Weather Prediction Using Data Mining

[1]Prashant Biradar, [2]Sarfraz Ansari, [3]Yashavant Paradkar, [4]Savita Lohiya

[1,2,3]Student, Department of Information Technology, SIES Graduate School of Technology, Nerul, Navi Mumbai, India
[4]Professor, Department of Information Technology, SIES Graduate School of Technology, Nerul, Navi Mumbai, India

Abstract*— Weather forecasting is the application of science and technology to predict the state of the atmosphere for a given location. Here this system will predict weather based on parameters such as temperature, humidity and wind. This system is a web application with effective graphical user interface. To predict the future's weather condition, the variation in the conditions in past years must be utilized. The probability that it will match within the span of adjacent fortnight of previous year is very high .We have proposed the use of K-medoids and Naive Bayes algorithm for weather forecasting system with parameters such as temperature, humidity, and wind. It will forecast weather based on previous record therefore this prediction will prove reliable. This system can be used in Air Traffic, Marine, Agriculture, Forestry, Military, and Navy etc.

Index Terms*— Data mining, Weather prediction, Weather forecasting.

I. INTRODUCTION

Weather forecasting is mainly concerned with the prediction of weather condition in the given future time. Weather forecasts provide critical information about future weather. There are various approaches available in weather forecasting, from relatively simple observation of the sky to highly complex computerized mathematical models. The prediction of weather condition is essential for various applications. Some of them are climate monitoring, drought detection, severe weather prediction, agriculture and production, planning in energy industry, aviation industry, communication, pollution dispersal, and so forth. In military operations, there is a considerable historical record of instances when weather conditions have altered the course of battles. Accurate prediction of weather conditions is a difficult task due to the dynamic nature of atmosphere.

The weather condition at any instance may be represented by some variables. Out of those variables, one found that the most significant are being selected to be involved in the process of prediction. The selection of variables is dependent on the location for which the prediction is to be made. The variables and their range always vary from place to place. The weather condition of any day has some relationship with the weather condition existed in the same tenure of precious year and previous week. Rainfall is a form of precipitation. Its accurate forecasts can help to identify possible floods in future e and to plan for better water management. Weather forecasts can be categorized as: Now forecasts which is forecasts up to few hours, Short term forecasts which is mainly Rainfall forecasts is 1 to 3 days forecasts, Forecasts for 4 to 10 days are Medium range forecasts and Long term forecasts are for more than 10 days. Short range and Medium Range rainfall forecasts are important for flood forecasting and water resource management.

1.1 Data Mining or knowledge discovery is process of finding facts which are not known. Classification is a supervised learning process which lies under the umbrella of Data Mining. It is used as model to distinguish samples with unknown class labels on the basis of their similarities and dissimilarities and predict a class label for them. Classification has been applied in many fields like for detecting frauds by banks and companies, by various service providers to predict their performance in future, to classify patients on the basis of their symptoms. The Naive Bayes approach and K-Medoids has been used in this project to forecast the Weather.

1.2 Bayesian approach for classification is a statistical and linear classifier which predicts class label for data instance on the basis of distribution of attribute values. This is a parametric classification where the size of classifier remains fixed.

1.3 The k-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoid shift algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers (medoids or exemplars)and works with an arbitrary metrics of distances between datapoints.

II. SYSTEM MODEL

A data flow diagram (DFD) is a graphical representation of the flow of data through an information system. A data flow diagram can also be used for the visualization of data processing (structured design). It is common practice for a designer to draw a context-level DFD first which shows the interaction between the system and outside entities. This context-level DFD is then exploded to show more detail of the system being modeled.

The four components of a data flow diagram (DFD) are:

- External Entities/Terminators are outside of the system being modeled. Terminators represent where information comes from and where it goes. In designing a system, we have no idea about what these terminators do or how they do it.
- Processes modify the inputs in the process of generating the outputs
- Data Stores represent a place in the process where data comes to rest. A DFD does not say anything about the relative timing of the processes, so a data store might be a place to accumulate data over a year for the annual accounting process.
- Data Flows shows how data moves between terminators, processes, and data stores (those that cross the system boundary are known as IO or Input Output Descriptions).

Figure 2.1and 2.2 represent the Level 0 and Level 1 Data Flow Diagrams respectively.

Context level Data Flow Diagram:

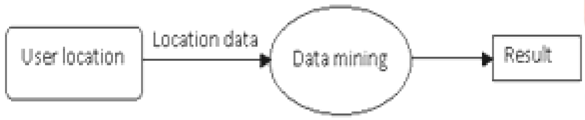


Figure 2.1

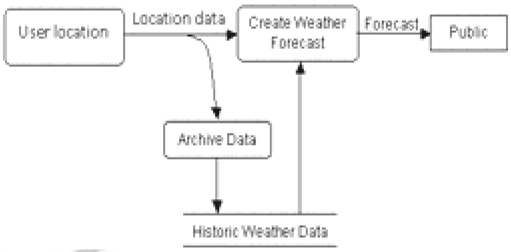


Figure 2.2

III. PREVIOUS WORK

Many works have done in forecasting of Temperature & Pressure. Some of them are presented here. This will lead us to make better understanding of project work. Some basic concepts, findings & facts will be extracted to make some conventions for our project.

Piyush Kapoor and Sarabjeet Singh Bedi [1] used sliding window approach and it has been found to be highly accurate except for the months of seasonal change where conditions are highly unpredictable. The results can be altered by changing the size of the window. Accuracy of the unpredictable months can be increased by increasing the window size to one month.

Jose L. Aznarte and Nils Siebert [2], a dynamic line rating experiment is presented in which four machine learning algorithms (Generalized Linear Models, Multivariate Adaptive Regression Splines, Random Forests and Quantile Random Forests) are used in conjunction with numerical weather predictions to model and predict the ampacity up to 27 hours ahead in two conductor lines located in Northern Ireland.

Deepti Gupta and Udayan Ghose [3], weather factors including mean temperature, dew point temperature, humidity, pressure of sea and speed of wind and have been used to forecasts the rainfall.

Mohsen Hayati & Zahra Mohebi [4] utilizes ANN for one day ahead prediction of temperature. They used MLP to train & test ten years (1996-2006) meteorological data. For accuracy of prediction they split data into four seasons and then for each seasons one network is presented. Two random unseen days in each season are selected to test the performance. The error in result varies between 0 to 2 MSE.

S.S. De [5] used ANN to forecast the Max. & Min. Temperature for Monsoon month. The temperature of June, July & August has been predicted with the help of January to May temperature. The data of three months of 1901 to 2003 is used. The ANN model generated here is a single hidden layer model with 2 nodes at hidden layer. After 500epochs the result is validates. The Max. Error appeared is 5%.

IV. PROPOSED METHODOLOGY

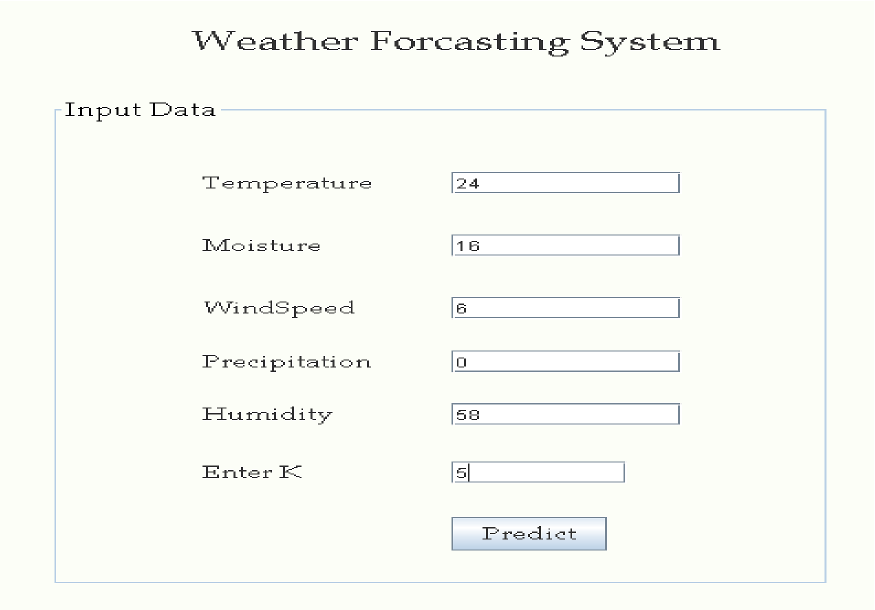
- **Naive Bayes:**

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

- **K- Medoids:**

The *k*-medoids algorithm is a clustering algorithm related to the *k*-means algorithm. Both the *k*-means and *k*-medoids algorithms are partitional (breaking the dataset up into groups). *K*-means attempts to minimize the total squared error, while *k*-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster.In contrast to the *k*-means algorithm, *k*-medoids chooses datapoints as centers (medoids or exemplars).

V. EXPERIMENTAL RESULTS



Weather Forcasting System

Input Data

Temperature 24

Moisture 16

WindSpeed 6

Precipitation 0

Humidity 58

Enter K 5

Figure 5.1 - Table used to store various parameters

The work proposes to predict a day's weather conditions. For this the previous seven days weather is taken into consideration along with fortnight weather conditions of past years.

VI. CONCLUSION

We conclude that using Data mining techniques for weather prediction yields good results and can be considered as an alternative to traditional metrological approaches. The study describes the capabilities of various algorithms in predicting several weather phenomena such as temperature, rainfall and concluded that major techniques like decision trees, clustering and regression algorithms are suitable to predict weather phenomena. A comparison is made in this project, which shows that decision trees and k-medoid clustering are best suited data mining technique for this application.

VII. ACKNOWLEDGMENT

We are very grateful to our guide Mrs. Savita Lohiya of SIES GST for her support and encouragement to write this paper.

REFERENCES

[1]. Piyush Kapoor and Sarabjeet Singh Bedi, Weather Forecasting Using Sliding Window Algorithm, ISRN Signal Processing Volume 2013, Article ID 156540.

[2]. Jose L. Aznarte and Nils Siebert, Dynamic Line Rating Using Numerical Weather Predictions and Machine Learning: a Case Study, IEEE Transactions on Power Delivery (Volume: PP, Issue: 99).

[3]. Deepti Gupta and Udayan Ghose, A Comparative Study of Classification Algorithms for Forecasting Rainfall, IEEE 978-1-4673-7231-2, ©IEEE Publications 2015.

[4]. Mohsen Hayati & Zahra Mohebi, Temperature Forecasting Based on Neural Network Approch, WorldApplied Science Journal 2(6) 613-620, 2007, ISSN 818-4952 ©IDOSI Publications 2007.

[5]. S.S. De, University of Kolkata, Artificial Neural Network Based Prediction of Max. & Min.Temperature in the Summer-Monsoon month over India, Applied Physics Research,Vol.1,No.2,Nov-2009.