IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Volume-3, Issue-5 E-ISSN: 2347-2693

A Survey On Meteorological Data Analysis Using Data Mining Techniques

M.Mayilvaganan[1*] and P.Vanitha[2]

[1*] *Department of Computer Science, PSG College of Arts and Science, India*
[2] *Department of Computer Science, Hindustan Arts and Science, India*

www.ijcseonline.org

Received: May/02/2015 Revised: May/10//2015 Accepted: May/24/2015 Published: May/30/ 2015

***Abstract*—** Meteorological data analysis in form of data mining is concerned to predict the knowledge of weather condition. To make an accurate prediction is one of the challenging of meteorologist to survey the weather condition efficiently. Rainfall prediction becomes an important issue in agricultural country like India. The atmospheric correlations play a significant role in determining the climate trends which are crucial in understanding the short and long-term trends in climate. The climate changes, so experienced today are mainly due to over ambitious strategies and actions of human being on the eco-system. In this paper, a survey of meteorological data analysis in various data mining techniques is presented. It can be provided for future direction for research.

Keywords— *Meteorological analysis; Data mining Process; Clustering Techniques; Classification methods; Association Rule*

I. Introduction

Weather occurs primarily due to density that is temperature and moisture is differences between one place and another. 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. Weather keeps changing from day to day and sometimes from hour to hour. When weather remains the same for a long period it calls as season. The commonly seasons are winter, spring, summer, autumn. The season of high temperatures, high winds and high moistures are resulting in potentially deadly weather. The most common data mining technique is used to identify and extract the weather condition based on Regression analysis, artificial neural networks, fuzzy logic techniques, k-Nearest Neighbor, multi linear regression analysis, Kmeans 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 prediction method.

II. RELATED WORKS

Data mining techniques should be able to handle noise in data or incomplete information. More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimensions increases.

Dathi et al. (1999) reported their findings on the trends present in the intensity of Canadian precipitation. It was found that southern areas of Canada partaken snowballing intensity for all seasons. Zhang et al. (2000) outlined Canadian precipitation trends along with temperature trends. It was seen that from 1900 to 1998, the annual mean temperature increased 0.5 to 1.5 in the southern part of Canada. Annual precipitation increased 5% to 35% during the same period. Finally, it was noted that there were significant negative trends in precipitation in southern regions during winter.

T. Sohn, [3] has developed a prediction model for the occurrence of heavy rain in South Korea using multiple linear and logistics regression, decision tree and artificial neural network.M. T. Mebrhatu [4] modeled for prediction categories of rainfall (below, above, normal) in the highlands of Eritrea. The most influential predictor of rainfall amount was the southern Indian Ocean SST. Experimental results showed that the hit rate for the model was 70%.

A. Regression Analysis

Ewona, presented the paper belongs to regression constants a and b were therefore extracted from the equations.

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. Monthly mean daily total rainfall shows marked latitudinal dependence as can be seen in the positive slopes of the graph of b against latitude. Rainfall data collected by the Nigerian Meteorological Agency. It shows consistent increase during the thirty years of this study. Constant b which is an indication of the volume of rainfall shows strong latitudinal dependence.

B. Artficial Neural Network

Deepak Ranjan Nayak, Amitav Mahapatra, Pranati Mishra presented in general, climate and rainfall are highly non-linear and complicated phenomena, which require advanced computer modeling and simulation for their accurate prediction. An Artificial Neural Network (ANN) can be used to predict the behavior of such nonlinear systems.

Nizar and Sanjay proposed an artificial neural network based model with wavelet decomposition for prediction of monthly rainfall on account of the preceding events of rainfall data. Wavelet transform an extraction of approximate and detail coefficient of the rainfall data series. The coefficients obtained from wavelet decomposition are used along with ANN for learning and knowledge extraction processes. After wavelet decomposition of rainfall time series, a multilayer perception with two hidden layer is found optimal for approximate coefficient prediction. Further focused time lag recurrent network with gamma memory is found optimal for prediction of detail coefficients. Thus a committee of two different ANN configurations is proposed for reliable rainfall prediction. The accuracy predicted for the rainfall model is reasonable.

C. K- Nearest neighbor

The classification algorithm k-Nearest Neighbor is used which is 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.

D. Multi linear Regression Analysis

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

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

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.

S. Nkrintra, [2] described the development of a statistical forecasting method for SMR over Thailand using multiple linear regression and local polynomial-based nonparametric approaches. SST, sea level pressure, wind speed, EiNino Southern Oscillation Index, IOD was chosen as predictors. The experiments indicated that the correlation between observed and forecast rainfall.

The weather data used for the research include daily temperature, daily pressure and monthly rainfall. 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.

III. DATA MINING INVOLVES METEOROLOGICAL DATA ANALYSIS

The meteorological data for use in air quality modeling consist of physical parameters that are measured directly by instrumentation, and include temperature, dew point, wind direction, wind speed, cloud cover, cloud layers, ceiling height, visibility, current weather, and precipitation amount. These data are used in air quality models to capture the atmospheric conditions occurring at a source and receptor location, and therefore, play an important role as they effect the concentration of pollutants at receptors of interest.

Data Mining, the process of analyzing data to find hidden patterns using automatic methodologies, is a powerful new technology with great potential to help companies focus on the most important in their data warehouse.

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.

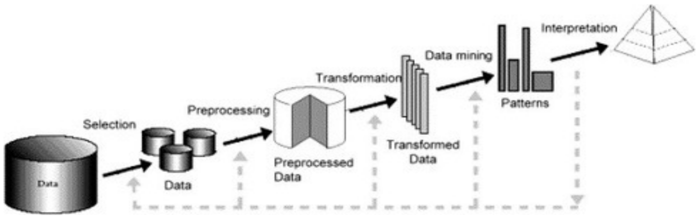

Fig.1 Data mining Task

The iterative process consists of the following steps:

1. Data cleaning**:** Also known as data cleansing, it is a phase in which noisy data and irrelevant data are removed from the collection.
2. Data integration**:** At this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3. Data selection**:** At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4. Data mining**:** It is the crucial step in which clever techniques are applied to extract data patterns potentially useful.
5. Pattern evaluation**:** In this step, strictly interesting patterns representing knowledge are identified based on given measures.

6. Knowledge representation**:** It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

The scope of the data mining is

- Automated prediction of trends and behaviors.
- Automated discovery of previously unknown patterns.

IV. Techniques involves in data mining

A. Clustering Technique

Cluster analysis organizes data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups, no predefined classification is required. The task is to learn a classification from the data. The representation can be investigated to see if the data group according to preconceived ideas or to suggest new experiments. Clustering algorithms can have different properties:

- Hierarchical or flat: hierarchical algorithms induce a hierarchy of clusters of decreasing generality, for flat algorithms, all clusters are the same.
- Iterative: the algorithm starts with initial set of clusters and improves them by reassigning instances to clusters.
- Hard and soft: hard clustering assigns each instance to exactly one cluster. Soft clustering assigns each instance a probability of belonging to a cluster.
- Disjunctive: instances can be part of more than one cluster.

The merging or division of clusters is performed according to some similarity measure, chosen so as to optimize some criterion.

B. Classification Method

Classification is a data mining technique used to classify each item in a set of data into one of predefined set of classes or groups. Data classification is a two step process.

- In the first step, a model is built by analyzing the data tuples from training data having a set of attributes. For each tuple in the training data, the value of class label attribute is known. Classification algorithm is applied on training data to create the model.
- In the second step of classification, test data is used to check the accuracy of the model. If the accuracy of the model is acceptable then the model can be used to classify the unknown data tuples.

In Classification algorithms there are several algorithms as follows

- Decision tree
- Rule-based induction
- Neural networks
- Memory(Case) based reasoning
- Genetic algorithms
- Bayesian networks

In Prediction Techniques, it is achieved with the help of regression. Regression analysis can be used to model the relationship between one or more independent or predictor variables and a dependent or response variable (continuous value).

- Linear regression
- Non Linear Regression

C. Rule induction

Rule induction is one of the most important techniques of machine learning. Since regularities hidden in data are frequently expressed in terms of rules, rule induction is one of the fundamental tools of Data Mining at the same time. Usually rules are expressions of the form

$i\ f$ (*attribute*-1, *value*-1) *and* (*attribute*-2, *value*-2) *and* ….. *and* (*attribute*$-n$, *value*$-n$) *then* (*decision, value*).

Some rule induction systems induce more complex rules, in which values of attributes may be expressed by negation of some values or by a value subset of the attribute domain.

D. Association rules mining

Essentially association mining is about discovering a set of rules that is shared among a large percentage of data. There are two ways of measuring usefulness, being objectively and subjectively. Objective measures involve statistical analysis of the data, such as Support and Confidence.

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

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 itemset satisfying user specified minimum support.

V. CONCLUSION AND FUTURE SCOPE

In this paper, it can be concluded that the particular survey presents knowledge discovery process to extract meteorological data from different region 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 Poisson distribution for identify the statistical analysis rainfall data 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 NeuralNetwork. 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] 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] Thair 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.