



Weather Damage Prophecy using Big-Data Analytics

¹G.Shyam Prasad, ²Sudheer Kumar Battula, ³CH. Divya Sruthi, ⁴A. Shruthi*, ⁵P.Bhagavathi and ⁶P.Akhila

Information Technology ACE engineering college
Hyderabad, AP, India

¹gscprasad@gmail.com, ²sudheer.itdict@gmail.com, ⁴sruthicheripally@gmail.com

Abstract: In today's world the prediction of weather is a challenging task. Weather prediction is the application of both science and technology in order to predict the weather for a particular location. Predictions include many parameters that are to be measured which are impossible with limited observations. The present predicting results are less accurate and the damage caused by it is not estimated because the time span between present moment and time for which forecasting is being made and released is varied. In order to find solution to these problems, we have come up with an algorithm using which would predict the scope of damage caused by the weather.

Keywords: KNN, prophecy, weather, big-data

I. INTRODUCTION

This paper aims at predicting damage caused by weather conditions such that people can have a brief knowledge about the weather conditions outside and make them aware of consequences. The predictions can be classified in to two types:

- Empirical approach: This approach used to scale the local weather conditions.
- Dynamic approach: It is used for large-scale measures.

The earlier researchers used data mining techniques where they used clustering and decision tree techniques by using which they predicted the future weather conditions. With the advent of advanced technology, we would like to predict the damage caused by weather using proposed algorithm with the help of big-data analytics.

II. EARLIER HISTORY

The forecasting of weather has begun in early civilizations to monitor the seasonal change using reoccurring astronomical and meteorological events. According to "NASA" facts forecasting weather has started from earlier ages which are explained in detail i.e. Chinese astronomer developed a calendar which includes 24 festivals with a difference based on weather conditions in 300 BC, at that time people did not know that the festivals can be celebrated based on the weather changes and a Greek philosopher Aristotle wrote meteorological, a philosopher treatise which includes theory about rain etc of earlier 340 BC whose theories were used for 2000 years, but these ideas were over thrown in 17th century and also Babylonians tried to predict the weather changes based on the appearance of clouds in 650 BC. Later on in the mid 15th century Nicholas-Cusa invented hygrometer which included properties to measure the parameters like moisture, temperature, pressure etc. In 1653 Torricelli introduced barometer which was helpful in finding the atmospheric pressure, from 17th to 19th century the refinement took place which lead to the invention of telegraphs. In mid of 19th century using the same crude weather maps were drawn.

As the years passed, synoptic forecasting based on analysis of observations taken over a wide area, in 1860's. During 1920 monitoring weather at high altitude was made with the invention of 'Radiosonde' which is filled with He or H₂ when increases to 30 kilometers indicated that it would burst. These were launched in to the space from across hundreds of ground stations for every 12 hours those days.

During 1862-1951 a Norwegian named Wilhelm Bjerknes presented an idea of predicting weather using some mathematical equations. In April 1950, for the first time Chaney's group made an attempt to record the weather forecasting for 24 hours.

III. LITERATURE SURVEY

a. KNN technique:

In spite of the fact that one cannot expect the accuracy of these estimations. Frequently weather predictions are used for forecasting future days in advance, but raise in technology slowly changed it from few days to yearly forecast.

According to this, Computer models of Sri Lanka weather may be statistical, uses the past data to expect the weather conditions by equating them to the sea and land surface conditions, on the other way, dynamical models calculates and differentiates atmospheric conditions with the help of laws present in physics

The aim of this paper is to use a data mining technique, i.e., "K-Nearest Neighbor (KNN)", to develop a Methodology which uses prior data to predict the climate, in advance, of a particular region (Sri Lanka). In this survey we came to know that, predicting weather is a prominent field.

Here in this research the researchers are applying KNN algorithm which is based on Euclidean distance formula, that is help full in finding out the closeness between the unknown samples whatever then mapped to the most common class in its k nearest neighbors

KNN is mostly used for estimation and prediction which categorizes the unknown data samples in to a class based on prior classified samples. KNN is called as Lazy Learner, because it performs process of learning at the

time when new sample is to be categorized, the KNN therefore requires more data processing techniques than eager learner technique which is nothing but classifying the samples before new sample is to be categorized. The KNN is mostly advantageous over dynamic data as it changes rapidly. In order to measure the distance between nearest neighbors Euclidean distance is the commonly used technique. All the values of distance obtained are then arranged so that the samples with the smallest distance to the new sample are known (estimated).

The classification of data is done depending upon whether it is continuous or categorical, based on which the mean or its average is used.

b. KNN algorithm:

- a) Firstly the distance (Euclidean) between the new data and the data to be trained is measured.
- b) Now the distance is sorted and the smallest value is selected.
- c) According to the application mean is calculated.

1. Euclidean distance calculation:
 Algorithm ED (No_of_Seq., Size_of_Seq.)
 While i < Total_element
 Do
 For j = 0 to Size_of_Sequence
 Sum = Sum + (Power (Base Sequence [j] – All Sequence [i]),2)
 i++
 End For
 Distance = Sqrt (Sum)
 End While

2. Sorting the distance and index array:
 Algorithm Sort (Distance Array, Index Array, Size)
 For i = 0 to size-1
 For j = 0 to (size- i) – 1
 If Distance[i] > Distance Array [i+1]
 Then Swap Distance Array[i]
 And Distance Array [i+1] Swap Index Array[i] and Index Array [i+1]
 End If
 End For
 End For

3. Applying k-means:

After the completion of sorting out of the value stored in the array index is obtained and values are summed up accordingly and then the mean or average of them are used for prediction.

In this survey it is observed that the performance of KNN is comparatively better than any other data mining techniques.

In this investigation we were able to observe two points : one of them is that they are majorly focusing on future weather and the other is they are not much concentrating on the weather parameter rain .Hence it is clear that they failed in calculating the damage caused by the weather specifically rain .

c. Decision tree:

Classification is an important function in data mining. Firstly the researchers created a model, and then it is trained by a sample of data, which is helpful in instructing them. Now the data which is unknown and to be tested is provided to the trained data for predicting the weather. So, that it might be helpful in obtaining accuracy.

Decision tree one of the data mining techniques, is frequently used for mining data. The primary aim is to generate a model which is efficient in predicting the value of the parameter that is to be subjected with the help of various parameters. The tree is easy to understand, by splitting the input data into number of subsets, based on the value to be tested and this is repeated as long as the desired subset of data sets are obtained which is called as recursive partitioning and it is said to be completed when the subset at a node has the value equal to the target variable. This process of top-down approach in decision trees is an example of a greedy algorithm.

Decision trees might be a combination of mathematical and computational techniques. In this technique from the independent variable, a dependent variable could be predicted. Decision trees are generally of following types.

- (a). Classification tree analysis: It is used in prediction of data in a class.
- (b). Regression tree analysis: It is useful in predicting the independent variable as a unit of number.
- (c). The Classification and Regression Tree (CART) analysis: It is a combination of both classification and regression analysis.

Though the trees used for regression and classification are similar in procedure but differ by their procedures of splitting a node, a decision tree structure resemble a flow-chart, in which each internal node represents test and their branch represent the outcome ,after the completion of test and also its leaf node denotes the label of the class. The first or node in level one represents the root node. There are many specific decision-tree algorithms.

One of them is Gini Index which is used in Classification and Regression Tree.

3.1 Gini index: It is calculated by summing up the probability of the datasets.

During this technique the following advantages and disadvantages of using decision tree has been identified:

- (a). Easy to understand and interpret.
- (b). Needs only training set and test set of data.
- (c). Can manage both numerical and categorical data.
- (d). Results are validated by statistical tests which add reliability to the model.
- (e). It cannot manage large datasets.
- (f). Accuracy is lost due to the above reason.

In the existing system, most of them were able to predict the weather but not the damage caused by it. They have used data mining techniques like clustering and decision trees, where they used the data i.e., parameters of atmosphere like temperature, humidity, pressure etc to estimate the occurrence of rainfall or to predict whether particular day are sunny or not but not the damage caused by the weather.

Now, a brief discussion was conducted, how the clustering and decision tree techniques are used by the earlier researchers to predict the future weather is given below:

A. Noval model:

a. Prediction:

Prediction is the task of predicting values for given input. For example we would like to predict value of damage in a location by collecting past data. The most common approach used for numerical data is regression.

Regression analysis can be used to model the relation between one or more independent (predictor) variable and dependent variable.

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. As it is used in both classification and regression hence we choose KNN in our proposed approach.

Whereas in KNN regression output is the property value for the object. This value is the average of the values of its k nearest neighbors.

KNN is a type of instance-based learning where the function is only approximated locally and all computation is deferred until classification. The KNN is most simple approach.

In regression, it can be useful to weight the contributions of the neighbors, so that the neighbors contribute more to the average than the more distant ones. For example, if the rain fall is continuous for three days and if at all it continues even for fourth day we are going to predict damage rate by considering the parameters like wind speed, wind direction, rain, humidity, temperature.

Neighbors are taken from a set of objects for which the class (for KNN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

In our way to find out the solution for these kinds of problems like, which are faced by the people who are nearer to the coastal area and to prevent them and to make them aware of the situations, where sometimes there is a probability of villages covered by water if at all there is a heavy rain fall and the damage caused by the rain due to its high force and intimidating the local people surrounding the dams, when there is a chance of its overflow and if any. These are the cases which we are considering i.e., in the first case we are going to consider the damage value directly which is readily available and in second case we are going to consider the nearest neighbor for the value to be considered and in the last case we will consider the mean value of the available damage values so that all the cases can be covered.

The following algorithm is going to be implemented for the prediction of damage value:

B. Proposed Algorithm:

Steps involved are:

- a. Previous data is collected.
- b. Parameters present in it are observed and processed.
- c. A particular locations data and parameters (majorly the parameters related to rain and its damage) are focused.

- d. Now using KNN the above steps are repeated such that the mean value of rain and its damage is calculated where following cases can be followed:

Case 1: The direct value of the damage is considered.

Case 2: nearer value to the Expected damage is taken.

Case 3: Average of below five and above five values were taken as the approximate value.

- e. After obtaining both the values from step4 the damage value can be predicted.

For example consider a nearby lake like Hussain sagar in Hyderabad. If there is heavy rain fall for two days of 8cm and it has reached its limit and if it continues for third day with 5cm of rain fall there may be a chance of water overflow which harms the people in those surroundings and even it disrupts the vehicular movement on the roads. Due to this reason there may be a loss which cannot be recovered. In order to provide a message of precaution we need to have a system which predicts the affect based on the previous data so that, the people may be intimated that this is the situation where it may take time to overflow (lake) before which they can shift their shelter to nearest safe place and even we would like to predict the damage of roads as it saves the time of going to a particular place and to estimate the figure of damage.

IV. CONCLUSION

The purpose of the current paper is to determine effects caused by the natural calamities like not only floods, cyclones, incessant rains but to given a prior warning or indication to the people which can minimize the heavy loss due to natural calamities with the data, we can put the situation under control and can also minimize the damage to be caused.

V. REFERENCES

- [1]. Applications of Data Mining Techniques in Weather Prediction and Climate Change Studies", 2012.
- [2]. "Meteorological Data Mining", 2012.
- [3]. "Data Mining using Rattle", 2013.
- [4]. "Data Mining Technique to analyze the Meteorological data", 2013.
- [5]. "Data warehousing and Data Mining application for Atmospheric Studies", 2013.
- [6]. "Effectiveness of using Data Mining for Predicting Climate Change in Sri Lanka".
- [7]. "Hadoop based ARIMA Algorithm and its Application in eather Forecast", 2013.