

A Critical Review of Data Mining Techniques in Weather Forecasting

Sara khan¹, Mohd Muqem², Nashra Javed³

Assistant Professor, Department of Computer Application, Integral University, Lucknow, India^{1,3}

Associate Professor, Department of Computer Application, Integral University, Lucknow, India²

Abstract: Analyzing large amounts of data has become an urgent need. Data could be scientific, medical, demographic, financial or marketing. Main objective is to identify hidden information from the data sets. Data Mining is a technique that helps in extracting relevant and meaningful information from the set of data. It can be further described as knowledge discovery process that can be applied on any set of data. Data mining techniques when applied on relational databases can be used to search certain trends or patterns. Weather forecasting is an area of meteorology that is done by collecting dynamic data related to current state of weather like fog, rainfall, temperature, wind etc. It is one of the most challenging problems to analyze this set of data to predict the behavior of weather. This paper provides a survey of different data mining techniques being used in weather prediction or forecasting. It also reviews and compares various techniques being used in a tabular format.

Keywords: Data Mining, demographic, Weather Forecasting, Meteorology.

I. INTRODUCTION

Meteorological department worldwide is putting a great effort in the research areas of weather prediction. Since, India is an agriculture based country most of the people are dependent on the weather conditions. Large amount of India population depend on monsoon. It is always a matter of concern for the scientists to exactly forecast the weather conditions. Weather forecast includes prediction of rain, fog, winds, clouds, lightening, storm etc. One the biggest challenge in weather forecasting is its unpredictable and dynamic climate data sets, which can frequently change according to global climatic changes. Many techniques have been applied and suggested out of which data mining is considered as the most feasible approach towards Weather forecasting [1]. Data Mining has the capability to mine hidden patterns, relations and provide verification/validation of data sets based on certain input conditions.

II. WEATHER PREDICTION APPROACH

Weather prediction has always been a fascinating domain of research as changes in climatic conditions directly impact the population. [2] has classified the approach of weather prediction in two category, Empirical approach and Numerical approach. Empirical approach collects present weather conditions through ground observations i.e. observations from ships, satellites etc. This present climatic condition observations taken are then forwarded to meteorological centres where it is analysed and is converted into multidimensional maps through the use of computers. Scientist predicts the changes in the map that will occur over the certain time period. Second approach is dynamical approach or numerical approach that uses mathematical equations over climatic variables to solve the prediction.

III. DATA MINING TECHNIQUE IN WEATHER FORECASTING

Data Mining is a knowledge discovery approach in Database (KDD) [3]. Data Mining is a recent approach applied in the field of Meteorology. Meteorology is the scientific study of atmosphere. Meteorological Data Mining is the approach of Data Mining that helps in analyzing meteorological data. Weather is one of the meteorological data. Meteorological data can be collected through Radar, Lidar, and Satellites etc. This high dimensional data set collected must be simplified for correct analysis. Process of Data Mining includes Data Collection, Data Preprocessing and Knowledge Discovery.

A. Data Collection

There are two approaches for meteorological Data collection, Manual approach and Automatic approach [4]. Manual observations are recorded in a manuscript form at the observing Meteorological Centers (MCs). There are at present 530 surface observatories, 45 radiation observatories, 10 air pollution monitoring observatories, 62 Pilot Balloon observatories, 39 Radio sound and Radio wind observatories, 219 agro-meteorological observatories and more than 700 hydro-meteorological observatories. Meteorologist collect data from the highly efficient devices such as Barometer, Thermometers, Radar systems, Rain gauges, Wind vanes, hygrometers, transmissometer etc. Weather balloons are also used to measure conditions such as humidity, temperature, pressure etc. Automatic Approach includes Automatic Weather Station (AWS) and Automatic Rain Gauge (ARG) installed at various centers throughout India. This method is increasingly becoming the ideal method for collection of meteorological data world over. AWS and ARG collects data from over the country on different parameters like pressure, temperature etc and this data is stored centrally in Climate Data

Management System (CMD) situated in India Meteorological Department (IMD), Pune.

B. Data Pre-processing

Due to vast number of Automatic weather observatories being built up, meteorological data collected is huge. Data pre-processing deals with quality management of these data. Meteorological Data collected can be both real time and non real time data. Only correct, complete and irredundant data is kept, suspicious data with errors are eliminated or cleaned. Various quality improvement tests are done on CMD that ensures that data kept is free from errors, and is the valid data required.

C. Data Analysis/Knowledge Discovery

Data Mining is an unsupervised learning approach i.e. it doesn't require any training. The most commonly used techniques of data mining in weather forecasting are, Artificial Neural Networks, Genetic Algorithms, Clustering and Memory-Based Reasoning, Decision Trees, Logistic Regression, Discriminant Analysis and Decision Trees. Weather Forecasting is done based on certain Weather parameters like maximum temperature, minimum temperature, extent of rainfall, cloud conditions, wind streams and their directions, projected using images taken by the meteorological satellites. Data Mining helps in identifying or assesses future trends. Few of the major techniques of Data mining used in Weather forecasting are analyzed below.

i) Use of Artificial Neural Network (ANN) For Weather Forecasting;

Artificial Neural Network is a learning process. It is a system inspired by biological nervous system. It consists of set of neurons that are interconnected to one another by means of a link that has weight assigned to it. ANN system are configured for particular purpose like pattern recognition, or data classification through learning system. In Neural networks a set of data is processed as a small units called neurons which are connected and these connections has a weight. These units can be trained by using different training and learning methods. These learning methods can be supervised or inductive. Neural Network learns by adjusting the weights so as to be able to correctly classify the training data and hence after testing phase, is used to classify unknown data.

Following Neural network techniques are used in data mining: For Classification LVQ and Kohonen are used, For Forecasting/Prediction BP, GRNN, and RBF are used, for pattern detection Recurrent Neural Networks are used. ANN neuron model system works in three layers (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 [5]. ANN is said to be the most powerful technique in weather forecasting as it can handle both linear and non-linear data sets. ANN can be used in different ways as per the requirement in weather forecasting. Feed Forward

multi-layered artificial neural network model is used to estimate the maximum surface temperature and relative humidity. Prediction error is computed and compared for single layer network and one hidden layer artificial neural network [6]. In rainfall forecasting ANN is used in multiple forms, Multilayer feed forward Network, Time delay feed forward network and partial recurrent neural network. Kaur and Singh [7] in his work uses Multi-Layer Perceptron (MLP) to model forecasting system and used Back Propagation algorithm to train the network. The network is trained and tested with actual data of the past ten years which comes from meteorological department. The results show that minimum temperature can be predicted with reasonable accuracy by using the Artificial Neural Network. Hence given enough data ANN's can detect the relationships between weather parameter and use these to predict future weather conditions.

ii) Use of Clustering Methods For Weather Forecasting;

Clustering is a technique that combines similar or most similar data from various data sets into a group known as Clusters. It is very successful unsupervised learning technique. Clustering analyses data objects without consulting a known class Label. Known class Label is a predefined pattern. Cluster analysis techniques have been applied to the aim of detecting any significant spatial variation in the skill of numerical weather prediction models. Clustering is useful in grouping, pattern analysis, decision making etc.

Typical pattern clustering activity involves the following steps [Jain and Dubes 1988] (1) Pattern representation (optionally including feature extraction and/or selection), (2) Definition of a pattern proximity measure appropriate to the data domain, (3) clustering or grouping, (4) Data abstraction (if needed), and (5) Assessment of output (if needed). Pattern representation defines number of classes, parameters, type of features available to the clustering algorithm. Pattern Proximity is defined by distance function calculated between two patterns. Grouping can be done hard line or fuzzy. Various grouping algorithms are used for clustering of patterns like Partitioning, Hierarchical, Incremental, Agglomerative, etc [Jain and Dubes 1988]. Data Abstraction refers to providing a compact definition to the clusters. Making it simpler for analysis. Clusters assessment should be done both with the data sets that won't be having any clusters and data sets that will also form clusters. It should be validated on both type of data sets. k-means clustering algorithm can be used successively for interpreting the weather.

Clustering techniques can analyze and predict data of temperature, humidity values for future. The outlier analysis can be considered to detect the outliers with respect to the data and clustering analysis can be done to partition the data based on the similarity of the objects. K-nearest neighbor algorithm can be used to predict the values of temperature and humidity parameters of climate. To find the distances between datasets in the K-nearest neighbor algorithm absolute distance and Euclidean distance measuring are used.

iii) Use of Decision Tree Network (DTN) For Weather Forecasting;

A Decision Tree is a flow chart like tree structure. It is used to describe data and represent it in a tree structure that is used for decision making. They are used to classify data and over a single attribute of time [8]. It provides a modeling techniques that is easy for humans to understand and it further simplifies the classification problem. It systematically analyzes the information contained in a large data source to extract valuable rules and relationships. Compared to other data-mining techniques, it is widely used in weather forecasting since it is robust to data scales or distributions. Decision tree algorithm recursively partitions a data set of records using Depth-first greedy approach [9] or Breadth-first approach, until all the data items belong to a particular class being identified. A decision tree structure is made up of root, internal and leaf nodes.

In Weather Forecasting, different decision tree algorithms exist and these include: Alternating Decision Tree, Logitboost Alternating Decision Tree (LAD), ID3 (Iterative dichotomiser3), and Classification and Regression Tree (CART) [10]. Depending upon the requirement different algorithm is used accordingly. One advantage of Decision Tree classifiers is that rule can be inferred from the trees generated that are very descriptive, helping users to understand their data. See5 software can generate both decision trees and decision tree rules depending on selected options. Weather data sets can be grouped based on some clustering algorithm and further decision tree can be used for predicting the observations.

iv) Use of Genetic Algorithm (GA) For Weather Forecasting;

Genetic Algorithms are adaptive method derived from Darwin's principle of survival of fittest in general fitness. For interaction among attributes, genetic algorithms (GAs) are strong, flexible and better than most of the rule induction algorithms in a global search.

The reason is that GA applies a population of candidate solutions (individuals), and it evaluates them as a whole using fitness function. Genetic algorithms can help to select appropriate features and build an optimum decision trees. GA includes an optimization process, in which many combinations of features and their interactions are considered. Because GA searches for solutions efficiently in high dimensional and difficult response surfaces, it can be utilized for feature selection in a variety of problems and multivariate calibration in Weather Forecasting [11]. GA performs well in single dimensional data, while its performance on multidimensional data is not impressive.

IV. COMPARATIVE ANALYSIS : DATA MINING APPROACH IN WEATHER FORECASTING

Table I compare and analyses various data mining techniques used in whether forecasting on the basis of various parameters.

Weather prediction approaches are challenged by complex weather phenomena with limited observations and past data.

From the below comparison based on various parameters, ANN comes out to be the most powerful technique that can give accurate and timely result. It can handle complex real and non real values, has greatest computation speed, provide more accurate results, and also provide better results for multidimensional data sets.

TABLE I COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES IN WHETHER FORECASTING

Parameter	ANN	DTN	Clustering	GA
Performance	Maximum	High	Moderate	High
Supervised Learning	Yes	Yes	No	Yes
Unsupervised Learning	Yes	No	Yes	Yes
Computation Speed	Fast	Fastest	Slow	Fast
Cost effectiveness	Yes	Yes	Yes	Yes
Accuracy	More	Less	More	Moderate
Complexity level	Complex	Less	Moderate	Complex
Fault Tolerance	Yes	Yes	Yes	Yes
Nature	Analytical	Analytical	Descriptive	Analytical
Domain Area	Complex	Simple	Complex	Complex
Parallel Computing	Yes	Yes	No	Yes

V. CONCLUSION

Weather Forecasting is one of the most technologically and scientifically challenging problems around the world. Entire Weather prediction relies on correct collection of meteorological data, and appropriate selection of data mining techniques. One of the major concerns of Weather prediction is its Accuracy and its Timely output. Since, Problem domain of Weather Forecasting is vast it is feasible to use data mining techniques that performs well with complex problem domain, give most accurate results and has higher evaluation speed. More than one data mining technique is however applied in parallel for better and accurate results.

REFERENCES

- [1] Ahrens, C. D., "Meteorology", Microsoft Corporation, 2007.
- [2] "Applications of Data Mining Techniques in weather prediction and climate change studies",MECS.DOI:10.5815/ijieeb,2012
- [3] "Introduction to Data Mining and Knowledge Discovery", Two Crowds Corporation,<http://www.twocrows.com/introdm.pdf>, 2009.
- [4] A K Jaiswal, N M Narkhede, Rachel Shaji "Atmospheric Data Collection, Processing and Database Management in India Meteorological Department India" Meteorological Department, India
- [5] Due R. A., "A Statistical Approach to Neural Networks for Pattern Recognition", 8th edition. New York: John Wiley and Sons publication, 2007.
- [6] M. M. Ali, P. S. V. Jagadeesh, I.-I. Lin, Je-Yuan Hsu, "A Neural Network Approach to Estimate Tropical Cyclone Heat Potential in the Indian Ocean" IEEE Geosciences and Remote Sensing Letters, Vol. 9, No. 6, Nov 2012, pp 1114-1117
- [7] Amanpreet Kaur, Harpreet Singh, "Artificial Neural Networks in Forecasting Minimum Temperature", International Journal of Electronics and Communication Technology, Vol. 2, Issue 3, pp 101-105, Sept. 2011
- [8] M. J. Berry and G. S. Linoff, 2000, "Mastering data mining", New York: John Wiley & Sons.
- [9] W. J Brawley, G. Piatetsky-Shapiro, and C. J Matheus, "Knowledge Discovery in data bases an overview", MIT Press, 1991
- [10] Rokach, Lior, Maimon, O, "Data mining with decision trees: theory and applications". World Scientific Pub Co Inc. ISBN 978-9812771711.
- [11] A.K. Jain, M.N. Murty, P. J. Flynn, "Data clustering: a review", ACM Computing Surveys (CSUR), Vol.31, Issue 3, 1999.