1st International Conference on Applied Soft Computing Techniques
22 & 23.04.2017
In association with
International Journal of Scientific Research in Science and Technology

Data Mining Techniques used in Climate Analysis - A Review

N. Krishnaveni*[1], Dr. A. Padma[2]
*[1] Department of Computer Science and Engineering,P.S.R Engineering College,Sivaksi,Tamil Nadu, India
veniengg@gmail.com[1]
[2] Department of Computer Science and Engineering, Muthayammal Engineering College, Rasipuram Tamil Nadu, India
giri.padma2000@gmail.com[2]

ABSTRACT

Data mining is technology is popular innovation. It converts files of data into useful information and knowledge that can help the data originators/users to pick choices and decide smart actions for data originators benefit. Data mining is the process to find for hidden patterns amongst vast sets of data. This data is useful to understand, predict behaviors for future. Overall, Data Mining is the set of methodologies used in analyzing data from various dimensions and perspectives, finding previously unknown hidden patterns, classifying and grouping the data and summarizing the identified relationships.

Keywords: Data mining, Clustering, Classification, Big data, R tool.

I. INTRODUCTION

Climate data analysis is an active research area, focused on analysis of change of climate conditions, analysis of climate phenomena, and analysis of interconnections of climate conditions. Data mining techniques provide better and faster analysis of large amounts of data in climatology. Climate data analysis, performed in order to understand climate change process and effect of different environmental factors in that change, has been focus of interest of researches for many years. The information extracted from database is used for the development of information technology. It generates large amount of data and huge volume of data in variety of research fields. To do research in variety of fields, data mining has given a various technology to store data and operate formerly stored data and to make decision based on mining process.

There are two methods to predict weather

1. **Empirical Approach:** This approach depends on investigation of past chronicled information of forecast which is gathered in meteorologist's focus and its connection to an collection of environmental variables over various parts of areas. The mostly utilized methodologies for climate predictions are Regression, decision tree, artificial neural network, fuzzy logic and LDA.

2. Dynamical Approach: In this approach, expectations are formed by physical models.By using these models, an arrangement of conditions that predict the future climate outline. To anticipate the climate by numeric means, meteorologist has generated atmosphere models that approximate the adjustment in temperature, weight. Climate speculation expectation is actualized. It is used with the utilization of exact measurable method.

II. DATAMINING PROCESS

Data mining is used to dig out inherent and earlier unknown information from data. Data Mining is the process which affords a concept to magnetize attention of users. Availability of massive amount of data, that needs to convert into useful information. So, many people bring into play the idiom “knowledge discovery device”or KDD for data mining. Seven sequential steps are used in data mining. [1]They are

1. Data cleaning: This phase removes noise and inconsistent data from collected raw data.
2. Data integration: This phase groups multiple data sources into single data store. It is called target data.

3. Data Selection: In this, data significant to examine task are get backed from data base as pre-processed data.
4. Data transformation: In this step, data is changed or merged into typical forms for mining by performing summing up and aggregation functions.
5. Data Mining: A range of smart techniques are applied in order to dig out data patterns.
6. Pattern evaluation: Data patterns are assessed.
7. Knowledge presentation: In this stage information is symbolized using representation techniques.

The purpose of knowledge finding and data mining process is to discover the patterns that are buried among the huge set of data and understand constructive knowledge and information.

III. PREDICTION TECHNIQUES

1. Artificial Neural Networks (ANN)

ANN is a mathematical model based on biological neural networks. It consists of an organized group of artificial neurons. It processes information using a connectionist move toward to calculation. Layers of Neurons are structured. The input layer consists of the original data, while the output layer represents the classes. There may be several hidden layers. Iterative learning process is the main feature of neural networks. In which data samples are offered to the network one at a time, and the weights are adjusted to foretell the correct class label. Advantages of neural networks include their high tolerance to noisy data. Ability to catalog patterns on which they have not been trained. Main concern of the training phase is to focus on the inner weights of the neural network, which is used according to the transactions used in the learning process. For each training transaction, the network receives in accumulation the expected output.

2. Bayesian Classifier

It is a, statistical classification approach based on the Bayes theorem.

Theorem:

To estimate probability of A given B,

P (B given A) = P (A and B)/P (A) the algorithm calculates the number of cases where A and B occurs concurrently and segregates it by the number of cases where A alone happens. Let X be a data tuple, X is judged Let H be some hypothesis, such that the data tuple X belongs to class C. P (H|X) is posterior probability, of H conditioned on X.P (H) is the prior probability of H in contract.

3. Decision Tree

It makes use of the simple divide-and conquer algorithm. In these tree structures, leaves represent

Classes and branches indicate conjunctions of features. It leads to those classes. The attribute most efficiently splits samples into different classes. A path to a leaf from the root is found depending on the assessment of the predicate at each node that is visited. To see coming the class label of an input. Decision tree is fast and easy method. It does not want any domain information. In the decision tree inputs are divided into two or more groups continue the steps till to complete the tree.[]

4. Hidden Markov Models

HMM is double implanted much more complex stochastic process. It provides mechanisms to teach the Markov Model <S, T> underlying a POMM from the sequence of clarification. Baum-Welch algorithm learns transition and observation probabilities. And also the state space (only the number of states has to be given). It is used to explain the observed training sequences. [10]

IV. DATA MINING TOOLS

1. Rapid Miner (erstwhile YALE)

Written in Java, gives advanced analytics methods. It is very popular because it is a readymade, open source and no-coding required. It incorporates multidimensional data mining functions such as data preprocessing, visualization, predictive analysis. It can be easily integrated with WEKA and R-tool to directly give models.

2. WEKA

It is a free JAVA based tool. It includes techniques for visualization, analysis, modeling, clustering, association, regression and classification.

3. R-Programming Tool

R programming is written in C. It permits the data miners to write scripts .So, it is utilized as to formulate statistical and analytical software. Graphical analysis,

classification, and clustering and time-based data analysis are maintained by R.

4. **Python based Orange and NTLK:**

   Due to ease of use and its powerful features Python is very popular. Orange, NTLK are the Python based open source tool that are very powerful language processing data mining tool .They are useful for data analytics, text analysis, and machine-learning and data scraping features. These features can easily be built up for adapted requirements.

5. **Knime:**

Mainly used for data preprocessing. Knime is a powerful tool with GUI that gives you an idea about the arrangement of nodes. Financial data analysts is a popular one and it has modular data pipe lining, machine learning, and data mining perceptions freely for building business intelligence reports.

V. WEATHER DATA MINING

The increasing research area in data mining technology is weather data mining. Data mining holds immense promising for weather forecasting to allow climate system to systematically use data and analysis to progress. In the weather forecasting managing data mining prediction are playing vigorous role. Some of the prediction based data mining techniques are as follows:

1. Artificial Neural network
2. Bayesian Classifiers
3. Decision tree and clustering.
4. Support Vector Machine (SVM)

In this system, R-Programming is used to analyze the climate data and used to predict weather.

VI. HADOOP AND MAP REDUCE PROGRAMING MODEL

[11] Hadoop and Map Reduce framework are the most widely used models today for Big Data processing. Hadoop is an open source large-scale data processing framework. It supports distributed processing of large chunks of data using simple programming models. The Apache Hadoop project consists of the HDFS file system and Hadoop Mapper and Reducer function. Hadoop is an open-source framework for dealing out a large amount of data across clusters of computers with the use of high-level languages like Java. Its modules provide easy to use languages, graphical interfaces and administration tools for managing data. Hadoop cluster is a set of machines networked together in one location. Data storage and processing all occur within this cloud of machines. User can submit jobs to Hadoop from his desktop machine in remote location from the Hadoop cluster[5].Two main components of Hadoop are Hadoop Distributed File System (HDFS) and MapReduce[11]. HDFS is a distributed file system management for large datasets of sizes of gigabytes and petabytes. And MapReduce is a programming framework for managing and processing huge amounts of unstructured data. Parallel processing of big dataset into smaller independent chunks.

VII. BIG DATA USING R

R is an open source software platform. Scope of big data analysis using R for statistical data analysis is popular. R is promptly adopted by statistics departments in universities Comprehensive R Archive Network and make it available to everyone. An excellent open – source interactive development environment has been created by R Studio for the R language. To increase the productivity of R users, [5] Google, Ford, Twit ter, US National Weather Service, The Rockefeller Institute of Government, The Human Rights Data Analysis Group makes use of R. Around the world, fascinated by its extensible nature as a platform for academic research. [5]Free in cost surely played a role as well. And it wasn't long before researchers in data science, statistics and machine learning started to publish papers in academic journals along with R code applying their new methods. R builds this process very easily and anyone can produce an R package to CRAN that stands for Comprehensive R Archive Network and make it available to everyone. An excellent open-source interactive development environment has been created by R Studio for the R language, further boosting the productivity of R users everywhere. [1]Google, Ford, Twitter, US National Weather Service, The Rockefeller Institute of Government, The Human Rights Data Analysis Group makes use of R.

VIII. CONCLUSION

To create a powerful and reliable statistical model, the following processes are more important. Like data transformation, evaluation of multiple model options and visualizing the results are essential. This is the reason why the R language has proven so popular: its interactive language uplifts investigation, explanation and presentation. Revolution R Enterprise gives the big - data support and speed to allow the data scientist to repeat through this process quickly. R programming model is used to analyze the climate data in very efficient manner.

IX. REFERENCES

[1] Joshi A, Kamble B, Joshi V, Kajale K, Dhange N. Weather forecasting and climate changing using data mining appli¬cation. International Journal of Advanced Research in Computer and Communication Engineering. 2015 Mar; 4(3) : 19–21.

[2] Caesar Wu, RajkumarBuyya, and Kotagiri Ramamohanarao. "Big data analytics = machine learning + cloud computing". CoRR, abs/1601.03115, 2016. Kaur, Anureet. "Big Data: A Review of Challenges, Tools and Techniques. IJSRSET, 2(2), 2016.

[3] Ijarcce Issn (Online) 2278-1021 Issn (Print) 2319 5940 International Journal Of Advanced Research In Computer And Communication Engineering Vol. 5, Issue 6, June 2016 Copyright

[4] To Ijarcce Doi 10.17148/Ijarcce.2016.56142 643 Weather Prediction Based On Big Data Using Hadoop Map Reduce Technique Basvanth Reddy 1, Prof. B.A Patil 2

[5] Big data analytics using r sanchita pati International research journal of engineering and technology (irjet) e-issn: 2395 -0056,volume:3 issue: 07 | july-2016www.irjet.netp-issn: 2395-0072

[6] Olaiya F. Application of data mining techniques in weather prediction and climate change studies. I. J. Information Engineering and Electronic Business. 2012 Jul ; 1: 51–59.

[7] Kalyankar MA, Alaspurkar SJ. Data mining technique to analyse the metrological data. International Journal of Advanced Research in Computer Science and Software Engineering. 2013 Feb; 3(2):114–118.

[8] Chauhan D, Thakur J. Data mining techniques for weather prediction: A review. International Journal on Recent and Innovation Trends in Computing and Communication. 2014 Aug; 2(4):2184–2189.

[9] Han, J., Micheline K., 2007, Data Mining: Concepts and Techniques, San Fransisco, CA: Morgan Kaufmann publishers.

[10] Rohit Kumar Yadav,Ravi Khatri 2016,"A weather Forecasting Model using the DataMining Technique,International Journal of Computer Application,Volume 139-No .14

[11] Dr. Doreswamy,and Ibrahim GadBig Data Techniques: Hadoop And Map Reduce For Weather Forecasting, International Journal of Latest Trends in Engineering and Technology Special Issue SACAIM 2016, pp. 194-199 e-ISSN:2278-621X.

[12] Basvanth Reddy1, Prof. B.A Patil2 Weather Prediction Based on Big Data Using Hadoop Map Reduce Technique, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016

[13] http://www.r-statistics.com/tag/hadley wickham/

[14] www.tutorialspoint.com