IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

A survey on Weather forecasting by Data Mining

Audireddy Gayathri[1], M. Revathi[2], J. Velmurugan[3]

M.Tech Scholar, Dept. of CSE, Sri Venkateswara College of Engineering and Technology, Chittoor, India[1]

Assistant Professor, Dept. of CSE, Sri Venkateswara College of Engineering and Technology, Chittoor, India[2]

Associate Professor, Dept. of CSE, Sri Venkateswara College of Engineering and Technology,
Chittoor and Research Scholar in VIT University, Vellore, India[3]

Abstract: Weather forecasting is one of the applications of data mining technology to predict the state of atmosphere for a future time and a given location as regards heat, cloudiness, dryness, wind, rain, etc. This paper focuses some of the data mining techniques for prediction of future weather. Classification algorithms such as Decision Tree Induction, Naïve Bayesian and Back propagation can be used to predict future by applied on the different parameters of weather. Prediction of weather must be accurate and also the weather should be forecasted earlier will be helpful for many applications like agriculture, air traffic, military and so on.

Keywords: Data Mining, Classification, Prediction, Naïve Bayesian, Back Propagation.

I. INTRODUCTION

a) Weather Forecasting

- Weather forecasting can be defined as the process of collecting data on atmospheric conditions, which records the humidity, rainfall, temperature, dew point, wind etc.
- Weather data can be collected from meteorological satellites and weather radars for weather forecasting. Currently Weather conditions changes rapidly due to the activities of humans and because of these natural disasters.
- To make an accurate forecast [5], a user must have data collected from the past. Must understand what processes are occurring in the atmosphere to produce the current weather at a particular location, which is then compared with the past data.
- Thus, the weather forecasting is considered as an important feature for day-to-day activities.

b) Types of Forecasting

The weather forecasts [7] are divided as,

- Now casting- in which the details about the current weather and forecasts up to a few hours ahead are given.
- Short range forecasts (1 to 3 days) – In which the weather in each successive 24 hr intervals may be predicted up to 3 days. This forecast range is mainly concerned with the weather systems observed in the latest weather charts, although generation of new systems is also considered.
- Medium range forecasts (4 to 10 days) – Average weather conditions and the weather on each day may be prescribed with progressively lesser details and accuracy than that for short range forecasts.
- Long range /Extended Range forecasts (more than 10 days to a season) - There is no rigid definition for Long Range Forecasting, which may range from a monthly to a seasonal forecast.

c) Parameters

- Weather is made up of multiple parameters [17] including air temperature, atmospheric pressure, humidity, precipitation, solar radiation and wind.
- Each of these factors can be measured to define typical weather patterns and to determine the quality of local atmospheric conditions.
- Based upon above parameters only estimate the extreme weather events like severe rainfall, heat waves, fog and droughts.

d) Data Mining Techniques

- Data Mining is the extraction of hidden patterns from data warehouses. It is a powerful technology with a great scope to analyze and predict vital information from the databases. Meteorological data are voluminous, dynamic, complex and high dimensional.
- There are several techniques [4] used in data mining as classification, Association [16], Regression , Cluster Analysis [20], Outlier Analysis and so on. Among these Classification technique is here mostly used.
- Classification is a supervised learning technique used to predict group membership for data instances. In case we may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy “or “cloudy”.
- Prediction models [3] continues-valued functions, it predicts unknown or missing values.

II. DATA MINING TOOLS

The Data Mining involves vast number of tools [18] used for the mining techniques. These tools comprises of different techniques as Classification, Prediction, Preprocessing, Clustering, Association and Visualization.

Some of the tools used are as follows,

DOI 10.17148/IJARCCE.2016.5262

- Weka, is a GUI and acts as one of the data mining tools. This includes the applications as Explorer, Experimenter, Knowledge Flow and Simple CLI. These possess the techniques as Classify, Preprocess, Associate, Visualize and so on.
- RapidMiner [8], is an open source tool with software platform that implements the integrated environment including mining and analytics. This not only processes the techniques such as classification, segmentation, regression but also validates the process for prediction automatically.
- OrangeCanvas, is one of the kinds of data mining tools that is used for unsupervised data. It includes the common classification, regression and the different types of regression. This also evaluates different plots, matrices and the predictions.

III. METHODOLOGIES USED

There are few common steps involved in implementing the algorithms as,

- Data Collection
- Data Preprocessing
- Data Transformation
- Applying Classification Algorithms
- Predicting the data

These mentioned are initial steps followed in each algorithm. Then it includes the individual steps and the methodologies involved.

a) Decision Tree Induction

Decision tree algorithm [2] is mainly used in data mining to analyses the data and to induce the tree and its rules that will be used to make predictions.

- There are different kinds of decision tree algorithms as ID3, C4.5, CART (Classification and Regression Tress), CHAID (Chi-squared Automatic Interaction Detection) and Quest.
- In decision tree construction, attribute selection measures are used to select the attribute that partition the data instances into distinct classes.
- ID3 uses information gain as its attribute selection measure. Information gain is the difference between original information requirement and the new requirement.
- C4.5 uses gain ratio as its attribute selection measure and CART uses gini index consider a binary split for each attribute.
- Therefore the above algorithms are used to analysis the past data and can predict the future data trends. In decision tree dependent variable is predicted from the independent variables.

b) Naïve Bayesian Classification

- Bayesian classifiers [11] are statistical classifiers that predict class membership probabilities. That is the probability that a given data instance belongs to particular class. This depends on Bayes' theorem.
- The basic idea of Bayes's theorem is that the outcome of a hypothesis or an event (H) can be predicted based on some evidences (X) that can be observed.
- A priori probability of H or P (H): This is the probability of an event before the evidence is observed.
- A posterior probability of H or P (H | X): This is the the probability of an event after the evidence is observed.

$$\mathbf{P(H/X)} = \frac{p\left(\frac{X}{H}\right)p(H)}{p(X)}$$

Bayesian Classifier works as follows

- Firstly D to be taken as training set of tuples or data instances and their associated class labels, tuple X =($x_{1,}$ $x_{2,\ldots.}x_{n)}$,depicts *n* measurements made on the tuple from n attributes, respectively, $A_{1,}$ $A_{2,\ldots..,}$ $A_{n.}$
- If there are m classes, $C_{1,}$ $C_{2\ldots\ldots}$ $C_{m.}$ For a given tuple X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. Therefore the naïve Bayesian classifiers predicts the tuple X belongs to the class C_i if and only if satisfy the below condition.

$$\mathbf{P\ (C_i\,/\,X) > P\ (C_j\,/\,X)}$$

- Therefore we maximize $P(C_i/X)$.The class C_i for which $P(C_i/X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' Theorem,

$$\mathbf{P(Ci/X)} = \frac{p\left(\frac{X}{Ci}\right)p(Ci)}{p(X)}$$

- In above equation P(X) is constant for all classes, so only $P(X/C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is $P(C_1) = P(C_2) = \ldots.\ P(C_m)$, therefore we want to maximize $P(X/C_i)$.
- Class prior probabilities are estimated by P $(C_i) = |C_{i,\ D}|\ /\ |D|$, here $|C_{i,\ D}|$ is the number of training instances of class C_i in D.
- The naïve assumption of class conditional independence is made. This deduces that the values of the attributes are conditionally independent of one another, given the class label of that tuple. In below equation $X_{\ k}$ refers to the value of attribute $A_{\ k}$.

Hence,

$$\mathbf{P(X/C_i) = P(X_1/C_i) \times P(X_2/C_i) \times \ldots\ldots. \times P\ (X_n/C_i)}$$

- So, the Bayesian classification algorithm takes past dataset for training and current dataset used as test data to predict [14] the values of dependent variable as class label. Probabilities of testing data tuple belongs to certain class value can be calculated by the use of above theorem.

c) Back propagation

Back propagation [1] is neural network learning algorithm. Neural network refers to the set of connected input or output units in which every connection has a weight associated with it. By adjusting the weights when learning phase network to be able to predict correct class label of given data instance.

DOI 10.17148/IJARCCE.2016.5262

- It is training or learning algorithm also called as Feed forward Networks or multilayer perceptrons(MLP). Back propagation network learns by example. Here we train the network, by giving new examples, change the network's weights and get the output (target).

1. The network is initialized by setting up all its weights to random numbers.
2. The input pattern is applied to get the output (forward pass)
3. Calculate the error of each neuron (target – actual value)
4. Error is mathematically changed, to minimize it.
5. Repeat steps b) to id) such that target is closer to actual value (reverse pass)

- Back propagation algorithm [6] is a supervised learning method because class label training tuples are known. This can be categorized into two phases as propagation and weight update. Until the performance of the network is satisfactory, the two phases are repeated.
- In this algorithm[19], the output is compared with the target to compute the value of predefined error-function. This error is then fed back to the network.
- The algorithm then adjusts the weights of each connection in order to reduce the value of the error function. This process is repeated for a sufficiently large number of training cycles, until the network converge at a state where the error function is relatively small. At this point, we can conclude that the network model is ready for test phase.
- By this we can predict [13] the values of parameters by changing value of any one parameter. For example in weather parameters like temperature varying with some unit how changing the other parameters like humidity, dew point, precipitation and so on.

IV. COMPARISON

Thus the weather forecasting can be done all these various methods specified such as Decision Tree, Bayesian Classification and Back Propagation.

The Decision Tree considers the past data and generates the tree and provides the best attribute to be considered for the future prediction. The Bayesian Classification provides the independence of attribute values for the given training data and based on this training data the future data is obtained for the test data considered. In the Back Propagation algorithm this considers the parameters through layers and processes information regarding the parameters in future and also depicts the changes that occur in different parameters when one parameter relates to any change gradually.

Among all these methodologies specified the Bayesian Classification is the one used mostly for the purpose of prediction as this can be applied for the large amount of data with ease. This also obtains data accuracy with time complexity when compared to the other methods.

V. FUTURE WORK

The Bayesian Classification encounters a disadvantage regarding the assumption of attribute values resulting in degradation of performance. Thus, the Weighted Bayesian [9] Classification can be considered as the future work that results in better computational performance by assigning the weights [10] to the attributes. Among these the important attributes are given with highest weights.

VI. CONCLUSION

Weather forecasting has predominant importance in our day-to-day life. In this survey some of the essential techniques and methodologies for this forecasting with few of their disadvantages and the different tools used in the data mining are given. Thus the data mining is one of the emerging technologies with various classification and the prediction techniques for real-time applications.

REFERENCES

1. Classification and Prediction of Future Weather by using Back Propagation Algorithm-An Approach by Sanjay D. Sawaitul, Prof. K. P. Wagh, Dr.P. N. Chatur Government College of Engineering, Amravati, Maharashtra, India.
2. Decision Tree for the Weather Forecasting by Rajesh Kumar, Ph.D Asst. Prof., Dept. of ECS Dronacharya College of Engineering, Gurgaon, India.
3. Rainfall Prediction using data mining by Sangari.R.S, Dr.M.Balamurugan School of CSE, Bharathidasan University, Trichy, India.
4. Data Mining Techniques for Weather Prediction by Divya Chauhan, Jawahar Thakur Dept. of Computer Science ,Himachal Pradesh University Shimla , India.
5. Towards a Self-Configurable Weather Research and Forecasting System by Khalid Saleem, S. Masoud Sadjadi, Shu-Ching Chen ,School of Computing and Information Sciences, Florida International University, Miami FL.
6. An Efficient Weather Forecasting System using Artificial Neural Network by Dr. S. Santhosh Baboo and I.Kadar Shereef.
7. Convective weather forecast accuracy analysis at center and sector levels by yao wang and banavar sridhar, nasa ames research center, moffett field, California.
8. Artificial Neural Networks' Application in Weather Forecasting – Using RapidMiner by **A** Geetha, G M Nasira **,**Mother Teresa Women's University ,Kodaikanal.
9. Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting by Nayyar A. Zaidi, Jesus Cerquides, Mark J. Carman Geoffrey I. Webb, Monash University VIC 3800, Australia.
10. Locally Weighted Naive Bayes by Eibe Frank, Mark Hall, and Bernhard P fahringer Department of Computer Science, University of Waikato Hamilton, New Zealand.
11. A Tutorial on Naive Bayes Classification by Choochart Haruechaiyasak.
12. Prediction of rainfall using Data mining technique over Assam by Pinky saikia dutta, hitesh tahbilder, Guwahati University ,Gauhati, Assam, India.
13. Prediction of Severe Thunderstorms applying Neural Network using RSRW Data by Himadri Chakrabarty, Sonia Bhattacharya Panihati Mahavidyalaya Barasat State University Kolkata, India.
14. Heart Disease Prediction System using Naive Bayes by Dhanashree S. Medhekar, Mayur P. Bote , Shruti D. Deshmukh.
15. Air Temperature Forecasting using Radial Basis Functional Artificial Neural Networks I.El-feghi, Zakaria Suliman zubi, A. Abozgaya, University of Tripoli,Tripoli-Libya.
16. Efficient Mining of Intertransaction Association Rules, Anthony K.H. Tung, Member, IEEE, Hongjun Lu, Member, IEEE, Jiawei Han, Member, IEEE, and Ling Feng, Member, IEEE.
17. Accuweather.com,http://www.accuweather.com
18. http://thenewstack.io/six-of-the-best-open-source-data-mining-tools.
19. Neural Networks and Back Propagation Algorithm Mirza Cilimkovic ,Institute of Technology Blanchardstown Blanchardstown Road North Dublin 15 Ireland.
20. Temperature and Humidity Data Analysis for Future Value Prediction using Clustering Technique: An Approach by Badhiye S. S., Dr. Chatur P. N. Wakode B. V. Government College of Engineering, Amravati, Maharashtra, India.

DOI 10.17148/IJARCCE.2016.5262