Sangari.R.S *et al*, International Journal of Computer Science and Mobile Applications,
Vol.2 Issue. 2, February- 2014, pg. 84-88 **ISSN: 2321-8363**

A SURVEY ON RAINFALL PREDICTION USING DATAMINING

Sangari.R.S[*1], Dr.M.Balamurugan[#2]

**1 School of Computer Science and Engineering, Bharathidasan University, Trichy, India*

#2 School of Computer Science and Engineering, Bharathidasan University, Trichy, India

1. sangari_selvam@yahoo.co.in

2. mmbalmurugan@gmail.com

Abstract

India is an agricultural country and its economy is largely based upon crop productivity. For analyzing the crop productivity, rainfall prediction is require and necessary. Rainfall Prediction is the application of science and technology to foretell the state of the atmosphere. It is important to exactly determine the rainfall for effective use of water resources, crop productivity and pre planning of water structures. Using data mining techniques we can predict rainfall. Data mining techniques are used to estimate the rainfall numerically. This paper focuses some of the data mining algorithms for rainfall prediction. Naive Bayes, K- Nearest Neighbour algorithm, Decision Tree, Neural Network and fuzzy logic are some of the algorithms compared in this paper. From that comparison, we can analyze which method gives better accuracy for rainfall prediction.

Keywords:- Naive Bayes, K- Nearest Neighbour algorithm, Decision Tree, Neural Network, Fuzzy Logic.

I. INTRODUCTION

Data mining is the process of extracting or mining knowledge from large amount of data. In other words, data mining is the efficient discovery of valuable, non-obvious information from a large collection of data. The goal of data mining is to extract information and convert them into useful knowledge for future information. Data-mining is a process that can be applied to any type of data ranging from weather forecasting, electric load prediction, product design, among others. 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 converted into usable information. Useful information can play vital role in understanding the climate variability and rainfall prediction. This understanding can be used to support many important sectors that are affected by climate like agriculture, water resources, forestry and tourism. Particularly, it is useful to foresee the natural disaster like flood and drought. Thus many data mining algorithms are used to predict the rainfall.

II. ALGORITHMS USED FOR RAINFALL PREDICTION

1. K- Nearest Neighbour algorithm

K-means algorithm is the most popular clustering tool used in scientific and industrial applications. The name implies from exhibiting each of k clusters by the mean or weighted average of its points, the so-called centroid. The

centroid of a cluster is a point whose coordinates are the mean of the coordinates of all the points in the clusters [1][2]. The *k*-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k *n*earest neighbors (*k* is a positive integer, typically small). If ***k* = 1**, then the object is simply assigned to the class of that single nearest neighbor. The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its *k* nearest neighbours. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. The *k*-nearest neighbor algorithm is sensitive to the local structure of the data. Nearest neighbor rules in effect implicitly compute the decision boundary. It is also possible to compute the decision boundary explicitly, and to do so efficiently, so that the computational complexity is a function of the boundary complexity.

2. Decision Tree

A Decision Tree is a flow-chart-like tree structure. Each internal node denotes a test on an attribute. Each branch represents an outcome of the test. Leaf nodes represent class distribution. The decision tree structure provides an explicit set of "if-then" rules (rather than abstract mathematical equations), making the results easy to interpret [4][5]. A decision tree is a predictive model most often used for classification. Decision trees partition the input space into cells where each cell belongs to one class. The partitioning is represented as a sequence of tests. Each interior node in the decision tree tests the value of some input variable, and the branches from the node are labeled with the possible results of the test. The leaf nodes represent the cells and specify the class to return if that leaf node is reached. The classification of a specific input instance is thus performed by starting at the root node and, depending on the results of the tests, following the appropriate branches until a leaf node is reached [6]. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities. Decision tree is a predictive model that can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of the data set with their classification. The author defines a Decision Tree as a schematic tree-shaped diagram used to determine a course of action or show a statistical probability [7].

3. Naïve Bayes

Bayesian classification is a kind of the statistical classification. It's an algorithm based on the probability. Bayesian algorithm theory is rediscovered and perfected by Laplace, the basic idea is using of the known prior probability and conditional probability density parameter, based on Bayes theorem to calculate the corresponding posterior probability, and then obtained the posterior probability to infer and make decisions[8]. Naive Bayes is a classification model based on the famous Bayes theorem. Naïve Bayes is a tree Bayesian network which contains a root node, a plurality of leaf nodes. In which the leaf node is an attribute variable. It descripts the properties of the object to be classified. The root node is a class variable that describes the object's categories[9]. The Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". A naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods. Despite their naive design and apparently oversimplified assumptions, Naive Bayes classifiers have worked quite well in many complex real-world situations. Bayes method—also called idiot's Bayes, simple Bayes, and independence Bayes. This method is

important for several reasons. It is very easy to construct, not needing any complicated iterative parameter estimation schemes. This means it may be readily applied to huge data sets. It is easy to interpret, so users unskilled in classifier technology can understand why it is making the classification it makes. And finally, it often does surprisingly well: it may not Probabilistic approaches to classification typically involve modeling the conditional probability distribution. An advantage of Naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

4. Artificial Neural Network

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a huge number of highly interconnected processing elements (neurons) working in unison to solve specific problems. An ANN is configured for a particular application, such as pattern recognition or data classification, through a learning process. The artificial neuron is an information processing unit that is fundamental to the operation of a neural network. The artificial neural networks not only analyze the data but also learn from it for future predictions making them suitable for weather forecasting. Neural networks provide a methodology for solving many types of non-linear problems that are difficult to be solved through traditional techniques. Furthermore neural networks are capable of extracting the relationship between inputs and outputs of a process without the physics being explicitly provided. Hence these characteristics of neural networks can be used for the prediction of the weather processes [10]. The utility of artificial neural network models lies in the fact that they can be used to infer a function from observations. This is particularly useful in applications where the complexity of the data or task makes the design of such a function by hand impractical.

5. Fuzzy logic

A fuzzy logic model is also known as a fuzzy inference system or fuzzy controller. The fuzzy logic model adopted in this work composed of two functional components. One is the knowledge base, which contains a number of fuzzy if–then rules and a database to define the membership functions of the fuzzy sets used in the fuzzy rules. Based on this knowledge base, the second component is the fuzzy reasoning or decision-making unit to perform the inference operations on the rules. Two operations are performed for fuzzy logic modeling. When data are ready, a fuzzification operation is processed to compare the input variables with the membership functions on the premise part to obtain the membership values of each linguistic fuzzy set. These membership values from the premise part are combined through a min operator to get firing strength (weight) of each rule in order to generate a qualified consequent (either fuzzy or crisp) of each rule depending on this firing strength. Then the second operation is the defuzzification to aggregate the qualified consequents to produce a crisp output [11]. FL is very useful in modeling complex and imprecise systems, and fuzzy set theory is a powerful tool and its applications have rapidly increased with establishing its utility in numerous areas of the scientific world. Any system consisting of vague and ambiguous input variables may contribute to an ultimate effect. The fuzzy logic possibility and its degree of effect due to the ambiguous input variables are considered by some as being generated in the human mind and is often referred to as expert knowledge. This expert knowledge is the accumulation of knowledge and ideas as a result of the expert's experience in a particular system; hence, decision-making processes may be considered as fuzzy expressions perceived by the expert[12].

III. COMPARATIVE PERFORMANCE OF VARIOUS DATAMINING TECHNIQUES

Table 1.Comparison of Data mining techniques and its accuracy

Data mining techniques	Accuracy
Naïve Bayes	82.81%
KNN (k=30)	81.81%
Decision Tree	81.40%
Neural Networks	85.77%
Fuzzy logic	68.93%

In Table 1, We analyze five data mining techniques and its accuracy to predict the rainfall[2]. From this table, we understand that the Neural Network has better accuracy in result than other data mining techniques. Accuracy is determined by the formula 100-RMSE. RMSE (Root Mean Square Error) is one of the measuring techniques for predicting rainfall. It is measured by the differences between values predicted by a model and the values actually observed from the model[11].

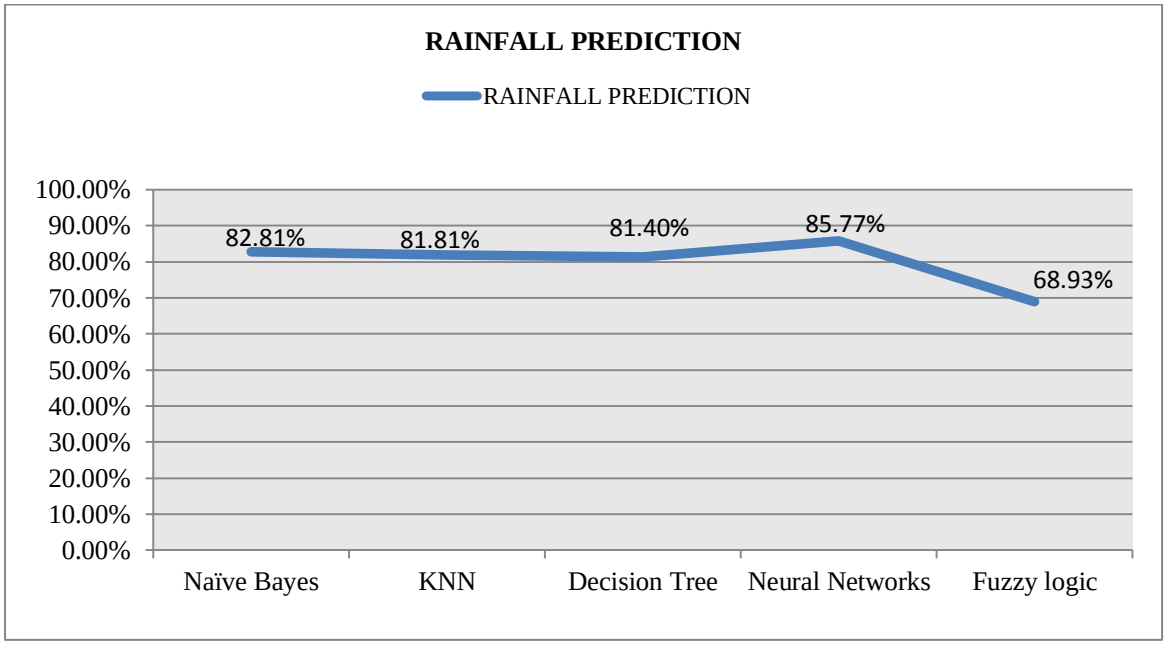

Figure 1.Graphical representation of Data mining techniques and its accuracy

IV. CONCLUSION

The rainfall which is an important factor for the use of water resources is a difficult variable to estimate. Basically, there are two approaches used for prediction. They are Empirical method and dynamical methods. The empirical approach is based on the study of historical data of the rainfall and its relationship to a variety of atmospheric and oceanic variables over different parts of the world. Regression, artificial neural network, fuzzy logic and group method of data handling are some of the empirical approaches used to predict rainfall. In dynamical approach, predictions are generated by physical models based on systems of equations that predict the evolution of the global climate system in response to initial atmospheric conditions. The Dynamical approaches are implemented using numerical rainfall forecasting method. In this study, one of the data-mining empirical approaches is used to predicate rainfall. While analyzing this paper Neural Network model from data-mining process gave more accuracy in result than any other algorithm.

REFERENCES

[1] Berkhin P, 2002, CA, Tech. Rep, *"Survey of clustering data mining techniques, Accrue Software, San Jose"*.

[2] Sarah N. Kohail, Alaa M. El-Halees, July 2011,*" Implementation of Data Mining Techniques for Meteorological Data Analysis".*

[3] Han, J., Micheline K, 2007, San Fransisco, CA: Morgan Kaufmann publishers, *"Data Mining: Concepts and Techniques"*.

[4] Folorunsho Olaiya and Adesesan Barnabas Adeyemo, February 2012, *"Application of Data Mining Techniques in Weather Prediction and Climate Change Studies".*

[5] Lior Rokach and Oded Maimon, 2008, World Scientific Publishing Company, *"Data Mining with Decision Trees: Theory and Applications (Series in Machine Perception and Artificial Intelligence)",* ISBN: 981-2771-719.

[6] Venkatadri.M and Lokanatha C. Reddy , Sept 2010, International Journal Of Computer Applications In Engineering, Technology And Sciences (IJCAETS), *"A comparative study on decision tree classification algorithm in data mining"*, Vol.- 2, no.- 2 , pp. 24- 29 .

[7] Y Zhang, CH Chu, Y Chen, H Zha, X Ji, 2006, *"Splice site prediction using support vector machines with a Bayes kernel. Expert Systems with Application", 30*: 73-81.

[8] McCallum A, Nigam K, 1998, AAAI-98 workshop on learning for text categorization, *"A comparison of event models for naive bayes text classification"*, 752: 41-48.

[9] Jasmeen Gill, Baljeet Singh and Shaminder Singh, 2010, IEEE 8th International symposium on intelligent systems and informatics Serbia, *"Training Back Propagation Neural Networks with Genetic Algorithm for Weather Forecasting".*

[10] Agboola A.H., Gabriel A. J., Aliyu E.O., Alese B.K, April, 2013, *" Development of a Fuzzy Logic Based Rainfall Prediction Model",*Volume 3 No. 4.

[12] Somia A. Asklany, Khaled Elhelow, I.K. Youssef, M. Abd El-wahab, February 2011*,"Rainfall events prediction using rule-based fuzzy inference system".*

