# The use of data mining technique and the uses of data to predict the weather

Technology

## ABSTRACT

In our daily life, weather is playing an important role. Weather forecasting provides the business with valuable information. Weather prediction can have significant impact on various sectors of society like Agriculture, Aviation, Broadcast Media, Chemical & Petroleum, Cross-Industry, Financial Markets, Government, Ground Transportation, Insurance, Media & Entertainment, Construction industry, Retail and Telecommunications. Accurate weather prediction is one of the most challenging problems around the globe. This work uses Clustering, Classification and Regression techniques. It had been tried to extract useful practical knowledge of weather data on monthly based historical analysis. It also presents the review of data mining techniques MLR, K-Mean, KNN for weather prediction and studies the benefit of using it.

## INTRODUCTION:

Weather prediction has one of the most crucial things around the world. The prediction of weather conditions can have significant impacts on various sectors of society in different parts of the country. Forecasts are used by government and industry to protect life, property and also to improve the efficiency of operations and by individuals to plan a wide range of daily activities. The notable improvement in forecast accuracy has been achieved since 1950s, that is a direct outgrowth of technological developments, basic and applied research and the application of new knowledge and methods by weather forecasters. The advance knowledge of weather parameters in a particular region is advantageous in effective planning. Several studies on forecasting weather variables based on time series data in reference to a

particular region have been carried out at national and international level both in the farm and non- farm sectors. It has been one of the most interesting and fascinating domain. The scientists have been trying to forecast 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.

One of the major challenges facing meteorologist around the world is to make an accurate prediction of weather. The weather data humidity, Pressure, Wind Speed, Wind Direction, Visibility, Temperature, Dew Point, Precipitation, Sunshine, Rainfall, Clouds Quality, Snow depth, Relative humidity, Radiation and Gust are observed by radiosondes are launched all over the world approximately and the information transmitted to the ground station. Also, the surface weather measurements are made at observing stations around the world, from ships and buoys at sea, commercial aircraft, weather radars and satellites. All these measurements are transmitted to different centers. These centers have very fast supercomputers they have programmed with equations to describe the atmosphere changes at every point. All of these equations run as a computer program are called a numerical weather forecast model.

The increasing availability of climate data during the last decades observational records, radar and satellite maps, proxy data, etc., makes it important to find effective and accurate tools to analyze and extract hidden knowledge from this huge data. 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. In turn, this understanding can be used to support many important sectors that are affected by climate.

## LITERATURE REVIEW

Guhathakurata [2006] – Weather is a continuous, data-intensive, multi-dimensional, dynamic and chaotic process, and these properties make weather forecasting a formidable challenge. It is one of the most imperative and demanding operational responsibilities carried out by meteorological services all over the world. At present, the assessment of the nature and causes of seasonal climate variability is still conception.

Sivakumar et al. [1999] – The field of meteorology all decisions are to be taken in the visage of uncertainty associated with local of and global climatic variables. Several authors have discussed the vagueness associated with the weather systems. Chaotic features associated with the atmospheric phenomena also have attracted the attention of the modern scientists.

Dilip C and Dr. K Thippeswamy [2016] – In the first step the generating local clusters on individual nodes will be done and in the second step local clusters are aggregated to form a global module. From several clusters some of the clusters acts as leaders, these leaders will do merging of local clusters into global one using overlay technique. This technique is continued until a resultant cluster is obtained. Distributed dynamic clustering are deals with

very large scale, distributed and heterogeneous datasets. The communication overhead is minimized by reducing the size of the dataset which is going to exchange between the systems. By using K-means algorithm local clusters are generated and analyzed. During aggregation the local clusters are merged and produce an ultimate accurate output.

Pinky Saikia Dutta, Hitesh Tabilder [2014] – Data mining techniques is used to predict the monthly rainfall of Assam. This carried out using traditional statistical technique Multiple Linear Regression. Regression model which contain more than two predictor variables are called Multiple Linear Regression. The period of 2007-2012 data collected from regional meteorological centre Guwahati. The model consider maximum temperature, minimum temperature, wind speed, mean sea level as predictors 63% accuracy in validation of rainfall for proposed model. The model can predict the monthly rainfall.

A. R. W. M. M. S. C. B. Amarakoon [2010] – Proposed a system that uses the authentic weather data and applies the data-mining calculation “ K-Nearest Neighbor (KNN)” for grouping of these chronicled data into a particular time traverse. The k closest time ranges is then additionally taken to anticipate the weather of Sri Lanka. The everyday weather data is gathered for finish one year. It creates exact outcomes inside a sensible time for a considerable length of time ahead of time. It is inferred that KNN is valuable to dynamic data, the data that progressions or updates quickly and gives better execution when contrasted with alternate procedures. Coordinating component choice strategies can even give more precise outcomes.

## DATA MINING IN METEROLOGY

Meteorological data mining is a form of data mining concerned with finding pattern inside largely available meteorological data. Weather can predict metrological parameters by using various techniques in data mining. While some of these algorithms are more exact prediction than others. In the recent years, the accessibility of weather data has been expanded. Such sources of weather data like observational records, understudy data, etc. makes it more crucial to find tools with higher accuracy rate to analyze different patterns from vast amount of data. Therefore, meteorological data mining is a type of mining which is concerned with finding hidden patterns inside massive data available.

So, the information extracted can be transformed into practical knowledge. This knowledge plays a vital role to predict the future of the weather. Having Knowledge of meteorological data is the key for variety of application to perform analysis and prediction of weather condition and it also does good job for prediction of temperature, humidity and irrigation system. These databases can become valuable information for analysts who use to perform different operations on this data. It requires higher scientific techniques like machine learning application for effective study and prediction of weather condition. This work is using K- Means, K- Nearest Neighbor and Multiple Linear Regression.

## STEPS FOR DATA MINING PROCESS

“ Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques.

Data mining is a knowledge discovery process of extracting previously unknown, actionable information from very large databases" (Moxon1996).

## A. Data Collection

Data collection is the systematic approach to gathering and measuring information from a variety of sources to get a complete and accurate data. Data collection enables evaluate outcomes and make predictions about future. One of most important and can't do anything without the dataset for this analysis. The Water underground website maintains the historical data. The data are between Elev 52ft 13 °N, 80. 18 °E that roughly covers the Chennai city, Tamil Nadu.

B. Data Pre- Processing

The data pre-processing challenge is knowledge discovery process in temperature, humidity, dew point and pressure data is poor quality. Relevant data may not be recorded due to misunderstanding or because of equipment faulty. The weather data is used which its having various parameters like Gust, Wind, Visibility, precipitation, Events, Temperature, Dew point, humidity and etc., here pre-processing means removing unwanted parameters from the dataset and remove noisy data. The following steps are,

## Data Cleaning

It is also known as data cleaning, in this phase noise data and irrelevant data are removed from the collection. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data. It's used for to improve the data quality.

## Data Transformation

Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system. The collected weather parameters are usually in Excel format it’s converted into comma separated values (CSV) file.

## Data Integration

Data integration combines data from multiple sources to form a coherent data store. The resolution of semantic heterogeneity, metadata, correlation analysis, tuple duplication detection, and data conflict detection contribute to smooth data integration. In this humidity and temperature analysis, weather data collected from one website because the dew point, wind speed, humidity and temperature parameters are available in www. wunderground. com this site.

## Data Reduction

It includes data cube aggregation, attribute subset selection, dimensionality reduction and discretization can be used to obtain a reduced representation of data while minimizing the loss of information’s content.

## Discretization

Data Discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Concept hierarchies can be used to reduce the data by collecting and replacing low-level with high-level concepts.

## Knowledge Discovery

For knowledge extraction various data mining techniques such as Clustering, Classification and Regression can be applied in Weka tool.

## Result of analysis

The future value of temperature and humidity predicted depending on the result of

K- Mean, K- Nearest Neighbor and Multiple Linear Regression algorithms.

## Data Mining Techniques

Data mining is a crucial analytic process indeed to explore data, the most imperative task in data mining is to extract non- trivial nuggets from vast amount of data. The dataset describes which contains details of various parameters about the weather. Temperature and humidity used for analysis of future weather condition. There are several major data mining techniques are Clustering, Classification, Outlier analysis, Association and Correlation. This work carried out using the three algorithms are,

## clustering

Clustering is an unsupervised learning algorithm and it deals with finding a structure in a collection of unlabeled data. “ The process of organizing data points into groups which is similar in some way”. The similar data points are one cluster and dissimilar data points are another cluster.

## Centroid based clustering

Centroid based clustering also called K- Means clustering. Clustering can uncover previously undetected relationships in a dataset. There are many applications for cluster analysis and an important issue in k- means

clustering to determine the similarity between two objects, so that clusters can be formed from objects with high similarity between clusters. Commonly, to measure similarity or dissimilarity objects, a distance measured by Euclidean Distance.

## Algorithm

The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster. Steps for working of k- means clustering,

STEP 1: Pick ' K' random points as cluster centers called centroids

STEP 2: Assign each xi to nearest cluster by calculating its distance to each centroid

STEP 3: Find new cluster center by taking the average of the assigned points

STEP 4: Repeat Step 2 and 3 until none of the cluster assignments change

## Input:

k: the number of clusters,

D: a data set containing n objects.

Output: A set of k clusters.

## CLASSIFICATION

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels.

**K- Nearest Neighbor**

**Nearest-neighbor classifiers are based on Lazy Learning function, a lazy learner simply stores and waits until it is given a test tuple. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. When given an unknown tuple, a k-nearest neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k " nearest neighbors" of the unknown tuple. " Closeness" is defined in terms of a distance metric, such as Euclidean distance. Steps for working of K- Nearest Neighbor algorithm is,**

**STEP 1: Determine parameter K= number of nearest neighbors**

**STEP 2: Calculate the distance between the query instance and all the training samples**

**STEP 3: Sort the distance and determine nearest neighbors based on the K-th minimum distance**

**STEP 4: Gather the category y of the nearest neighbors**

**STEP 5: Use simple majority of the category of nearest neighbors as the prediction value of the query instance.**

**Regression**

Regression is a data mining function that predicts a number like Profit, sales, mortgage rates, house values, square footage, temperature, or distance could all be predicted using regression techniques. For example, a regression model that predicts temperature values could be developed based on observed data for temperature over a period of time.

**Multiple Linear Regression**

Regression model which contain more than two predictor variables are called Multiple Regression Model. Equation of the Multiple linear regression model is,

$$Y= b0+b1x1 +b2x2 +b3x3+ b4x4+...e \quad (1)$$

Where,

**Y is Dependent variable**

b0, b1, b2, b3, b4 are Regression Coefficient

x1, x2, x3, x4 are predictor or regressor or explanatory variables

Multiple linear regression fits a model to predict a dependent (Y) variable from two or more independent (X) variables. The predictors can be understood as independent variables and the target as a dependent variable. The error, also called the residual, is the difference between the expected and predicted value of the dependent variable. The regression parameters are also called as regression coefficients.

Weather Data System Architecture

## Analysis and Result

The Weather parameters max, Min, Avg Humidity, Max, Min, Avg Temperature, Pressure and Dew point are fed into K- Means, K- Nearest Neighbor and Multiple Linear Regression algorithms using weka tool. January temperature data file fed in k- Means algorithm.

## Conclusion and Future Work.

Climate affects the human society in all the possible ways. A reliable weather forecast can help many sectors like Agriculture, Aviation, Broadcast Media, Insurance, Media & Entertainment, Construction industry and so on. Analysis on weather data describes the use of data mining technique and the uses of historical data to predict the weather in a particular region or city. This work makes use of Classification and clustering technique to predict weather for a month in a particular region with past data set. After applying clustering and

classification for weather prediction KNN is most feasible then other techniques. Future work, In future dynamic data mining methods can be used to predict nature, rapid changes and sudden events with dynamical data set. We can enlarge the database with other important attributes.