Weather Forecasting using Data Mining Techniques

Venkata Naga Satya Sangeeta Eluri
Department of Computer Science and Engineering
GITAM University
Visakhapatnam, Andhra Pradesh, India.
sangeetaevns@gmail.com

***Abstract*—Weather forecasting is a vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last century. Weather is a continuous, data-intensive, multidimensional, dynamic process that makes weather forecasting a formidable challenge. The rate of acquiring data is exploding and managing such data so as to infer useful knowledge that can be put to use, is becoming important. Data Mining is one such technology that is employed in inferring useful knowledge that can be put to use from a vast amount of data. Data mining is an interesting field of computer science that can be used for various applications. This project aims at developing a weather information system as a web service that can be used by any type of application and uses the prediction technique of data mining for weather forecasting.**

Keywords—Weather Forecasting, Data mining, Decision tree, ID3 algorithm, Information Gain, Entropy, Random Forest, Temperature, Pressure, Humidity, Wind speed, Java, R.

I. INTRODUCTION

Data mining, also called Knowledge Discovery in Databases (KDD), is the process of extracting or mining knowledge from large amount of data. In other words Data mining is the efficient discovery of valuable, non-obvious, novel and potentially useful information from a large collection of data. It extracts hidden predictive information from large databases, thus is a powerful new technology with great potential to help in analysis of data and for decision making. Data mining is an interesting technique that can be implemented to many areas. 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. 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. Useful knowledge can play important role in understanding the climate variability and climate prediction. This understanding can be used to support many important sectors that are affected by climate like agriculture, water resources and tourism. To make an accurate prediction is one of the major challenges facing meteorologist all over the world. Weather is one of the meteorological data that is rich by important knowledge.

Weather forecasting is a prediction of what the weather will be like in future, it had been invented many years ago. A weather forecast involves five steps: observation, collection and transformation of data, plotting of weather data, analysis of data and extrapolation to find the future state of the atmosphere, and prediction of particular variables. There are various classification methods that can be used, like Decision Trees, Random Forest, Naïve Bayes classifier, Artificial Neural Networks and Support Vector Machines which come under the category of supervised methods, whereas the unsupervised method is an adaptation of the K-means clustering method.

II. METHODOLOGY

Accurate prediction of weather parameters is a difficult task due to the dynamic nature of atmosphere. In this work, two approaches are used i.e. Decision trees, implemented using ID3 algorithm and Random Forest algorithm for forecasting the weather.

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. In the tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. 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. Decision trees can be implemented using algorithms like ID3, ASSISTANT and C4.5.

We calculate entropy and Information gain for the classification.

Entropy and Information Gain: Entropy is the sum of the probability of each label times the log probability of that same label.

- Select the attribute with the highest information gain
- Assume there are two classes, *P* and *N*

Let the set of examples *S* contain *p* elements of class *P* and *n* elements of class *N*

The amount of information, needed to decide if an arbitrary example in *S* belongs to *P* or *N* is defined as

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Assume that using attribute A a set *S* will be partitioned into sets $\{S_1, S_2, \ldots, S_v\}$

- If S_i, the entropy, or the expected information needed to classify objects in all sub-trees S_i is :
- The encoding information that would be gained by branching on *A*

$$Gain(A) = I(p,n) - E(A)$$

ID3 algorithm is a simple decision tree learning algorithm, developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to build a decision tree by employing a top-down search through the given sets to test each attribute at every tree node. The algorithm uses a greedy search, i.e., it chooses the best attribute and never reconsiders the choice already made.

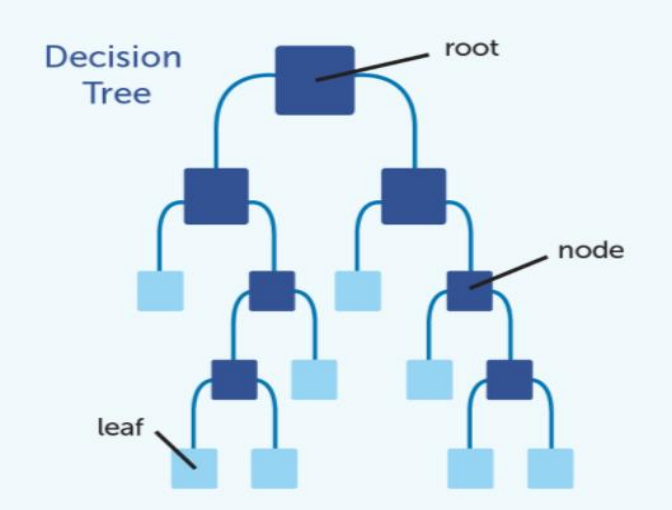
The main steps of the ID3 algorithm are:

- For each attribute in the database, compute its entropy.
- The current node is the attribute (A) with the highest information gain.
- For every value of the attribute A build a sub-tree;

  If A=value 1 then generate subtree1

  If A=value 2 then generate subtree2, etc.
- For each sub tree, repeat this process from the first step.
- Every time a new node is created in the tree with a variable, the attribute is removed from the variables group.
- Chooses attribute that has the lowest entropy is minimum or when information gain is maximum

The process stops when there are no attributes left.



Decision trees are preferred over other classification techniques because:

- Understandable prediction rules are created from the training data, which are simple to be interpreted.
- Builds the fastest tree.
- Builds a short tree.
- Only need to test enough attributes until all data is classified.

Random Forest:

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Algorithm:

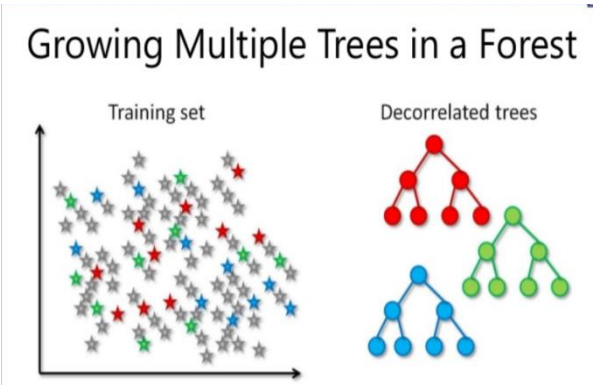
It sample N cases at random with replacement to create a subset of the data. The subset should be about 66% of the total set. At each node:

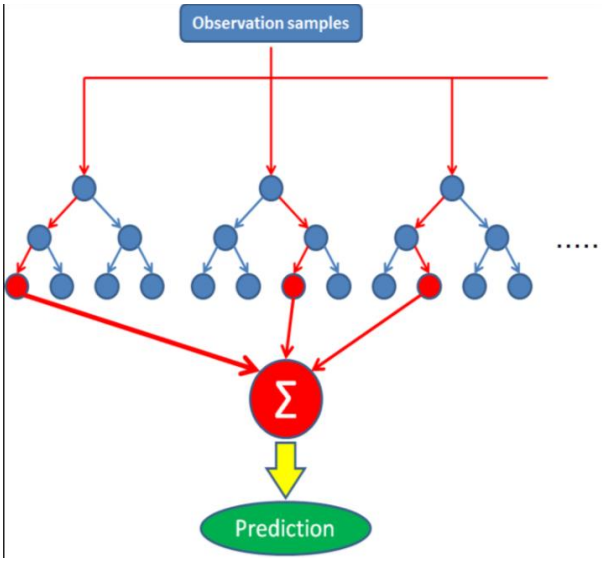
- For some number m, m predictor variables are selected at random from all the predicted variables.
- The predictor variable that provides the best split, according to some objective function, is used to do a binary split on that node.
- At the next node, choose another 'm' variable at random from all predictor variables and do the same.

The training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This OOB (Out –Of – Bag) data is used to get a running unbiased estimate of the classification error as trees

are added to the forest. It is also used to get estimates of variable importance.

After each tree is built, all of the data are run down the tree and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers and producing illuminating low dimensional views of the data.





Random Forest Classification

Out -Of -Bag (OOB) error estimate: In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run as follows:

Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree. Put each case left out in the construction of the kth tree down the kth tree to get classification. In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, take j to be the class that got most of the votes every time case 'n' was OOB. The proportion of times that j is not equal to the true class of 'n' averaged over all cases is the OOB error estimate. This has proven to be unbiased in many tests.

Variable Importance: In every tree grown in the forest, put down the OOB cases and count the number of votes cast for the correct class. Now randomly permute the values of variable 'm' in the OOB cases and put these cases down the tree. Subtract the number of votes for correct class in the variable 'm' permuted OOB data from the number of votes for the correct class in the untouched OOB data. The average of this number over all trees in the forest is the raw importance score for variable 'm'.

If the values of this score from tree to tree are independent, then the standard error can be computed by a standard computation. The correlations of these scores between trees have been computed for a number of data sets and proved to be quite low, therefore we compute standard errors in the classical way, divide the raw score by its standard error to get a z-score, ands assign a significance level to the z-score assuming normality.

If the number of variables is very large, forests can be run once with all the variables, then run again using only the most important variables from the first run.

For Decision Tree approach, Net Beans IDE is used for the complete implementation of the classification. A User-Interface is to be created which allows the user to input the weather parameters we ought to classify and as well display the classified results. NetBeans is an integrated development environment (IDE) for developing primarily with Java, but also with other languages, in particular PHP, C/C++, and HTML5. It was brought in by the Sun Microsystems in 1999. The philosophy behind Net Beans is to provide an extensible IDE that provides all the tools necessary to develop desktop, enterprise, web and mobile applications. The ability to install plug-ins allows developers to tailor the IDE to their individual development tastes. It also serves as an application platform framework for Java desktop applications and others. The Net Beans IDE is written in Java and can run on Windows, OS X, Linux, Solaris and other platforms supporting a compatible JVM. This platform allows applications to be developed from a set of modular software components called modules.

For Random Forest approach, Random Forest package in R is used for the implementation of weather data classification. Random Forest package uses existing weather data for training and builds a classification model. Using the classification model, weather label is classified for test data.

III. LITERATURE REVIEW

Meteorological Department has progressively expanded its infrastructure for meteorological observations, communications, forecasting and weather services and it has concurrently contributed to scientific growth. Denis Riordan

and Bjarne K Hansen (2002) stated that Case-based reasoning is emerging as a leading methodology for the application of artificial intelligence. They described an investigation into the application of case-based reasoning in airport weather forecasting. Knowledge about temporal features that human forecasters use to construct analogous climatic scenarios is encoded in a fuzzy similarity measure. A fuzzy case-based system for weather prediction similarity measure is used to locate the k-nearest neighbours from the historical database. These nearest neighbours are in turn adapted to produce values for the forecast parameters, which increase the accuracy of predictions. This approach of fuzzification was further developed and a new method for classification was also implemented using Artificial Neural Networks.

Imran Maqsood et.al, (2004) have proposed an ensemble of neural networks approach for weather forecasting. This study presented the weather forecasting in southern Saskatchewan; Canada. The proposed method used back then, for weather forecasting was advantageous over other techniques like linear combination. Generally, the output of an ensemble is a weighted sum, which is weight-fixed, with the weights being determined from the training or validation data. In the proposed approach, weights are determined dynamically from the respective certainties of the outputs. Further, the use of ANN's in this sector was implemented.

Elia Georgiana Petre in (2009) has implemented a decision tree, which represents a decision support tool very often used because it is simple to understand and interpret. Classification and Regression Trees -CART -is a technique formed by a collection of rules based on values of certain variables in the modelling data set. His paper presents a small application of CART for whether prediction. It had been chosen the data collection registered over Hong Kong. The data was recorded between 2002 and 2005. To build the decision tree he has used a free data mining software available under the GNU General Public License– Weka. Then, there are presented the decision tree, the results and the statistical information about the data used to generate the decision model.

Dr. S. Santhosh Baboo and I.Kadar Shereef(2011) also, have stated that temperature warnings are important forecasts because they are used to protect life and property. Temperature forecasts are made by collecting quantitative data about the current state of the atmosphere. They have present a neural network-based algorithm for predicting the temperature. The Neural Networks package supports different types of training or learning algorithms. One such algorithm is Back Propagation Neural Network (BPN) technique.The main advantage of the BPN neural network method is that it can fairly approximate a large class of functions. This method is more efficient than numerical differentiation.

Several steps followed to predict the temperature were:

a. Data collection (atmospheric pressure, temperature, wind speed and direction, humidity, precipitation) as inputs.
b. Data assimilation and analysis
c. Numerical weather prediction
d. Model output post processing

Through the implementation of this system, they illustrated how an intelligent system can be efficiently integrated with a neural network prediction model to predict the temperature. The algorithm improved convergence and damped the oscillations. The results obtained were compared with actual working of meteorological department and those results confirm that their model had the potential for successful application to temperature forecasting. Real time processing of weather data indicate that the BPN based weather forecast have shown improvement not only over guidance forecasts from numerical models, but over official local weather service forecasts as well.

Folorunsho Olaiya, Adesesan Barnabas Adeyemo (2012) have investigated the usage of data mining techniques in forecasting maximum temperature, rainfall, evaporation and wind speed. This was carried out using Artificial Neural Network and Decision Tree algorithms and meteorological data collected between 2000 and 2009 from the city of Ibadan, Nigeria. A data model for the meteorological data was developed and this was used to train the classifier algorithms. The performances of these algorithms were compared using standard performance metrics, and the algorithm which gave the best results used to generate classification rules for the mean weather variables. A predictive Neural Network model was also developed for the weather prediction program and the results compared with actual weather data for the predicted periods. The results show that given enough case data, Data Mining techniques can be used for weather forecasting and climate change studies. Weather Forecast is one of the important areas where classification technique plays a major role.

IV. MATERIALS AND METHODS

A. Data Collection:

The data used for this work was collected from website of University of Cambridge.

B. 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. Random Forest is an efficient algorithm to find the missing values in data. Finally, the cleaned data were transformed into a format suitable for data mining.

C. Data Analysis and Mining Stage:

In this stage, the ID3 algorithm for decision trees and Random forest algorithm is implemented; trees are generated in order to predict the weather, using the current or already present data.

V. REQUIREMENTS & RESULTS

Various requirements for this project are weather observations of

- Temperature
- Pressure
- Humidity
- Wind speed

Constraints used for Decision tree:

Temperature:

- If temp >= 55 and temp <= 70, temp = low.
- If temp >= 71 and temp <= 85, temp = medium.
- If temp >= 86 and temp <= 100, temp = high.

Pressure:

- If pressure >= 29.50 and pressure <= 29.70, pressure = low.
- If pressure >= 29.71 and pressure <= 29.90, pressure = medium.
- If pressure >= 29.91, pressure = high.

Humidity:

- If humidity >= 50 and humidity <= 65, humidity=low
- If humidity >= 66 and humidity <= 80, humidity=medium
- If humidity >= 81 and humidity <= 95, humidity=high.

The sample of training data format is

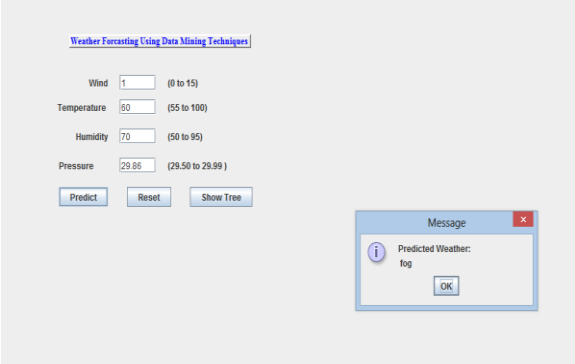
Temperature	Wind	Pressure	Humidity	weather
82.4	5	29.59	89	rain,thunderstorm
86	6	29.51	89	rain
86	7	29.55	83	rain
86	7	29.51	80	rain
84.2	12	29.65	80	rain
84.2	12	29.64	82	rain,thunderstorm
87.8	9	29.64	81	rain
86	5	29.55	74	haze
87.8	5	29.55	72	haze
89.6	5	29.57	68	haze
88	6	29.62	81	rain,thunderstorm
84	3	29.61	81	rain,thunderstorm
84	5	29.61	86	rain,thunderstorm
84	9	29.66	82	rain,thunderstorm
88	9	29.68	84	haze
89	12	29.61	75	rain
88	11	29.61	78	haze
86	6	29.69	87	rain
82	6	29.76	87	rain

Weather Results will be in the form of

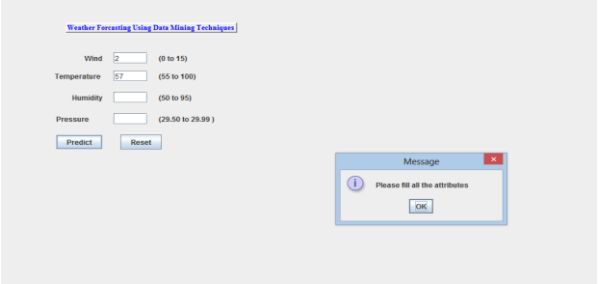
- Rain
- Rain, Thunderstorm
- Haze
- Haze, mist
- Fog

Weather Results using Decision Tree Approach:

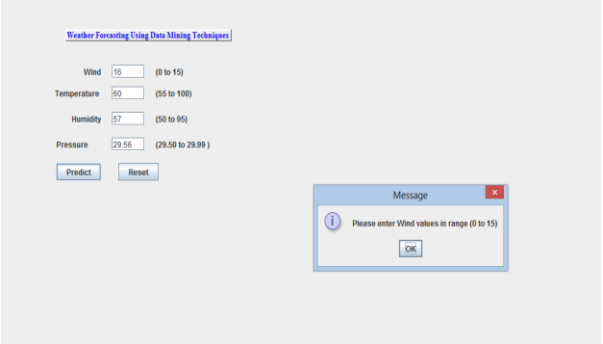
1. In Decision Tree approach, we have to specify the values of Temperature, Wind, Pressure, Humidity in user interface.



2. Every value must be filled in user interface.



3. If any value is given out-of-bound, the classification will not be successful and the error will be thrown.



Weather Prediction Results using Random Forest Approach:

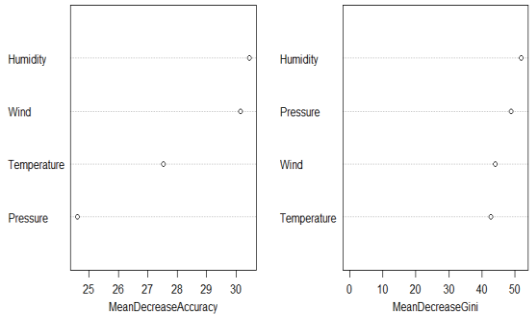
1. The training and testing data format is same for both Decision tree and Random forest. For random forest, it should be in csv format.
2. In Random Forest package, the function called randomForest is used to classify the data.

```
# Apply the Random Forest Algorithm
my_train1 <- randomForest(as.factor(weather) ~ T
```

3. To predict the labels of test data using classification model, predict function is used.

```
# Make your prediction using the test set1
my_prediction1 <- predict(my_train1, test1, type = "class")
```

4. Variable Importance and OOB(Out-of-Bag) error-rate is calculated for the prediction. Variable – importance plot is as follows:



In the above figure, we can see that Humidity has a highest importance followed by Wind.

5. Prediction result will be in the form of

weather
rain
rain
rain,thunderstorm
rain
rain
rain,thunderstorm
rain
haze
haze
haze
rain,thunderstorm
fog
rain,thunderstorm
rain,thunderstorm
rain
rain

VI. Conclusion and Future Scope

A decision tree building process and Random Forest is one of the most common data mining techniques. We can predict, with certain accuracy, the weather condition if we have data regarding some important aspects of the weather. For that we can use a decision tree built with ID3 algorithm. The working of this application is largely dependent on the daily estimates of weather and changes dynamically from time to time and season to season. This application is limited to only few places. The implementation here also states that the algorithm used for decision tree construction works well for the classification of weather data we have considered here.

Future work can include the extension of the database with other important weather parameters like dew point, wind direction or radiation. Besides this aspect, we can enlarge our database with records from other years and not just pertaining to an year. Having all this improvements in mind, we can increase the precision in building the decision tree and the weather prediction based on it. The appearance and the user interface can be further enhanced.

VII. References

[1] A Decision Tree for Weather Predictio- Elia Georgiana Petre.

[2] Application of Data Mining Techniques in Weather Prediction and Climate Change Studies - Folorunsho Olaiya.

[3] An Incremental Method for Finding Multivariate Splits for Decision Trees1 PAUL E. UTGOFF and CARLA E. BRODLEY Department of Computer and Information Science, University of Massachusetts, Amherst, MA 01003 U.S.A.

[4] An Implementation of ID3 --- Decision Tree Learning Algorithm Wei Peng, Juhua Chen and Haiping Zhou.

[5] Induction of Decision Trees by J.R. QUINLAN, Centre for Advanced Computing Sciences, New South Wales Institute of Technology, Sydney 2007, Australia.

[6] Medical Data Mining Based on Decision Tree Algorithm

Ruijuan Hu ,Dep. of Foundation ,PLA University of Foreign Languages, Computer and Information Science Vol. 4, No. 5; September 2011.

[7] Extension and Evaluation of ID3 – Decision Tree Algorithm. Anand Bahety.Department of Computer Science University of Maryland, College Park.

[8] International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 1 Issue 7, September – 2012 Horizontal partitioning ID3 algorithm A new approach of detecting network anomalies using decision tree by Sonika Tiwari, Prof. Roopali Soni.

[9] Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Morgan Kaufman Publications.

[10] CSE5230 Tutorial: The ID3 Decision Tree Algorithm MONASH UNIVERSITY Faculty of Information Technology.