IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

RAINFALL PREDICTION USING MACHINE LEARNING CLASSIFICATION ALGORITHMS

[1]B.Revathi, [2]C.Usharani

[1]Asst.Prof, [2]Asst.Prof
[1]Computer Science and Engineering,
[1]Mangayarkarasi college of Engineering, Madurai, India.

Abstract: Machine learning entails artificial intelligence, and it's far used in fixing many problems in facts technology. One common Machine Learning application is the prediction of an outcome based upon existing data. Rainfall prediction is important as heavy rainfall can lead to many natural disasters. The main challenge is to build a model for long term rainfall prediction from the training data set**.** Decision tree classification is one of the best Machine Learning Algorithms. A decision tree is looks like a tree structure. Many decision trees have been formulated. In this paper we predict the rainfall dataset using both CART and IDA decision tree algorithms. Using these algorithms which one provides highest predictive accuracy using performance measure.

***Index Terms* – Machine Learning, CART, IDA, Prediction.**

I. INTRODUCTION

Rainfall speculating is very important because high and irregular rainfall can have many issues like destruction of crops and farms, damage of property so a better speculating model is essential for an early warning that can reduce risks to life and property and also managing the agricultural farms in better way. These conventional methods cannot work in an efficient way so by using machine learning techniques we can produce accurate results. We can just do it by having the historical data analysis of rainfall and can predict the rainfall for future seasons. We can apply many methods such as, classification, regression according to the needed requirements and also we can calculate the error between the actual and prediction and also the accuracy. Various techniques provide different accuracies so it is important to choose the right algorithm and model it according to the requirements. Rainfall depends on humidity, temperature, pressure, wind speed, dew point, etc. Decision tree algorithm using Gini Index in order to predict the condensation with accuracy. The decision tree is to be fabricated and classification rules are then generated. To improve accuracy CART technique is applied to this result thereby obtaining a result with enlarged accuracy rate.

Machine learning (ML) is the study of computerized algorithms. Machine learning algorithms develop a model based on sample data, called as "training data", in order to make decisions. Machine learning can also be used in the projection systems. Considering the rainfall prediction example, to compute probability of a fault, the system will need to classify the available data in groups. One of the Supervised Machine Learning techniques is decision tree, where the data is continuously split according to a certain parameter (Max-depth, Min-samples and etc). The decision tree can be described by two entities, namely decision nodes and leaves. The leaves are the decisions and the decision nodes are where the data is split.

Classification is one of the data mining techniques used to predict group membership for data samples. There are two types of classification approaches.1.Rule based Classification 2. Decision tree based classification. Decision Tree is used for both classification and Regression problems is a tree-structured classifier, where internal nodes represent the characteristics of a dataset, branches represent the decision rules and each leaf node represents the output. In a Decision tree, there are two nodes, which are the Decision and Leaf. The decisions are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible results to a problem based on given conditions. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

II. LITERATURE SURVEY

In this survey various decision tree based classification algorithms are discussed and compared with the proposed work**.** Ayisha Siddiqua L, Senthil kumar N C [1], proposes a Data mining approach in order to predict rainfall based reflectivity of radar, in this paper uses data mining algorithms for some conditions like, temperature, humidity, pressure and etc from data to predict rainfall. Emilcy Hern´andez, Victor Sanchez-Anguix and Javier Palanca [2], they presented Forecasts of meteorological time series can help decision-making processes carried out by organizations responsible of disaster prevention auto encoder for reducing and capturing non-linear relationships between attributes, and a multilayer perceptron for the prediction task. Seung-HyunMoon , Yong-HyukKim, Yong HeeLee, Byung-RoMoon [3] proposes a method for an effective Early Warning System (EWS) for very short-term heavy rainfall with machine learning techniques. The EWS produces a warning signal when it is expected to reach the criterion for a heavy rain advisory. As a classifier, logistic regression is used to predict whether or not a warning is required. A comparative evaluation was performed on the EWS models generated by various classifiers. Vikrant Singh [4], the various models and techniques are developed to estimate rainfall in various researches using data mining techniques. The accurate and exact estimation of rainfall prediction and estimation of precipitation is not possible though many techniques are available. This Paper suggests various techniques of rainfall prediction and estimation and their results with the actual rainfall value. Charles X. Ling, Victor S. Sheng, and Qiang Yang [5], they proposed to design cost-sensitive machine learning algorithms to model this learning and diagnosis process. Medical tests are like attributes in machine learning whose values may be obtained at a cost and it is empirically evaluate these test strategies. The results can be readily applied to real-world diagnosis tasks. Ezekiel T. Ogidan, Kamil Dimililer, Yoney Kirsal Ever [6], they presented Machine learning being a powerful tool for automation can be merged with data science and analysis to make for a more effective faster way to analyze data. In this paper, an application of expert systems for data analysis. Nikhil Sethi, Dr.Kanwal Garg [7], Rainfall plays an important role in agriculture so early prediction of rainfall plays an important role in the economy of India. Rainfall prediction has been the one of the most challenging issue around the world in last year. Widely used techniques for prediction are Regression analysis, clustering, and Artificial Neural Network (ANN). This paper represents multiple linear regression (MLR) technique for the early prediction of rainfall. Siddharth s. Bhatkande ,Roopa G. Hubballi [8], they developed a model using decision tree to predict weather phenomena like full cold, full hot and snow fall which can be lifesaving information. Generally maximum temperature and minimum temperature are mainly responsible for the weather prediction. On the percentage of these parameters we predict there is a full cold or full hot or snow fall.

III. PROPOSED SYSTEM

For Rainfall prediction using Decision tree techniques Such as CART and IDA Algorithms are used. Decision Tree can be used for both classification and Regression problems, but mostly it is preferred for solving Classification issues. It is a tree-structured classifier. It is called a decision tree because, similar to a hierarchical structure, it starts with the root node, which expands on further split and constructs a tree-like structure. The datasets are taken to produce the list of rules from decision tree .They include Weather, Contact Lens, Rainfall Prediction and etc. This IDA Algorithm applied to predict the rules from decision tree based classification, which will provide the best list of rules instead of using list of rules. The Proposed System follows the CART Compared with IDA.

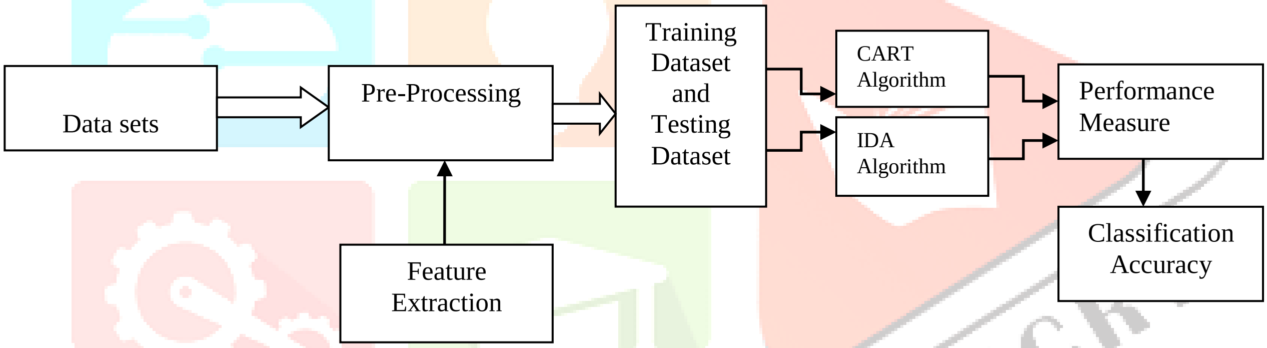

Fig 1. Block Diagram of Proposed Work

IV. CART

CART stands for classification and regression trees. It is a decision tree learning approach that outputs either classification or regression trees. CART is a supervised learning approach, since it is produce a labeled training dataset samples in order to construct the classification or regression tree model. Classification and Regression Trees software was used to develop models that can classify subjects into various risk categories. (e.g.,) presence to create a decision tree, with the goal of correctly classifying members of the population, each independent variable is examined and a split is made to maximize the vulnerability and individuality of the classification, resulting in a decision tree. The CART method is able control the complex communications among variables in the final tree, in contrast to identifying and defining the interactions in a multivariable logistic regression model. In CART Analysis, simple chance sampling without replacement was used to divide the sample into same sized developmental and validation samples.

The CART decision tree is a binary repeated scheduling procedure capable of processing continuous and nominal attributes as targets and predictors. Trees are extending to a maximum size without the use of an ending rule. The CART mechanism is calculated to produce not one tree, but a continuation of nested pruned trees, each of which is a candidate to be the proper tree. CART does not use training-data-based performance measure for tree selection. The CART mechanism includes automatic class evaluating and automatic missing value holding. CART algorithm is as follows,

```
INPUT: Dataset D
Step 1: Tree = {}
Step 2: MinLoss = 0
Step 3: for all Attribute k in D do:
        Step 3.1: loss = GiniIndex(k, d)
        Step 3.2: if loss<MinLoss then
                Step 3.2.1: MinLoss = loss
                Step 3.2.2: Tree' = {k}
Step 4: Partition (Tree, Tree')
Step 5: until all partitions processed
Step 6: return Tree
OUTPUT: Optimal Decision Tree
```

V. IDA

The divergence measure is used by the IDA algorithm and not the entropy measure. Divergence is defined as the act of diverging or the degree by which things diverge. There are various ways to define divergence for different types of variables. A few of them are, for probability distributions R and S of a discrete random variable their divergence is defined to be

$$Div\left(\frac{R}{S}\right) = \sum R(i)\log\left(\frac{R(i)}{S(i)}\right) \qquad (1)$$

In words, it is the average of the logarithmic difference between the probabilities R and S, where the average is taken using the probabilities R. The divergence is only defined if R and S both sum to 1.For distributions R and S of a continuous random variable; divergence is defined to be the integral:

$$Div\left(\frac{R}{S}\right) = \int_{-\infty}^{\infty} r(x)\log\left(\frac{r(x)}{s(x)}\right)dx \qquad (2)$$

Where r and s denote the densities of R and S. More generally, if R and S are probability measures over a set X, and S is absolutely continuous with respect to R, then the divergence from R to S is defined as,

$$Div\left(\frac{R}{S}\right) = -\int_{-x}^{x}\log\left(\frac{dS}{dR}\right)dR \qquad (3)$$

Where dS/dR is the derivative of S with respect to R, and provided the expression on the right-hand side exists. To take the dependency between attributes into consideration IDA utilizes the global dependency structure as a classification criterion. For the entire data set the global dependency structure is obtained and measured by the conditional divergence function. The locally best solutions calculated may be substituted by other neighboring alternatives as required to better the classification performance globally. The IDA uses the nearest neighbor dependency .An attribute is selected on the basis of individual classification effect and also its combined classification effect with other attribute using the look ahead method. If their combined classification effects with other attributes are poor, then even the individually best attributes will not be selected through this process. Given a collection of n objects A= (a1, a2, ….. , an, C) where C is the class to which an object belongs ,the

Intelligent Decision-tree Algorithm:

Step 1: The divergence measures Div (a_j | $a_{d)}$ is computed for each value a_d with the remaining attributes a_j. The largest Divergence measure Div (a_j^L|a_d) is selected structure.

Step 2: Average divergence measure E (ad)[Div(a_j^L|a_d)] =P(ad) Div(a_j^L|a_d) is computed for each value of $a_{i\,,}$ a_d

Step 3: The first node in the tree is the attribute with the largest divergence measure. E [Div (ajL|a_i)]

Step 4:Sub-nodes are created for each value a_d of a_i.

Step 5: The attribute a_j with the largest divergence measure Div (ajL | ad), will be next attribute for each non terminal sub node. Which in turn creates sub-nodes a_{js}. Repeated checks to be performed

Step 6: For each nonterminal sub-node a_{js} generate the next attribute a_k. Which is checked against ad its grandparent node. If attributes ah and a_i are the same or attribute a_k and a_i are too closely related, go back to Step (6) to find other neighboring Alternatives else proceed to Step (7). Attribute a_k is closely related to a_d if Div (a_k | a_d) < i where i, is the median of {Div(a_k |a_d) k = 1,. . . m} .

Step 7: All attributes with conditional divergence measures greater than or equal to i, in the neighborhood are to be considered. Right from the nearest neighbor with the closest divergence measure till the end. If this attribute is not dependent on its grandparent node too closely and has also not been used, continue, otherwise find the next nearest neighbor until an applicable attribute is found. The last attempted attribute is chosen if no neighboring attribute is found to be applicable.

Step 8: Sub-nodes are created and terminal sub-nodes are marked.

Step 9: For each of the remaining nonterminal nodes go to Step (5).

VI. EXPERIMENTAL RESULTS

The major difference between the CART and IDA is to is to provide the best list of rules instead of producing the list of rules using the decision tree algorithm. By compared with IDA algorithm, CART provides good performance results. Gini Index is measured by,

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

In IDA, we trained the dataset by using the following attributes.

Table 6.1: IDA Cross validation

Splitter	Max_depth	Min_samples	Percentage
300	4	0.04	45
500	2	0.06	50
400	5	0.08	35
200	8	0.05	45
250	6	0.20	40
300	3	0.15	75
250	7	0.10	45

Where p_i is the possibility of an object being classified to a specific class, while constructing the decision tree, we would prefer choosing the attribute/feature with the least Gini index as the root node.

$$Q = \frac{True\ Positive}{Covered}$$

Here Q is the Quality. In this, number of correctly classified records and total number of records which is used in the performance measure calculation. Here we are taken the samples to the cross validation. For example, IDA, CART analysis is weather dataset. In this there are 3 Attributes are used. They are Play, temperature, humidity, outlook, windy. The first attribute is predictable one. Rule can be generated either from a decision tree or directly from the training data using CART algorithm. In this algorithm using the decision based classification. The graph represented as the analysis of both IDA and CART. For the analysis process 4 parameters have been used.

Table 6.2: CART Cross validation

Splitter	Max_depth	Min_samples	Percentage
350	5	0.02	80
500	2	0.07	50
250	4	0.10	30
100	3	0.25	45
300	7	0.10	60
400	6	0.08	60
250	2	0.20	75

Final Cross validation Comparison results of IDA and CART algorithm will be followed as,

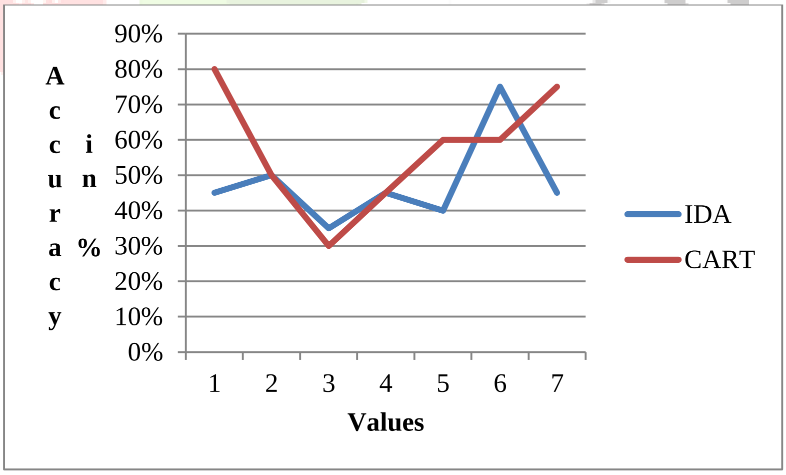

Fig 2. Comparison Results

VII. CONCLUSION

The proposed method consists of two modules i.e. IDA and CART. At first the IDA algorithm is applied for discovering the list of rules from the decision tree. By using this approach the best rule is found by the CART. The CART has been tested for various dataset records, which supports only nominal and continuous attributes. CART is used to provide smaller number of greater rules. Rules are uncovered once at a time, the resulting rule will not affect the previous set of rules because the search space is changed due to the removal of training examples covered by the previous rules. That is why the accuracy of CART is greater than that of IDA. Both the modules are completed and it is concluded that the CART is the most accurate algorithm, which achieves statistically and significantly higher predictive accuracy than IDA.

REFERENCES

[1] L.Ayisha Siddiqua , N.C Senthil kumar , “Heavy Rainfall Prediction using Gini Index in Decision Tree”, Volume-8, Issue-4. ISSN: 2277-3878, Nov 2019.

[2] B. Emilcy Hern´andez, Victor Sanchez-Anguix, Vicente Julian,Javier Palanca, and N´estor Duque,” Rainfall Prediction: A Deep Learning Approach”, this work is partially supported by the MINECO/FEDERTIN2012-36586-C03-01 of the Spanish Government,2012.

[3] Seung-Hyun Moon, Yong-Hyuk Kim, Yong Hee Lee, Byung-Ro Moon,” Application of machine learning to an early warning System for very short term heavy rainfall”, 568 1042-1054, 2019.

[4] Vikrant Singh,” Study of Various Rainfall Estimation &Prediction Techniques using Data Mining”, ISSN: 2278-0181, Vol. 9 Issue 07, July-2020

[5] Charles X. Ling, Victor S. Sheng,” Cost-Sensitive Learning and the Class Imbalance Problem”, Springer. 2008

[6] Ezekiel T. Ogidan, Kamil Dimililer, Yoney Kirsal Ever,” Machine Learning for Expert Systems in Data Analysis”, DOI 10.1109/ISMSIT.2018.8567251,Oct 2019.

[7] Nikhil Sethi, Dr.Kanwal Garg,” Exploiting Data Mining Technique for Rainfall Prediction”, ISSN: 0975-9646, Vol. 5 (3), 3982-3984, 2014.

[8] S.Siddharth,Bhatkand,G.Roopa,Hubballi,”Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques”, ISSN (Online) 2278-1021, Vol. 5, Issue 5, May 2016