] Godfrey C. Onwubolu1, Petr Buryan, Sitaram Garimella, Visagaperuman Ramachandran,Viti Buadromo and Ajith Abraham *"Self-organizing data mining for weather Forecasting"* IADIS European Conference Data Mining 2007 pp. 81-88

[12] Sarah N. Kohail, Alaa M. El-Halees, "*Implementation of Data Mining Techniques for Meteorological Data Analysis",* IJICT Journal Volume 1 No. 3, July 2011

[13] Meghali A. Kalyankar, Prof. S. J. Alaspurkar, *"Data Mining Technique to Analyse the Metrological Data"*. International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, pp. 114-118, Feb 2013

[14] "WMO, NNDC Climate Data Online, National Climatic Data Center, NESDIS, NOAA". Available from: http://www7.ncdc.noaa.gov/CDO/dataproduct; [cited 5 July 2011].

[15] University of Alberta, Osmar R. Zaïane, 1999, "*Chapter I: Introduction to Data Mining",* CMPUT690 Principles of Knowledge Discovery in Databases

[16] J. Han and M. Kamber. *Data Mining: Concepts and Techniques.* Morgan Kaufmann, 2000.

[17] Lakshminarayan K., S. Harp & T. Samad, *"Imputation of Missing Data in Industrial Databases"*. Applied Intelligence 11, 259-275, 1999.

[18] Ritu Sharma,M. Afshar Alam,Anita Rani, \K-Means Clustering in Spatial Data Min-ing using Weka Interface". International Conference on Advances in Communication and Computing Technologies (ICACACT), 2012.