International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

A Survey on Ensemble Computing Method for Rainfall Prediction in Different Regions of Chhattisgarh

Shahista Navaz[1], Huma Khan[2], Dr. S. M. Ghosh[3]

[1]PhD Research Scholar, CVRU, Bilaspur, Chhattisgarh India

[2]PhD Research Scholar, CVRU, Bilaspur, Chhattisgarh India

[3]Professor, Dept of Computer Science and Engineering, CVRU, Bilaspur, Chhattisgarh India

Abstract: ***India is an agricultural country and most of economy of India depends upon the agriculture. Rainfall plays an important role in agriculture so early prediction of rainfall is necessary for the better economic growth of our country. Rainfall Forecasting is one of the most challenging topics across the globe. Unlike traditional methods, modern Weather forecasting involves a combination of knowledge of trends and patterns and computer models. Using these methods, accurate forecasting could be done. Forecasting System is based on any architecture or technology such as Artificial Neural Network Forecasting, Sensor-Based, Numerical weather prediction model, Fuzzy sub-system or a used friendly web based system. The weather predictions in advance of 2-8 days are also possible. Forecasting could be carried out globally or region based. This paper represents a review of different rainfall prediction techniques for the early prediction of rainfall prediction of rainfall. It also focuses on giving different ways to forecast the weather in different regions. The paper presents the review of Rainfall Forecasting using different techniques and studies the benefit of using them. It provides a survey of available literatures of some techniques given by different researchers. The technical success, that have been achieved by various researchers in the field of rainfall forecasting, has been reviewed and presented in this survey paper.***

Keywords: NN, SVM,ANN,ARIMA, K-NN, Naïve Bayesian, Regression,ID3, WEKA tools

1. Introduction

Agriculture is the backbone of Indian economy. Irrigation facility is still not so good in India and most of agriculture depends upon the rain. A good rainfall result in the occurrence of a dry period for a long time or heavy rain both affect the crop yield as well as the economy of country, so due to that early prediction of rainfall is very crucial . A wide range of rainfall forecast methods are employed in weather prediction at regional and national levels. Fundamentally there are two approaches to predict Rainfall. They are Empirical and Dynamical Methods. The Empirical approach is based on analysis of past historical data of weather and its relationship to a variety of atmospheric variables over different parts of Chhattisgarh. The most widely use empirical approaches used for climate prediction are Regression, artificial neural network, fuzzy logic and group method of data handling. The dynamical approach, predictions are generated by physical models based on system of equations that predict the future Rainfall. The forecasting of weather by computer using equations are known as numerical weather prediction. To predict the weather by numeric means, meteorologist has develop atmospheric models that approximate the change in temperature, pressure etc using mathematical equations.

Data mining [13] is a process which finds useful patterns from large amount of data. Data mining can also be defined as the process of extracting implicit, previously unknown and useful information and knowledge from large quantities of noisy, ambiguous, random, incomplete data for practical application. It is a powerful new technology with great potential to help companies focus on the most important information in their databases. It uses machine learning, statistical and visualization technique to discover and predict knowledge in a form which is understandable to the user. Prediction is the most important technique of data mining which employs a set of pre-classified examples to develop a model that can classify the data and discover relationship between independent and dependent data.

2. Background Study

A. Data Mining

Data mining is the science and technology of exploring data in order to discover unexplored patterns. Traditionally, data meteorological instruments were being refined during the previous centuries. Other related developments that are, theoretical, and technological developments, also contributed to our knowledge of the atmospheric weather conditions. Weather prediction is an important goal of atmospheric research. Hence changes weather condition is risky for human society [3,5,15].It affects the human society in all the possible ways. Weather prediction is usually done using the data gathered by remote sensing satellites. Various weather parameters like temperature, rainfall, and cloud conditions are projected using image taken by meteorological satellites to access future trends. The satellite based systems are expensive and requires complete support systems. The variables defining weather conditions varies continuously with time, prediction model can be developed either statistically or by using some other means like decision tree, artificial neural networks, regression, clustering techniques of data mining. Weather prediction is a form of data mining which is concerned with finding hidden patterns inside largely available meteorological data.

There are various data mining techniques [7,2,3] such as: Classification, Prediction, Clustering, Association, Outlier Detection and Regression. The prediction discovers relationship between independent variables and relationship between dependent and independent variables. There are various algorithms of classification and prediction [8,5,6]. Some of them are Decision Tree, Artificial Neural Networks, Support Vector Machines (SVM), Bayesian Classification and Regression. There are several criteria for evaluating the prediction performance of algorithm [3].

B. Weather Prediction

The various methods used in prediction of weather are:

1) *Synoptic weather prediction*: It is the traditional approach in weather prediction. Synoptic refers to the observation of different weather elements within the specific time of observation. In order to keep track of the changing weather, a meteorological center prepares a series of synoptic charts every day, which forms the very basic of weather forecasts. It involves huge collection and analysis of observational data obtained from thousands of weather stations.
2) *Numerical weather prediction*: It uses the power of computer to predict the weather. Complex computer programs are run on supercomputers and provide predictions on many atmospheric parameters. One flaw is that the equations used are not precise. If the initial stage of the weather is not completely known, the prediction will not be entirely accurate.
3) *Statistical weather prediction*: They are used along with the numerical methods. It uses the past records of weather data on the assumption that future will be a repetition of past weather. The main purpose is to find out those aspects of weather that are good indicators of the future events. Only the overall weather can be predicted in this way.

C. Different Methods of Rainfall Prediction

i. **Multiple Linear Regression**- Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable usually denoted by Y and a series of other changing variables known as independent variables. Regression model which contain more than two predictor variables are called Multiple Regression Model.

ii. Autoregressive Integrated Moving Average (Arima) Model - ARIMA is used to predict a value in a response time series as a linear combination of its own past values, past errors, and current and past values of other time series. The ARIMA procedure provides a comprehensive set of tools for uni-variant time series model identification, parameter estimation, and forecasting, and it offers great flexibility in the kinds of ARIMA or ARIMAX models that can be analyzed.

iii. Genetic Algorithm-Genetic algorithms are algorithms that attempt to apply an understanding of the natural evolution in problem-solving tasks (problem solving). The approach taken by this algorithm is to combine a wide selection of solutions randomly within a population and then evaluate them to get the best solution By doing this process repeatedly, these algorithms simulate the process of evolution as the desired number of generations. This generation will represent improvements on previous population. In the end, we will get the best solutions appropriate to the problems faced. To use a genetic algorithm, solutions to problems represented as a set of genes that make up chromosomes. This chromosome was randomly based coding techniques are used. Chromosomes will be evolved in several stages iterations called generations. The new generation is obtained By cross breeding techniques (crossover) and mutation (mutation). Crossover includes cutting two pieces of chromosomes based on the desired number of points and then combine half of each chromosome with other couples. While mutations include the replacement value of the gene in a chromosome with the value of other genes from other chromosomes become partner. The chromosomes are then evolved to a suitability criterion (fitness) and the set will be selected the best results while others are ignored. Furthermore, the process repeated until you have a chromosome that has the best fit (best fitness) to be taken as the best solution of the problem.

iv. Support Vector Machine (SVM)- A Support Vector Machine (SVM) is a computer algorithm that learns by example to find the best function of classifier hyperplane to separate the two classes in the Input space. The SVM analyzed two kinds of data, i.e. linearly and non- linearly separable data. Support Vector Machine is one of the important category of perceptrons and radial basis function networks, support vector machines can be used for pattern classification and nonlinear regression.

v. Fuzzy Logic (Fuzzy)- Fuzzy Logic is a type of reasoning based on the recognition that logical statements are not only true or false (white or black areas of probability) but can also range from “almost certain” to “very unlikely” .Fuzzy logic has proven to be particularly useful in expert system applications.

d. Weather Research and Forecasting Mode

In general data mining prediction model first we collect the historical weather data. Data were collected from Indian metrological department .The collected data consist of different features including daily dew point temperature (Celsius), relative humidity, wind speed (KM/H), Station level pressure, Mean sea level, wind speed, pressure and rainfall observation. Creating a target data set selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed. Then important step in the data mining is data preprocessing. One of the challenges that face the knowledge discovery process in meteorological data is poor data quality. For this reason we try to prepare our data carefully to obtain accurate and correct results. First we choose the most related attributes to our mining task. purpose we neglect the wind direction. Then we remove the missing value records. In our data we have little missing, because we are working with weather data. Then finding useful features to represent the data depending on the goal of the task. After preprocessing and transforming the weather data choosing the data mining task i.e. classification, regression and decision tree. Then applying different data mining techniques i.e. K-NN, Naïve Bayesian, Multiple Regression and ID3 on weather data set and makes

the rainfall prediction i.e. Rainfall Category or No Rainfall Category.

3. Related Work

There are many studies that support the applicability of data mining techniques for weather prediction Several researchers in computer science have compared several forecasting methods in the prediction of rainfall and weather forecasting. A few of them are outlined here :

Table 1: Comparison of A Data Mining Techniques for Weather Prediction

Authors	Applications	Techniques	Algorithms	Attributes	Time Period	Dataset Size	Accuracy	Advantages	Disadvantages
P. Hemalatha [27]	Weather prediction for ship navigation	Decision tree	C4.5, ID3	Climate, Humidity, Stormy, Temperature	4-5 location	20- 30 instances	-	Verifiable performance	Do not handle continuous range data directly.
E. G. Petre [10]	Weather prediction	Decision tree	CART	Pressure, clouds quantity, humidity, precipitation, temperature	4 years	48 instances	83%	Good prediction accuracy	Data transformation is required. Extra computation required.
S Yeon *et al.*[33]	Hourly rainfall prediction	Decision tree	C4.5, CART	Temperature, wind direction, speed, gust, humidity, pressure	3 years	26280 instances	99%, 93%	High prediction accuracy	Small data is left for prediction.
S Kannan , S Ghosh [29]	Daily rainfall prediction in river basin	Decision tree, Clustering	CART, k-Mean clustering	Temperature, MSLP, pressure, wind, rainfall	50 years	432000 instances	-	Grouping of multisite rainfall data in clusters	Small data is left for prediction. No verification is done.
F Oliya, AB Adeyemo [17]	Weather Prediction and Climate Change Studies	Decision tree, ANN	C4.5, CART, TLFN	temperature, rainfall, evaporation, wind speed	10 years	36000 instances	82%	Best network is selected for prediction	Accuracy varies highly with size of training dataset
P Sallis, S Shanmuganathan [34]	Wind gust prediction	Decision tree, ANN	C5.0, CRT, QUEST, CHAID, SOM	Dew point, humidity, temperature, wind direction, wind speed	4 years	86418 instances	99%, 85%	Good for analyzing ad hoc dataset	Data recorded at irregular intervals. Do not handle continuous data.
GJ Sawale [12]	Weather prediction general	ANN	BPN, Hopfield networks	Temperature, humidity, wind speed	3 years	15000 instances	-	Combining both gives better prediction accuracy	Attribute normalization is required
Amarakoon [1]	Climate prediction in Sri Lanka	ANN	KNN	Temperature, humidity, precipitation, wind speed	1 year	365 instances	-	Beneficial for dynamic data.	Need to integrate feature selection techniques
S Badhiye *et al* [4].	humidity and temperature prediction	Lazy learning, clustering	KNN, K-mean clustering	Temperature, humidity	-	-	100% approx	Suitable for multi-modal classes.	Cannot predict data in remote areas
Z Jan *et al.* [38]	Inter annual climate prediction	Lazy learning	KNN	Wind speed, dew point, seal level, snow depth, rain	10 years	40000 instances	96%	Long term accurate results with large set of attributes.	Cannot incorporate to reflect global changes.
M. A. Kalyankar, S. J. Alaspurkar [23]	Meteorological data analysis	Clustering	K- mean clustering	Temperature, humidity, rain, wind speed	4years	8660 instances	-	Good prediction accuracy	Dynamic data mining methods required.
K Pabreja [16]	Cloud burst predicion	Clustering	K- mean clustering	Temperature, humidity	2 days		100% clustering	Supplement with NWP models.	Not good for long term predictions.
PS Dutta, H Tahbilder [28]	Rainfall prediction	Regression	MLR	Min and max temperature,	6 years	72 instances	63%	Acceptable accuracy.	Attribute elimination

				wind direction, humidity, rainfall					required for better accuracy
M Kannan *et al.*[32]	Short Term Rainfall prediction	Regression	MLR	Min and max temperature, wind direction, humidity, rainfall	3 months for 5 years	450 instances	52%	Can work even with small dataset	Instead of accurate, an approximated value is retrieved.
N Khandelwal, R Davey [25]	Drought prediction	Regression	MLR	Rainfall, sea level, humidity, temperature	1 year	365 instances	-	Coorelation and statistical analysis is also applied.	Verification is not done.

4. Data Collection and Preprocessing

4.1 Feature Extraction

It is the technique of selecting a subset of relevant features for building robust learning models. Many features like Temperature, Evaporation, Wind Speed, Terrain features, Height from sea level, humidity, Precipitable water affects the rainfall .Out of it, the most relevant five features are considered in this paper. The following are the features selected.

4.1.1 Relative Humidity

Relative humidity is a term used to describe the amount of water vapor in a mixture of air and water vapor. It is defined as the ratio of the partial pressure of water vapor in the air-water mixture to the saturated vapor pressure of water at the prescribed temperature. The relative humidity of air depends not only on temperature but also on the pressure of the system of interest. Relative humidity is often used instead of absolute humidity in situations where the rate of water evaporation is important, as it takes into account the variation in saturated vapor pressure.

4.1.2 Pressure

Air pressure varies over time and from place to place and these temporal differences are usually caused by the

temperature of the air. Cool air is denser (heavier) than warm air. Warm air is less dense (lighter) than cool air and will therefore rise above it. Areas of high pressure can be caused when cool air is sinking and pressing on the ground. At this time, the weather is usually dry and clear. In contrast, when warm air rises, it causes a region of low pressure. With low pressure, the weather is often wet and cloudy.

4.1.3 Temperature

Atmospheric temperature is a measure of temperature at different levels of the Earth's atmosphere. It is governed by many factors, including incoming solar radiation, humidity and altitude. Air temperature is the intensity aspect of sun's energy that strikes the earth's surface. Because the amount of energy from the sun reaching the earth varies from day to day, from season to season, and from latitude to latitude, temperatures also vary. The earth as a whole receives a constant flow of radiant short-wave energy from the sun. The earth also radiates long-wave energy to space. During the day, the flow of short-wave radiation absorbed exceeds long -wave energy emitted, and the surface temperature increases.

5. Methodology

Rainfall prediction has become one of the most scientifically and technologically challenging problems in the world. A wide variety of rainfall forecast methods are available. This paper uses data mining techniques such as clustering and classification techniques for rainfall prediction. Prediction can be done by considering the data training and testing them than accordingly building the model shown below:

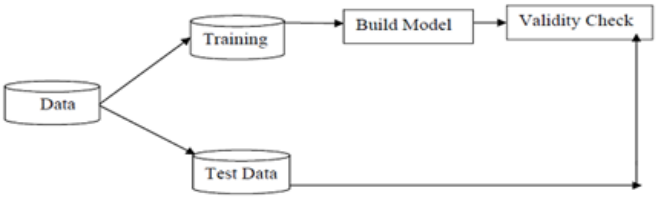

Figure: Overview of the forecasting model

There are various data mining techniques such as: Classification, Prediction, Clustering, Association, Outlier Detection and Regression. The prediction discovers relationship between independent variables and relationship between dependent and independent variables. There are various algorithms of classification and prediction .Some of them are Decision Tree, Artificial Neural Networks, Support Vector Machines (SVM), Bayesian Classification and Regression. There are several criteria for evaluating the prediction performance of algorithm.

The Sample size of the dataset will be of last 10 years i.e. from 2006- 2016 for the rainfall prediction. The dataset will be collected from the meteorological department of Chhattisgarh of the districts Raipur, Gariyaband, Baloda Bazar, Mahasamund, Dhamtari, Durg, Balod, Bemetara, Rajnandgaon, Kabirdham, Bilaspur, Mungeli, Korba, Janjgir and Raigarh.

The methodology used will follow the following steps that are as follows.

- Data Collection
- Data Preprocessing
- Data Transformation
- Applying Classification
- Algorithms.
- Predicting the data

The experiment carried out will be on the data set taken from the meteorological department and after collecting those data the data mining Ensemble technique will be applied to extract the pattern and then by the help of classifier those

pattern are trained on the classification model and then with the help of ensemble model it will be tested so that the accurate forecast in advance can be done to avoid the various problems and disaster that can be happened.

6. Proposed Methodology

In the proposed work the time series data set is analyzed to forecast rain precisely than the existing models. The work will be carried out in two different faces that is firstly collecting the data from weather forecast department from year 2006- 2016 and then applying the data mining ensemble techniques.

As data mining is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process will be extract information from a data set and transform it into an understandable structure for further use, Data mining is the process of extracting or mining knowledge from large amount 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. The basic rainfall prediction method as follows:

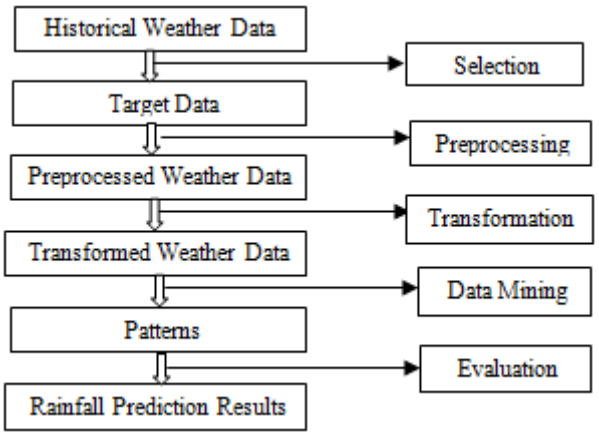

Figure: Overview of the forecasting model

The work proposed is based on Ensemble Prediction for prediction of the annual rainfall record of Chhattisgarh region by taking into account the data of all districts. This Ensemble computing approach is used to increase accuracy on your dataset and is to combine the predictions of multiple different models together. After the result achieved then finally by the help of WEKA tools the result will be compared and will be trained to give much better and accurate result

7. Conclusion

The conclusion that can be made after data collected should be analyzed and trained in a proper manner so that it can be tested by ensemble algorithm more efficiently in order to get the predicted result very nearness to a measured value or the standard set. The attempt of the work will be to analyze the time series data set inorder to forecast rain precisely than the existing model. The developed method will be targeted to be so simple that it can tested and validated without any complexity also in order to increase the accuracy of the model proposed by combining the model predictions with Ensemble Predictions so that it will helpful in the agriculture sector for increasing the productivity.

References

[1] P.Hemalatha, "Implementation of Data Mining Techniques for Weather Report Guidance for Ships Using Global Positioning System", International Journal Of Computational Engineering Research Vol. 3 Issue. 3 , march 2013.
[2] Elia Georgiana Petre "A Decision Tree for Weather Prediction", Buletinul, Vol. LXI No. 1, 77-82, 2009.
[3] Soo-Yeon Ji, Sharad Sharma, Byunggu Yu, Dong Hyun Jeong, "Designing a Rule-Based Hourly Rainfall Prediction Model", IEEE IRI 2012, August – 2012.
[4] S. Kannan , Subimal Ghosh, "Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output", Springer-Verlag, July- 2010.
[5] Kaya, E.; Barutçu, B.; Menteş, S. "A method based on the van der Hoven spectrum for performance evaluation in prediction of wind speed". Turk. J. Earth Science, 22, 1–9, 2013.
[6] Subana Shanmuganathan and Philip Sallis, "Data Mining Methods to Generate Severe Wind Gust Models", 5, 60-80, Atmosphere 2014.
[7] Gaurav J. Sawale, Dr. Sunil R. Gupta, "Use of Artificial Neural Network in Data Mining For Weather Forecasting", International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013.
[8] A.R.W.M.M.S.C.B. Amarakoon, "Effectiveness of Using Data Mining for Predicting Climate Change in Sri Lanka", 2010.
[9] Badhiye S. S., Dr. Chatur P. N., Wakode B. V., "Temperature and Humidity Data Analysis for Future Value Prediction using Clustering Technique: An Approach", International Journal of Emerging Technology and Advanced Engineering, 2250-2459, Volume 2, Issue 1, January 2012.
[10] Zahoor Jan, M. Abrar, Shariq Bashir, and Anwar M. Mirza, "Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique", Springer-Verlag Berlin Heidelberg, CCIS 20, 40.
[11] Meghali A. Kalyankar, S. J. Alaspurkar, " Data Mining Technique to Analyse the Metrological Data", International Journal of Advanced Research in Computer Science and Software Engineering 3(2), 114-118, February – 2013.

[12] Kavita Pabreja, "Clustering technique to interpret Numerical Weather Prediction output products for forecast of Cloudburst", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 3 (1) , 2996 - 2999, 2012
[13] Pinky Saikia Dutta, Hitesh Tahbilder, "Prediction Of Rainfall Using Data mining Technique Over Assam", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 5 No.2 Apr-May 2014.
[14] Simon S. Haykin, "Neural Networks: A Comprehensive Foundation", Second Edition, Prentice Hall International, 1999.
[15] Neha Khandelwal, Ruchi Davey, " Climatic Assessment Of Rajasthan‟s Region For Drought With Concern Of Data Mining Techniques", International Journal Of Engineering Research and Applications (IJERA), Vol. 2, Issue 5, 1695-1697,September- October 2012