Implementation and Analysis of Data Mining Techniques for Weather Prediction

Nikita Gupta  
Pimpri Chinchwad College of Engineering  
Rashmi Narayanan  
Pimpri Chinchwad College of Engineering  
Anagha Chaudhari  
Pimpri Chinchwad College of Engineering

ABSTRACT

Data Mining is the process of discovering new patterns from large data sets, this technology. Some efficient techniques can be developed for solving complex problems using data mining which is employed in inferring useful knowledge that can be put to use from a vast amount of data. Meteorological data mining involves finding hidden patterns in available meteorological data and transforming it into usable knowledge. In this study, we have focused on three main algorithms including Time-series, K-means and Naive Bayes Classifier which help in prediction. Meteorological data analysis is a system which considers real time data while making predictions and giving out weather forecasts. It should be scalable, portable and should work on variety of client systems. We propose a solution to this by using R programming language for analysis of weather data using Microsoft Azure HDInsight for good long term predictions.

KEYWORDS
Data Mining Techniques, Weather data, Meteorological data, Time-Series, Naïve Bayes, K-means clustering

1. INTRODUCTION

Data mining is employed in inferring useful knowledge that can be put to use from a vast amount of data, various data mining techniques such as Classification, Prediction, Clustering and Outlier analysis can be used for the purpose [9]. Climate has a significant impact on the human society. Weather data or climate data in a region is essential for business, society, agriculture and energy applications [1]. The main aim of the study is to overview on data mining process for weather data and to study on weather data using data mining technique. By using clustering, Naïve Bayes and K-means technique we can acquire weather data and can find hidden patterns inside the large dataset so as to transfer the retrieved information into usable knowledge for classification and prediction of climate condition [9].

2. THRUST AREA: AGRICULTURE

- Climate affects the human society in all the possible ways. Knowledge of weather data or climate data in a region is essential for business, society, agriculture and energy applications.
- The understanding can be used to support many important sectors that are affected by climate like agriculture, vegetation, water resources and tourism. Using data mining techniques we can develop a system that uses numeric historical data to forecast the climate of a specific region or city.
- This study provides, long-range forecasts to accurately predict future weather useful for agriculture.
3. DIFFERENCE BETWEEN FORECAST AND PREDICTION

**PREDICTION**: An estimation of any event happening (in past, present or future).

**FORECASTING**: It is always associated with a time dimension in the future i.e estimation for some specific future duration or over a period of time.

For Forecasting, use of R is highly beneficial as it is the statistical programming language used mainly for prediction and forecasting purpose. In R, we will be using three algorithms:

### 3.1 TIME SERIES

Time series is used for understanding the trend in the historical dataset. Time series is a series of data points in which each data point is associated with a timestamp. A simple example is the price of a stock in the stock market at different points of time on a given day. Another example is the amount of rainfall in a region at different months of the year [10].

![Time-series plot](image)

Fig. 1: An example of Time-series plot

R language uses many functions to create, manipulate and plot the time series data. The data for the time series is stored in an R object called **time-series object**. It is also R data object like a vector or data frame. Using the `ts()` function the time series object is created.

### 3.2 NAÏVE FORECAST

Output of time series is provided as an input to this function. Gives forecasts ranges depending on the value of the number of years for which you want to forecast. It computes the conditional a-posterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule [12]. In naïve Bayes classifiers it is assumed that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence.

Bayes’ rule states that

\[
P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}
\]

where \(P(A|B)\) is defined as the probability of observing A given that B occurs. \(P(A|B)\) is called posterior probability, and \(P(B|A), P(A)\) and \(P(B)\) are called prior probabilities.

### 3.3 K-MEANS CLUSTERING.

Clustering to partition the data into classes that can provided for classification techniques [1]. K-Means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining.

**Input:**

1. \(K\): the number of clusters,
2. \(D\): a data set containing \(n\) objects.

**Output:** a set of \(k\) clusters.
Method:

1. Arbitrarily choose k objects from D as the initial cluster centers;
2. Repeat
   3. (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
   4. Update the cluster mean;
   5. Until no change;

4. SYSTEM ARCHITECTURE

As a first step, by uploading dataset on Azure through azure blob storage from local file system [2]. Next, with R for accessing tables and operation. Once the connection is established, we can read dataset into R and then perform operations and analysis on R.

We will be using three algorithms in R for forecasting purpose. The very first and important algorithm is time series that is used for analyzing the trend in the given data its object known as time series object is used as input for other algorithms. Next is Naïve method used for forecasting of weather based on the input of years. Then, after naïve we will be doing clustering and classification so that final output is understandable by user [2].

4. CONCLUSION

- As MS Azure is a recent platform, any new R&D is a surge to giving customers new services in a user friendly way and on a software like Microsoft.
- Therefore now this system would give predictions for datasets that show basic trend in its behavior.
- It will give good weather prediction system with good accuracy and faster processing with the three algorithms i.e Time series, K-Means (), Naïve Bayes.
In our proposed system, we have provided a solution to a problem that says we have various weather mobile applications that have been developed using clustering and regression algorithms, but real time analysis is still a big challenge. Also more accurate prediction of weather on dataset is required.

5. FUTURE WORK

- API of dataset can be created.
- Can work on Linux OS providing Linux cluster is available.
- It can be used for various smart city applications.
- Giving forecast by identifying the users location.

REFERENCES

[1]. Meghali A. Kalyankar and Prof. S. J. Alaspurkar , “Data Mining Technique to analyse the Metrological Data”, IJARCSSE VOLUME 3, ISSUE 2, FEBRUARY 2013
[12]. https://www.rdocumentation.org