Volume 3, Issue 2, February 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Data Mining Technique to Analyse the Metrological Data

Meghali A. Kalyankar
Master of Engineering in Computer Science,
University of Amravati,
Maharashtra, India

Prof. S. J. Alaspurkar
ME-I (CSE) GHRCEM Course Coordinator
University of Amravati
Maharashtra, India

Abstract— Data Mining is the process of discovering new patterns from large data sets, this technology which is employed in inferring useful knowledge that can be put to use from a vast amount of data, various data mining techniques such as Classification, Prediction, Clustering and Outlier analysis can be used for the purpose. Weather is one of the meteorological data that is rich by important knowledge. 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. We know sometimes Climate affects the human society in all the possible ways. Knowledge of weather data or climate data in a region is essential for business, society, agriculture and energy applications. The main aim of this paper is to overview on Data mining Process for weather data and to study on weather data using data mining technique like clustering technique. By using this technique we can acquire Weather data and can find the hidden patterns inside the large dataset so as to transfer the retrieved information into usable knowledge for classification and prediction of climate condition. We discussed how to use a data mining technique to analyze the Metrological data like Weather data.

Keywords— Data Mining, Data Mining Techniques, weather data, meteorological data

I. INTRODUCTION

Weather data has Synoptic data or climate data are the two classifications. Climate data is the official data record, usually provided after some quality control is performed on it. Synoptic data is the real-time data provided for use in aviation safety and forecast modelling. We know the Climate and weather affects the human society in all the possible ways. For example: Crop production in agriculture, the most important factor for water resources i.e. Rain, an element of weather, and the proportion of these elements increases or decreases due to change in climate. The effect of frost on the growth and quality of crops is leading potentially to total harvest failure. Energy sources, e.g. natural gas and electricity are depends on weather conditions. Hence changes Climate or weather condition is risky for human society as in all the possible ways [1] [7].

The increasing availability of climate data during the last decades (observational records, radar and satellite maps, proxy data, etc.) makes it important to find effective and accurate tools to analyze and extract hidden knowledge from this huge data. 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 [2]. Useful knowledge can play important role in understanding the climate variability and climate prediction. In turn, this understanding can be used to support many important sectors that are affected by climate like agriculture, vegetation, water resources and tourism.

The paper describes how to use a data mining technique and how to develop a system that uses numeric historical data to forecast the climate of a specific region or city. In this paper we try to extract

useful knowledge from weather data by using Clustering technique i.e. k-Means Partitioning Method and we discussed Data mining Process for Weather Data Analysis System. The rest of this paper is organized as follows: Section 3 presents some related works. Section 4 shows the introduction about Data mining in Meteorology. Section 5 shows steps for data mining process. Section 6 presents and discusses data mining techniques. Section 7 presents Cluster Analysis using k-Means Partitioning Method. Section 8 shows applications of data mining in metrological data like weather data. Finally conclusion and future work are presented in the last section.

II. RELATED WORKS

An easy Many researchers have tried to use data mining technologies in areas related to meteorology and weather prediction. Kotsiantis et al. [3] 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 [8]. 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 [4].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 [2].

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 [5]. These are all related work about this paper.

III. DATA MINING IN METEOROLOGY

Meteorology 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. Weather is one of the meteorological data that is rich by important knowledge [2].

IV. STEPS FOR DATA MINING PROCESS

There are different steps in which this paper will be implemented and various methodologies are used in each step as shown in the figure below to predict the values of weather data for ex. temperature and humidity parameters of climate with higher accuracy, and prove the prediction ability of data mining technique in the same context [1]. Title and Author Details

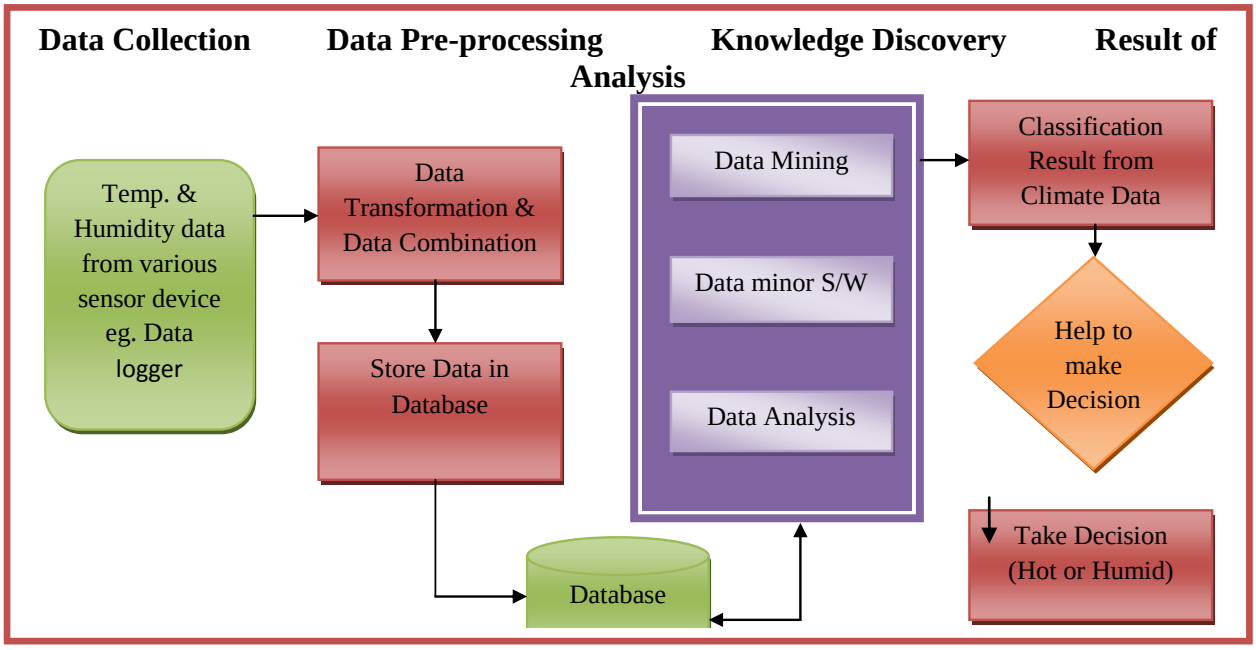


Figure.1 Design of Weather Data Analysis System **[2]**

A. Data Collection

This is the most important part while implementing any of the data mining technique and thus for this purpose we are using 10 channel midi-data logger system. This system provides weather data in form of excel sheets. Data Loggers are based on digital processor. It is an electronic device that record 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 [1]. 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.

B. Data Pre-Processing

The next important step in data mining is data pre-processing the challenge faced in knowledge discovery process in temperature and humidity data is poor data quality. Thus, data is to be pre-processed so as to remove the noisy and unwanted data. In this study, the weather data is used which consists of various parameters as temperature, humidity, rain, wind speed etc., pre-processing means removing the other unwanted parameters from the dataset. Data pre-processing includes following processes:

1) Data cleansing: It is also known as data cleaning, it is a phase in which noise data and irrelevant data are removed from the collection [9]. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data

2) *Data transformation:* It is also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure. Data transformation can involve the following [7]:

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.

C. Knowledge Discovery

For knowledge extraction various data mining techniques such as Outlier Analysis, Clustering, Prediction and Classification and Association rules can be applied in Statistical Data Miner Software

D. Result of Analysis

The future values of temperature and humidity are predicted depending on the result of the classification algorithm

V. DATA MINING TECHNIQUES

There are various data mining techniques such as:

A. *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.

B. *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.

C. *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.

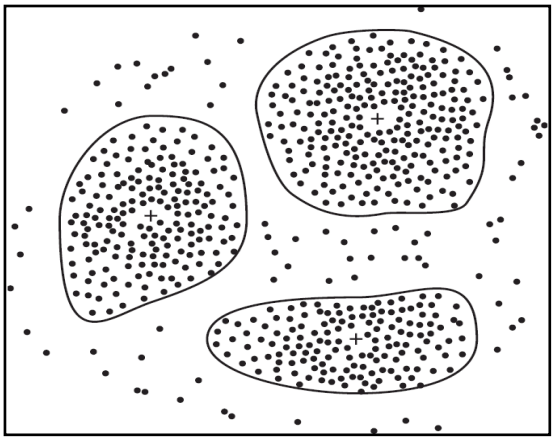
D. *Outlier analysis*

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.

VI. CLUSTER ANALYSIS

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. 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.



For example: Cluster Analysis can be performed on All Electronics Customer Data in order to identify homogeneous subpopulations of customers. This cluster may represent individual target groups for marketing. Figure 2 Cluster Analysis shows a 2-D plot of Customers with respect to customer locations in a city. Showing three clusters, each cluster "center" is marked with a "+".

In general, the major clustering methods can be classified into the following categories:

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods and
- Model-based methods

Here we will discuss about Partitioning method i.e. the k-Means Method:

The k-means algorithm takes the input parameter, k, and partitions set of n objects into k cluster so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster which can be viewed as the cluster's centroid or center of gravity.

For example: Suppose that there is a set of objects located in space as depicted in the rectangle figure 3.a). Let k=3; that is, the user would like the objects to be partitioned into three clusters.

Algorithm: The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

1. k: the number of clusters,
2. D: a data set containing n objects.

Output: A set of k clusters.

Method:

1. Arbitrarily choose k objects from D as the initial cluster centers;
2. Repeat
3. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. Update the cluster mean;
5. Until no change;

VII. APPLICATION

Application and use of data mining technologies in areas related to meteorology and weather prediction are:

A. Estimating of minimum, maximum and mean temperature values at a specific 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. DM 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 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, heavy rain, 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 techniques together with the use of computational neural networks as a means to modeling and then predicting frost.

These applications can maintain public safety and welfare.

VIII. CONCLUSION

Ideally, the market needs timely and accurate weather data. In order to achieve this, data should be continuously recorded from stations that are properly identified, manned by trained staff or automated with regular maintenance, in good working order and secure from tampering. The stations should also have a long history and not be prone to relocation. The collection and archiving of weather data is important because it provides an economic benefit but the local/national economic needs are not as dependent on high data quality as is the weather risk market.

In this paper we applied knowledge discovery process to extract knowledge from Gaza city weather dataset. We went through all knowledge discovery process and applied data mining technique i.e. clustering. Data mining tasks provide a very useful and accurate knowledge in a form of rules, models, and visual graphs. This knowledge can be used to obtain useful prediction and support the decision making for different sectors. Our future work includes building adaptive and dynamic data mining methods that can learn dynamically to match the nature of rapidly changeable weather nature and sudden events.

REFERENCES

[1] Badhiye S. S., Wakode B. V., Chatur P. N. "*Analysis of Temperature and Humidity Data for Future value prediction"*, IJCSIT Vol. 3 (1), 2012, 3012 – 3014

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

[3] 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

[4] 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

[5] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias, *"A Hybrid Data Mining Technique for Estimating Mean Daily Temperature Values",* IJICT Journal Volume 1 (5) pp. 54-59

[6] Thair Nu Phyu, *"Survey of classification techniques in Data Mining",* IMECS 2009 Volume 1 Hong Kong pp. 1-5

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

[8] 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.

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

[10] A.S. Coin and J.M. Gutierrez, B. Jakublak and M. Melonek, *"Implementation of Data Mining Techniques for Meteorological Applications"* DataMiningPaper.doc World Scientific,215-240,2003

[11] Berkhin P., *"Survey of clustering data mining techniques, Accrue Software, San Jose",* CA, Tech. Rep., 2002.

