Weather forecast prediction: a Data Mining application

Ms. Ashwini Mandale, Mrs. Jadhawar B.A.

Assistant professor, Dr.Daulatrao Aher College of engg,karad,Ashwini.mandale@gmail.com,8407974457

Abstract— Weather forecasting is an important application in meteorology and has been one of the most scientifically and technologically challenging problems around the world. In this paper, we analyse the use of data mining techniques in forecasting weather. This can be carried out using Artificial Neural Network and Decision tree Algorithms and meteorological data collected in specific time. The performance of these algorithms was compared using standard performance metrics, and the algorithm which gave the best results used to generate classification rules for the mean weather variables. The results show that given enough case data mining techniques can be used for weather forecasting.

Keywords—ANN, CART algorithm, Data mining, Decision Tree, KDD, LAD, Weather prediction

INTRODUCTION

Weather prediction [9] has been one of the most interesting and fascinating domain. The scientists have been trying to forecast the meteorological characteristics using a large set of methods, some of them more accurate than others. Lately, there has been discovered that data mining, a method developed recently, can be successfully applied in this domain. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid prediction.

Contents-

Data means collection of information. Database means organized collection of data. Data warehouse means which provides enterprise with memory. Data Mining- It is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from huge amount of data .It is an interesting technique that can be implemented in various areas to generate useful information from the existing large volumes of data. Data mining has thus far been successfully implemented to bring success in commercial applications. Some of the applications of data mining include discovery of interesting patterns, clustering of data based on parameters and prediction of results by using the existing data. There are diverse techniques and algorithms available in data mining that can be implemented for various applications. This paper proposes an efficient data mining technique for weather forecast. Knowledge Discovery in Databases (KDD) is the whole process of finding useful information and patterns in data. Typical data mining architecture is as shown in Fig 1.

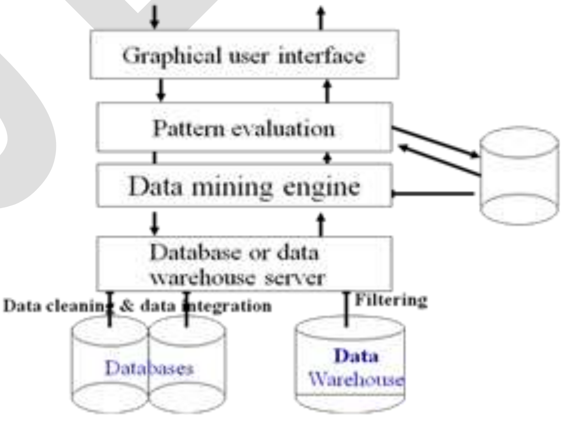


Fig 1. Typical data mining architecture

IJERGS

Data Mining Techniques- Data mining techniques search for interesting information without demanding a priori hypotheses, the kind of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data and predictive data mining tasks that attempt to do predictions based on inference on available data. This techniques are often more powerful, flexible, and efficient for exploratory analysis than the statistical techniques. The most commonly used techniques in data mining are: artificial neural networks, genetic algorithms, rule induction, nearestneighbour method and memory-based reasoning, logistic regression, discriminant analysis and decision trees.

Weather Forecasting –

Weather forecasting plays a significant role in meteorology. Weather forecasting remains a formidable challenge because of its data intensive and frenzied nature. Generally two methods are used to forecast weather: a) the empirical approach and b) the dynamical approach. The first approach is based on the occurrence of analogues and it is often referred to as analogue forecasting. This approach is useful in predicting local scale weather if recorded cases are plentiful. The second case is based upon equations and forward simulations of the atmosphere and is often referred to as computer modeling. The dynamical approach is useful to predict large scale weather phenomena and may not predict short term weather efficiently. Most weather prediction systems use a combination of both the techniques.

In this paper there are two techniques described for weather prediction:-

a) algorithm using both Artificial Neural Networks (ANN) and Decision Trees (DT) were used to analyze meteorological data gathered in-order to develop classification rules for the Application of Data Mining Techniques in Weather Prediction. Weather parameters over the study period and for the prediction of future weather conditions using available historical data. The targets for the prediction are those weather changes that affect us daily like changes in minimum and maximum temperature, rainfall, evaporation and wind speed. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a huge number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a particular application, such as pattern recognition or data classification, through a learning process. The artificial neuron is an information processing unit that is fundamental to the operation of a neural network. There are three basic elements of a neuron model. Fig.2 shows the basic elements of neuron model with the help of a perceptron model, which are, (i) a set of synapses connecting links, each of which is characterized by a weight or strength of its own, (ii) an adder for summing the input signals weighted by the respective synapses of the neuron and (iii) an activation function for limiting the amplitude of the output of a neuron. A typical input-output relation can be expressed as shown in Equation 1.

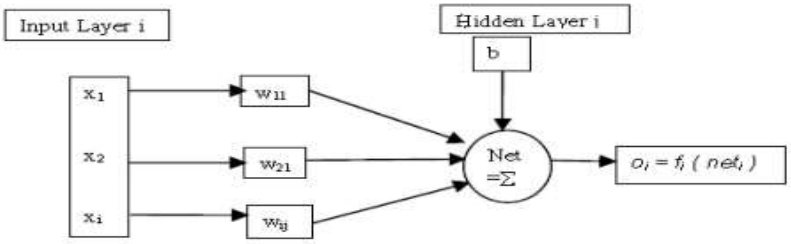


Fig 2: Model of a perceptron

$$net_j = \sum_{j=1}^{n} w_{ij} x_i + b_j$$
$$o_i = f_i\ (net_i) \qquad \ldots\ldots\ldots\ (1)$$

Where Xi = inputs to ith node in input, Wij = weight between ith input node and jth hidden node, b – bias at jth node, net = adder, f = activation function.

The type of transfer or activation function affects size of steps taken in weight space. ANN's architecture requires determination of the number of connection weights and the way information flows through the network, this is carried out by choosing the number of layers, number of nodes in each layer and their connectivity. The numbers of output nodes are fixed by the quantities to be estimated. The number of input nodes is dependent on the problem under consideration and the modeler's discretion to utilize domain knowledge. The number of neurons in the hidden layer is increased gradually and the performance of the network in the form of an error is monitored. A Decision Tree is a flow-chart-like tree structure. Each internal node denotes a test on an attribute. Each branch represents an outcome of the test. Leaf nodes represent class distribution. The decision tree structure provides an explicit set of "if-then" rules (rather than abstract mathematical equations), making the results easy to interpret. Depending on the algorithm, each node may have two or more branches. For example, CART [11] generates trees with only two branches at each node. Such a tree is called a binary tree. When more than two branches are allowed this is called a multiway tree [10]. In the tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications.

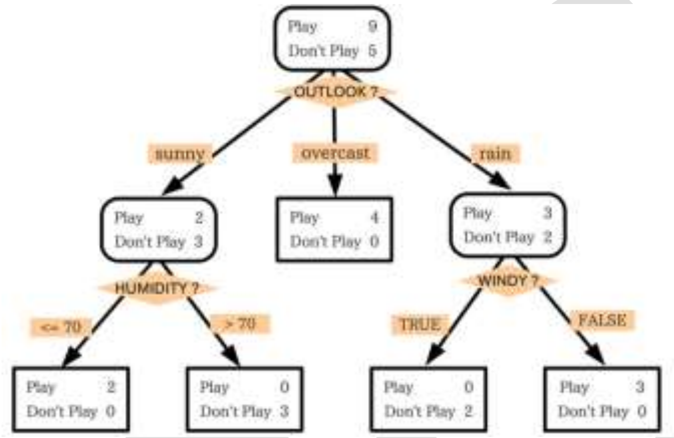


Fig 3:Decision tree

As shown in fig.3 provided here is a cannonical example in data mining, involving the decision to play or not play based on climate conditions. In this case, outlook is in the position of the root node. The degrees of the node are attribute values. In this example, the child nodes are tests of humidity and windy, leading to the leaf nodes which are the actual classifications. This example also includes the corresponding data, also referred to as instances. In our example, there are 9 "play" days and 5 "no play" days.

In decision analysis, a decision tree can be used visually and explicitly to represent decisions and decision making. The concept of information gain is used to decide the splitting value at an internal node. The splitting value that would provide the most information gain is chosen. Formally, information gain is defined by entropy. In other to improve the accuracy and generalization of classification and regression trees, various techniques were introduced like boosting and pruning. Boosting is a technique for improving the accuracy of a predictive function by applying the function repeatedly in a series and combining the output of each function with weighting so that the total error of the prediction is minimized or growing a number of independent trees in parallel and combine them after all the trees have been developed. Pruning is carried out on the tree to optimize the size of trees and thus reduce overfitting which is a problem in large, single-tree models where the model begins to fit noise in the data. When such a model is applied to data that was not used to build the model, the model will not be able to generalize. Many decision tree algorithms exist and these include: Alternating Decision Tree, Logitboost Alternating Decision Tree (LAD), C4.5 and Classification and Regression Tree (CART).

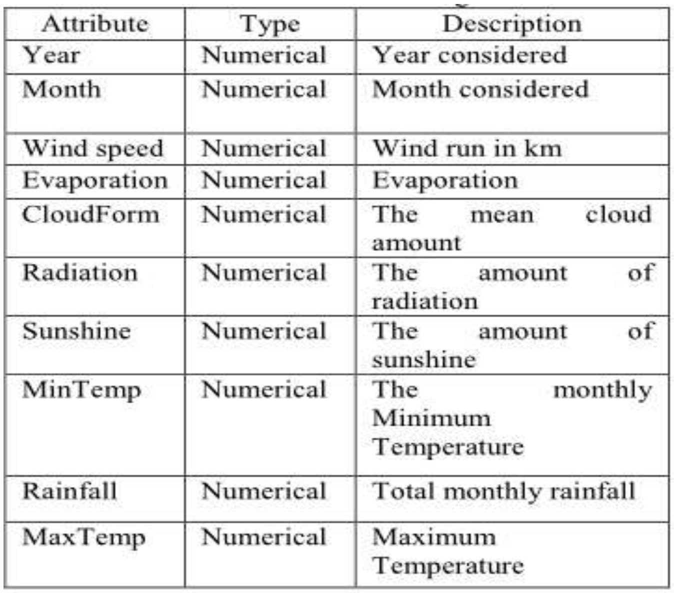
Materials and Methods-

Data Collection -The data used for this work was collected from specific region. Following stages of the research applied on collected data: Data Cleaning, Data Selection, Data Transformation and Data Mining.

Data Cleaning- In this stage, a consistent format for the data model was developed which took care of missing data, finding duplicated data, and weeding out of bad data. Finally, the cleaned data were transformed into a format suitable for data mining. A very low-quality information is available in various data sources and on the Web; many organizations are interested in how to transform the data into cleaned forms which can be used for high-profit purposes. This goal generates an urgent need for data analysis aimed at cleaning the raw data.

Data Selection - At this stage, data relevant to the analysis was decided on and retrieved from the dataset. The meteorological dataset had ten (10) attributes, their type and description is presented in Table 1. Due to the nature of the Cloud Form data where all the values are the same and the high percentage of missing values in the sunshine data both were not used in the analysis.

Table 1: Attributes of Meteorological Dataset



Attribute	Type	Description
Year	Numerical	Year considered
Month	Numerical	Month considered
Wind speed	Numerical	Wind run in km
Evaporation	Numerical	Evaporation
CloudForm	Numerical	The mean cloud amount
Radiation	Numerical	The amount of radiation
Sunshine	Numerical	The amount of sunshine
MinTemp	Numerical	The monthly Minimum Temperature
Rainfall	Numerical	Total monthly rainfall
MaxTemp	Numerical	Maximum Temperature

Data Transformation-This is also known as data consolidation. It is the stage in which the selected data is transformed into forms appropriate for data mining. The data file was saved in Commas Separated Value (CVS) file format and the datasets were normalized to reduce the effect of scaling on the data.

Data Mining Stage -The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the meteorological datasets. The testing method adopted for this research was percentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Thereafter interesting patterns representing knowledge were identified.

Evaluation Metrics

In selecting the appropriate algorithms and parameters that best model the weather forecasting variable, the following performance metrics were used

1. Correlation Coefficient: This measures the statistical correlation between the predicted and actual values. This method is unique in that it does not change with a scale in values for the test cases. A higher number means a better model, with a 1 meaning a perfect statistical correlation and a 0 meaning there is no correlation at all.

2. Mean Squared Error: Mean-squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value.

3. The Mean-squared Error: is simply the square root of the mean-squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values.

% Error: The percent error is defined by the following formula.

$$\%Error = \frac{100}{NP} \sum_{j=0}^{P} \sum_{i=0}^{N} \frac{\left| dy_{ij} - dd_{ij} \right|}{dd_{ij}} \quad \text{.} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where P = number of output processing elements

N = number of exemplars in the data set

dyij = denormalised network output for exemplar i at processing element j .

ddij = denormalised desired output for exemplar I at processing element j .

Experimental Design

C5 Decision Tree classifier algorithm which is implemented in See5 is used to analyze the meteorological data. The C5 algorithm is selected application of Data Mining Techniques in Weather Prediction, after comparison of results of tests carried out using CART and C4.5 algorithms [12]. The ANN algorithms used were those capable of carrying out time series analysis namely: the Time Lagged Feedforward Network (TLFN) and Recurrent networks implemented in NeuroSolutions 6 ANN development and simulation software). The ANN networks were used to predict future values of Wind speed, Evaporation, Radiation, Minimum Temperature, Maximum Temperature and Rainfall given the Month and Year.

ACKNOWLEDGMENT

I would like to express our sincere gratitude and thanks to Mrs.Birnale M.A., HOD, Computer Science and Engg department, Dr.Daulatrao Aher College of Engg, karad, for her help, Encouragement and intellectual influence, which made this paper possible. Her invaluable guidance, reviews and suggestions have been instrumental in the successful completion of this work.

CONCLUSION

In this the C5 decision tree classification algorithm was used to generate decision trees and rules for classifying weather parameters such as maximum temperature, minimum temperature, rainfall, evaporation and wind speed in terms of the month and years. Given enough data the observed trend over time could be studied and important deviations which show changes in climatic patterns can be identified**.** Artificial Neural Networks can detect the relationships between the input variables and generate outputs based on the observed patterns inherent in the data without any need for programming or developing complex equations to model these relationships. Hence given enough data ANN’s can detect the relationships between weather parameter and use these to predict future weather conditions This is important to climatic change studies because the variation in weather conditions in term of temperature, rainfall and wind speed can be studied using these data mining techniques.

REFERENCES:

[1] Bregman, J.I., Mackenthun K.M., 2006, Environmental Impact Statements, Chelsea: MI Lewis Publication.

[2] Casas D. M, Gonzalez A.T, Rodrígue J. E. A., Pet J. V., 2009, “Using Data-Mining for Short-Term Rainfall Forecasting”, Notes in Computer Science, Volume 5518, 487-490

[3] Due R. A., 2007, A Statistical Approach to Neural Networks for Pattern Recognition, 8th edition. New York: John Wiley and Sons publication.

[4] Elia G. P., 2009, “A Decision Tree for Weather Prediction”, Universities Petrol-Gaze din Ploiesti, Bd. Bucuresti 39, Ploiesti, Catedra de Informatică, Vol. LXI, No. 1

[5] Fairbridge R. W., 2007, "Climate" Microsoft® Student 2008 [DVD], Redmond, WA: Microsoft Corporation, 2007.

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

[7] Witten IH, Frank E, Data Mining: Practical Machine Learning Tools and Techniques. Second edition, 2005. Morgan Kaufmann.

[8] A Decision Tree for Weather Prediction by Elia Georgiana Petre, Vol. LXI No. 1/2009

[9] Introduction to Data Mining and Knowledge Discovery, Third Edition, Two Crowds Corporation, http://www.twocrows.com/intro-dm.pdf, accessed on 12 April 2009.

[10] Data mining Models and Algorithms, http://www.huaat.com/english/datamining/D_App.htm,accessed on 13 April 2009.

[11] Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont.

[12] Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Mateo.11

IJERGS