Original Research Article

SACRED HEART RESEARCH PUBLICATIONS

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in

ISSN: 2456-9496

A HYBRID MODEL OF K-MEANS CLUSTERING AND MULTI LAYER PERCEPTRON FOR RAINFALL

R. Pugazendi[#1], P. Usha[#2]

Received on 26th Aug 2017, Accepted on 8th Dec 2017

Abstract — Abstract— *India is a farming country which mostly depends on monsoon used for irrigation intention. An enormous sum of water is consumed for industrial creation, crop yield, and domestic use. Rainfall prediction is thus very essential and necessary for the development of the country. Weather factors including mean, precipitation, minimum temperature, average temperature, maximum temperature, cloud cover, vapor pressure, wet day frequency and diurnal temperature have been used to forecasts the rainfall and improve the accuracy of the rainfall prediction from the weather point. Continuous collection of historical weather data will assist in improving the accuracy and finding the rainfall prediction. The condition of weather data can be detected more accurately and in a fast manner, to terminate the problems and give a better solution. By using the hybrid methods K-Means clustering and Multilayer perceptron one can find the various issues and its consequences related to weather by analyzing and classifying the weather data there by producing an accurate rainfall prediction within a limited span of time and high-level accuracy.*

Keywords: Rainfall Prediction, K-Means clustering, Multi Layer Perceptron (MLP), Indian Metrological Weather Data.

I INTRODUCTION

In olden days, mining of pattern from datasets are done manually. It leads to a huge difficulty in Pattern Recognition. In the modern computer era, data set group, handling, and its storage is significantly made easy and also increased. Data mining is an inbuilt computerized algorithm in order to classify specific pattern from the large set of data. A model developed by learning concept using a computer is the division of data mining called as machine learning. For the given large data set, this model learns by training and testing and by using the learning concept, it predicts new instances of data.

The method of generating classification among the models in order to predict a result is commonly called as modeling. Classification is data mining process which predicts the value of a categorical variable by developing a model based on one or more attributes associated with the categorical variable and by classifying the identified data based on the training set and class labels. The clustering is mostly used for the unsupervised learning problems. So, as in which the problems are, dealt with discovering a formation in a group of unlabeled data. Therefore the clustering is a collection of things which are "related" and "unrelated" to the things which are belonging to further clusters. And also the clustering model is used to find natural grouping among things [11].

A time sequence is an order of data points scheduled in time order. Most usually, a time sequence is an order taken at successive uniformly spaced points in time. In a nutshell, it is an order of discrete-time information. And in order to extract the meaningful information, it's vital to analyze the data in a time sequence. The time series forecast new values based on existing data set. These methods divide into two types of techniques; one is a parameter and another one non-parameter. The parameter technique is used for a small number of factors and another one is obviously approximate to covariance [10]. Predicting the individuality of one item is based only on the description of another, related item. Prediction is necessary for predicting the future events which are unknowns, based on the connection between an object [12].

In, water related studies. "A long rainfall data series" is the essential part. Uniformity and continuity are played the main role to get dependable results. It includes gaps and missing value because of various reasons like as absence of observers, the problem with measuring procedure, loss of records etc. Arithmetical values are used to find out about the accuracy of study while the missing values. Some other methods can be used to replace the missing values. A serious could be replaced to make the water related study more believable. Basically, the measures can be grouped into three major classes as deterministic, stochastic and artificial intelligence. The deterministic technique is used to very suitable to work because of their robustness. It was implemented easily and computational efficiency.

* Corresponding author: E-mail: pugazendi_r@gmail.com, ushapoomalai@gmail.com.

[1]Assistant Professor, Dept. of Computer Science, Government Arts College, Periyar University, Salem – 7, Tamilnadu, India.

[2]Research Scholar, Dept. of Computer Science, Government Arts College, Periyar University, Salem – 7, Tamilnadu, India.

Some e.g. Arithmetic mean technique, normal ratio technique and inverse distance weighting technique. The stochastic technique is used to give probabilistic estimates of the result, complex and cost effective technique. Artificial intelligence methods such as artificial neural networks (ANNs) have a difficult mathematical formulation, cost effective and complex to implement.

Agriculture is the main occupation in India, accounting for concerning 52% of employment. Most of the subcontinent depends on the rainfall for the farming needs. The rainfall data accessible in the data mining techniques can be helpful for rainfall prediction.

II RELATED WORK

The uses of patterns in predictive models have received a group of attention in recent years. Pattern mining can facilitate to attain models for the structured domain, such as graphs and sequences. This paper presents a perspective on the evolving area. It supplied an overview of pattern-based classification methods. Which was done along the following dimensions: whether they used a pre-computed group of patterns or execute pattern mining algorithms and whether they chose patterns model-independently or whether the pattern selection was guided by a model, the authors revealed through their study that SCIP rule based classifiers presented more accurate higher classification than the other existing methods. It was not overly sensitive to the setting of a minimum support and confidence threshold, which was also proved to be a scalable and valuable tool to find representative patterns [1].

The author discussed Meteorologists process and studied with the help of visualization. This work presented the identity of the problems and connected to the weather data with meteorological forecasting. It extended a group of techniques as a starting step toward straight visualizing connections. It connected of multiple features over ensemble forecast. Finally, it has given a conclusion about the about the integration of the contributions into a functional of prototype tool, and also the different practical challenges that arose while working with weather data [2].

This paper presented the ensemble method based on ANFIS (Adaptive Neuro Fuzzy Inference System) and ARIMA (Autoregressive Integrated Moving Average) algorithm. It was used to predicting the monthly rainfall data for a particular area. For e.g. Indonesia, specifically Pujon and Wagner. In this learning, Gaussian, Gbell, and Triangular role were used in ANFIS membership method. ARIMA and ANFIS are the best individual method while compared with another method. It yields correct prediction in monthly pujon's rainfall data. It was based on the root of mean square errors (RMSE) at difficult datasets, the result showed that an entity ANFIS and ARIMA method yields better correct prediction in monthly Pujon's rainfall data and Wagner's rainfall data respectively [3].

This paper presented Bayesian method algorithm to cumulative forecasting time series rainfall information. ANN- filter was implemented to produce Bayesian rainfall information. The time series data was used to distribution assumption technique to all parameter values and possible models included their posterior distribution. Hurst parameter was denoted by H taking into a description that the time series predicted the same real time series H was the same series. This approach of performance was a sample of the Mackey glass tested over a time obtained. Mackey-Glass delay discrepancy equations and cumulative rainfall time series data. It was considered the geographical points of Cordoba, Argentina. The training set could apply the online heuristic law and change the NN topology, modify the number of patterns and iteration addition to the Bayesian inference in accordance with Hurst parameter [4].

This paper presented the decision tree algorithm to predict weather parameters such as fog, rainfall, cyclones, and thunderstorms. It was a lifesaving method and it making intelligent decisions. This model was used to improve more relevant. The attributes were used to predicting the dependent variables. The proposed model was implemented using the open source data mining tool Rapidminer and performed at weather forecasting with high accuracy rate. To comparing soft computing techniques it gave best results [5].

This paper presented a novel modular-type Support Vector Machine (SVM) to simulate rainfall prediction. There are more four techniques can be used for rainfall prediction. The first one was bagging sampling method. It was used to generate different types of training sets. The second one SVM method of a different kernel function with a different parameter. For e.g. base model. It was for different regression model trained by the formulated in a different training set. The third method Partial Least Square (PLS) method was used to select the exact number of SVR combined objects. The v-SVM model is the last base model to produce the learning process very easy. There is all the technique would be implemented to predict monthly rainfall data set in the place of Guangxi, China. The experimental result showed in achieving better prediction accuracy and increase prediction quality in the method of v-SVM [6].

This paper presented a Bayesian algorithm to predict weather data Indian Meteorological Department (IMD) Pune from the author. Some knowledgeable information extracted from historical weather data. The weather data collected from 36 attributes. Among the 36 attributes and only 7 attributes were used in rainfall prediction, a raw weather data was used to data pre- processing and data transformation of weather data. The rainfall prediction was used to prediction in data mining model. The high performance of computing and Super-computing power used for meteorological weather center to run a weather prediction.

To implement data used data intensive model in a rainfall prediction. To reduce address issues of the computed intensive model. The moderate compute resource used to predict the rainfall and the experimental result worked with good accuracy [7].

This paper presented K-means clustering and J48 algorithm to predict meteorological data. To predict the weather was important to help, prepare for best and worst climate. This algorithm used for identifying the various weather conditions. The J48 algorithm for classifying weather data and expectation maximization algorithm for finding the maximum value of weather condition and K-mean applied for finding maximum and minimum weather condition [8].

This paper presented AdaNaive and AdaSVM algorithm to predict weather data in the long periods, to measure the accuracy and classification error. In this paper, the author compared many algorithms to find which would the best accuracy and also found that AdaSVM algorithm is produced the best accuracy compared to Naive, SVM and AdaNaive algorithm [9].

III METHODOLOGY

A. Data Collection And Pre-processing

The data set has been collected from the source https://www.wunderground.com. One of the main challenging issues in the predicted accuracy is consistent weather data. The early stage in data mining procedure is collecting data and pre-processing. The crucial stage for rainfall prediction is fitting data for accurate output. The selected weather data from 1998 to 2002 models are used for rainfall prediction. In the weather, a lot of considers more than weather attributes have to be considered is Attributes Relationship File Format using time series data with parameter techniques. The parameter technique is a particular structure in which particular a time interval data set have been taking. The weather data have been indicated by numerical values and expected result i.e. rainfall prediction has been mentioned by binary values (1, 0). The model used to predict rainfall. The starting is to process load time series weather data set, then the training data set using particular sample historical weather data has to be done and at then providing a test data set for predicting rainfall with accuracy by adopting clustering and classification model as shown in *Figure I*.

B. Hybrid Method of K-Means Clustering and Multilayer Perceptron

Hybrid technique included is the best technique for acquiring accuracy in rainfall prediction. K-Means clustering and Multilayer perceptron are called a hybrid method of K-Means clustering and Multilayer perceptron. In this paper, K-Means and Multilayer perceptron have been implemented using JAVA. The K-means Cluster and Multilayer perceptron algorithm used to find the best accuracy and reduce an execution time.

S.NO	ATTRIBUTE	TYPE	DESCRIPTION
1	Month	String	Month
2	Year	Integer	Year
3	Precipitation	Numeric	Mean Total precipitation in inches.
4	MinimumTemp	Numeric	Mean Minimum temperature in F
5	Average temperature	Numeric	Mean Average temperature in F
6	MaximumTemp	Numeric	Mean Maximum temperature in F
7	CloudCover	Numeric	Mean Cloud Cover
8	vapour pressure	Real	Mean Vapour pressure
9	WetDayFrequency	Real	Mean Wet day Frequency
10	DiurnalTemperature	Real	Mean Diurnal Temperature
11	GroundFrostFreq	Real	Mean Ground Temperature
12	ReferenceCropEvap	Real	Mean Reference Crop Evapotranspiration
13	PotentialEvap	Real	Mean Potential Evapotranspiration

Table 1 - Weather data and description

K-Means Clustering

K-Means clustering is one of the unsupervised cluster algorithms. To group the elements without any prior knowledge of the individual relationship. K-Means is used, to cluster the similar weather data for using the partition-based cluster. The similarity measurement is clustering with using training and testing dataset, by applying centroid clustering methods

Multilayer perceptron

A multilayer perceptron is a supervised learning classification. It consists of three layers such as input layer, a hidden layer, and an output layer. It is either multilayer linear or nonlinear. In this paper feed forward multilayer, neural networks in developing the model for rainfall prediction to improve the accuracy and reduce time has been implemented.

C. Algorithm

The step by step of process Proposed Algorithm.

Step1: Load Rain Fall Dataset to Clustering Variables and Maximum Number of Clusters (K in Means Clustering).

Step2: Initialize cluster centroid. In this example, the value of K is considered as 5. Cluster centroids are initialized with first observations.

Step3: To Calculate Euclidean Distance.

$$\mathbf{J\,(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||y_i - z_j||)^2}$$

Where,

'$||y_i - z_j||$' is the Euclidean distance between y_i is a training data set and z_j is a testing data set. A 'c_i' is the number of data points in the i^{th} cluster. c=5 is the number of cluster centers.

Step4: Remove Unsimiliar Attributes on both train and test dataset.

Step5: Randomly initialize the weights in the system.

Step6: To apply the weather data to the system and compute it out for rainfall prediction.

Step7: To calculate the error (E) E= desired-computed.

Step8: To calculate the Δwi for all weights in a backward pass from hidden layer to output layer.

$$\mathbf{net_{i1}=w_1*i_1+w_2*i_2+b_1*1}$$

$$\mathbf{output_{i1}=1/1+e^{-}net\ i1}$$

Step 7: To calculate the Δwh for all weights in a backward pass from hidden layer to output layer.

$$\mathbf{net_{h1}=w_1*h_1+w_2*h_2+b_2*1}$$

$$\mathbf{output_{h1}=1/1+e^{-}net\ h1}$$

Step 8: To calculate the Δwo for all weights in a backward pass from the input layer to hidden layer. Update the weights in the network.

$$\mathbf{net_{o1}=w_1*o_1+w_2*o_2+b_3*1}$$

$$\mathbf{output_{o1}=1/1+e^{-}net\ o1}$$

Step 9: Repeat the step 4 to 8 for every training pattern until all pattern classified correctly

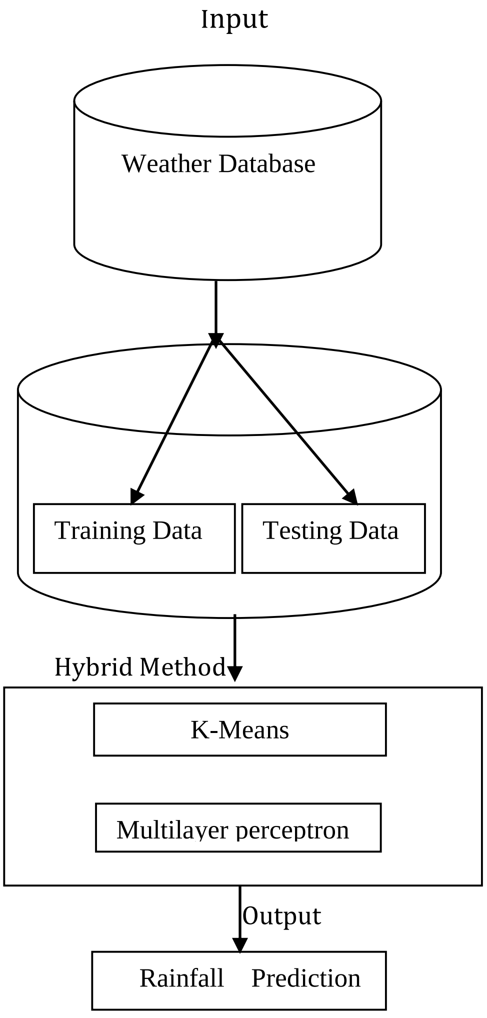

Figure I - System Architecture Diagram

IV EXPERIMENTAL RESULT

A. *Accuracy*

Accuracy means correct classification value for rainfall prediction and close measurement of accepted value as shown in *Figure II*. Which comparing the existing algorithm, a hybrid method of K-Means clustering and Multilayer perceptron accuracy value is high for rainfall prediction.

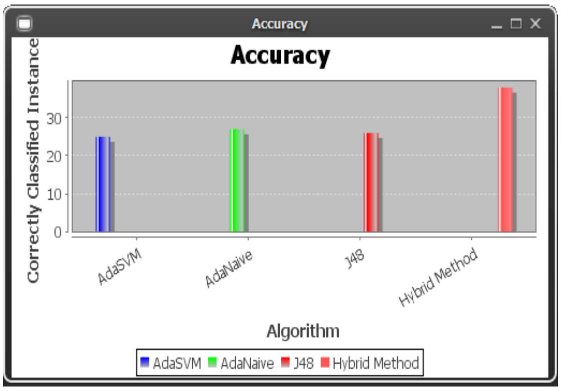

Figure II - shows Accuracy value comparison with existing and proposed approaches graphically

B. *Execution Time*

Execution time is known as building process in rainfall prediction using weather data as shown in *Figure III*. The existing algorithm, a hybrid method of K-Means clustering and Multilayer perceptron algorithm reduced execution time for rainfall prediction than the other algorithms.

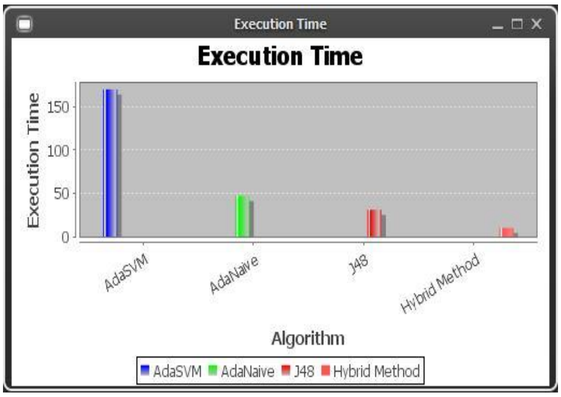

Figure III - Shows Execution Time comparison with existing and proposed approaches graphically.

C. *Recall*

Recall is the relation of the number of relevant information retrieved to the total number of relevant information in the weather database. Recall is usually expressed as a one hundredth as shown in *Figure IV*. A hybrid method of K-Means clustering and Multilayer perceptron algorithm recall performance is high for rainfall prediction while comparing the existing algorithms.

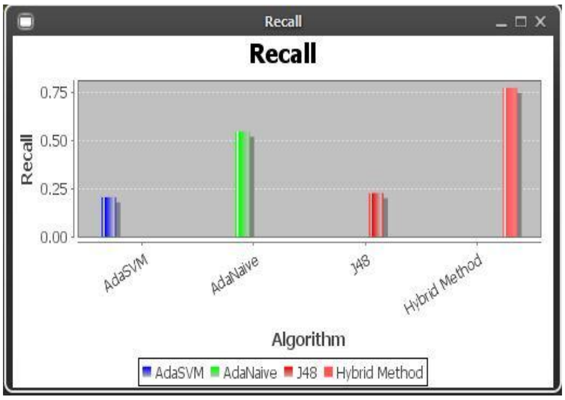

Figure IV - shows Recall value comparison with existing and proposed approaches graphically

D. Precision

Precision means reproducibility of several measurements. Precision is usually described in standard deviation and standard error. Comparing existing algorithm, a hybrid method of *K*-Means clustering and Multilayer perceptron is best and it's shown in *Figure V*.

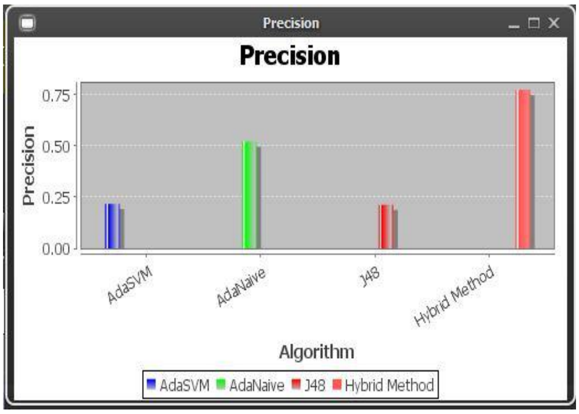

Figure V - shows Precision value comparison with existing and proposed approaches graphically.

V CONCLUSION

Rainfall prediction is an imperative application in the meteorological department and also one the challenging issues around the world. The proposed system analyzed the use of data mining techniques for predicting rainfall. This is carried out by using data mining techniques like clustering and classification and the hybrid method of K-Mean clustering and Multilayer perceptron algorithm which has been applied to the weather data in the particular time to compare existing algorithm for rainfall prediction using weather data. The algorithm which predicts the best result in accuracy and also reduces execution time for rainfall prediction is a hybrid method of K-Means clustering and Multilayer perceptron.

REFERENCES

[1] Cheng Zhou, Boris Cule, Bart Goethals “Pattern Based Sequence Classification”, IEEE Transactions on Knowledge and Data Engineering, Vol. 28, 5, pp.1285-1298, 2016.

[2] P.Samuel Quinan, Miriah Meyer “Visually Comparing Weather Features in Forecasts”, IEEE Transactions on Visualization and Computer Graphics, Vol. 22, 1, pp. 389-398, 2016.

[3] Suhartono, Ria Faulina, Dwi Ayu Lusia, Bambang W. Otok, Sutikno, Heri Kuswanto “Ensemble Method based on ANFIS-ARIMA for Rainfall Prediction”, IEEE International conference on statistics in Science, Business and Engineering (ICSSBE), pp.1-4, 2012.

[4] C. R. Rivero, J. Pucheta, S. Laboret, M. Herrera and V. Sauchelli “Time Series Forecasting Using Bayesian Method: Application to Cumulative Rainfall”, IEEE Latin America Transactions, Vol. 11, 1, pp. 359-364, 2013.

[5] A.Geetha, G.M Nasira " Data Mining for Meteorological Applications: Decision Trees for Modeling Rainfall Prediction", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp.1-4, 2014.

[6] Kesheng Lu, Lingzhi Wang “A Novel Nonlinear Combination Model Based on Support Vector Machine for Rainfall Prediction", IEEE 4th International Conference on Computational Sciences and Optimization (CSO), pp.1343-1346, 2011.

[7] V. B. Nikam and B. B. Meshram, "Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach", Fifth International Conference on Computational Intelligence, Modeling and Simulation, Seoul, pp.132-136, 2013**.**

[8] P. Kalaiselvi, D. Geetha, “Weather Prediction Using J48, EM And K-Means Clustering Algorithms”, International journal of innovative Research In Computer and Communication Engineering, Vol. 4, pp.20889-20895**,** Issue12, December 2016.

[9] B.Narayanan, M.Govindarajan,” Rainfall Prediction based on Ensemble Model”, International Journal of Innovative Research in Science Engineering and Technology, Vol.5, Issue 5, May2016.

[10] Time Series information Available URL is https://en.wikipedia.org/wiki/Time_series”.

[11] Clustering information Avialable URL is https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/.

[12] Data mining Prediction information avilable URL is “www.cs.stir.ac.uk/courses/ITNP60/.../1%20Data%20Mining/4%20-%20Prediction.pd “.

[13] Multilayer Preceptron information available URL is https://en.wikipedia.org/wiki/Multilayer_perceptron.