A Comparative Study of Classification Algorithms for Forecasting Rainfall

Deepti Gupta
USICT, G.G.S Indraprastha University
New Delhi, India
deeptigupta72@gmail.com

Udayan Ghose
USICT, G.G.S Indraprastha University
New Delhi, India
udayan@ipu.ac.in

***Abstract* - India is an agricultural country which largely depends on monsoon for irrigation purpose. A large amount of water is consumed for industrial production, crop yield and domestic use. Rainfall forecasting is thus very important and necessary for growth of the country. Weather factors including mean temperature, dew point temperature, humidity, pressure of sea and speed of wind and have been used to forecasts the rainfall. The dataset of 2245 samples of New Delhi from June to September (rainfall period) from 1996 to 2014 has been collected from a website named Weather Underground. The training dataset is used to train the classifier using Classification and Regression Tree algorithm, Naive Bayes approach, K nearest Neighbour and 5-10-1 Pattern Recognition Neural Network and its accuracy is tested on a test dataset. Pattern Recognition networks has given 82.1% accurate results, KNN with 80.7% correct forecasts ranks second, Classification and Regression Tree(CART) gives 80.3% while Naive Bayes provides 78.9% correctly classified samples.**

***Keywords*: Rainfall Prediction, Decision tree, Naive Bayes, K Nearest Neighbour, Neural Network**

I. INTRODUCTION

Rainfall is a form of precipitation. Its accurate forecasts can help to identify possible floods in future and to plan for better water management. Weather forecasts can be categorized as: Now forecasts which is forecasts up to few hours, Short term forecasts which is mainly Rainfall forecasts is 1 to 3 days forecasts, Forecasts for 4 to 10 days are Medium range forecasts and Long term forecasts are for more than 10 days. [1] Short range and Medium Range rainfall forecasts are important for flood forecasting and water resource management. Data Mining or knowledge discovery is process of finding facts which are not known. Classification is a supervised learning process which lies under the umbrella of Data Mining. It is used as model to distinguish samples with unknown class labels on the basis of their similarities and dissimilarities and predict a class label for them. Classification has been applied in many fields like for detecting frauds by banks and companies, by various service providers to predict their performance in future, to classify patients on the basis of their symptoms. The Decision tree (CART), Naive Bayes approach, K-Nearest Neighbour and neural networks has been used in this paper to forecast the Rainfall to class label y or n. The concept of Gini Index is used to create classification trees and Naive Bayes uses Bayes theorem for predictions. Nearest neighbours have been computed using Euclidean distance and neural network uses SCG descent as Backpropagation algorithm for learning. A Brief overview of past Related work is mentioned in Section II, Premise have been explained in Section III, Process and Outcomes are mentioned in Section IV and Conclusion and Future Work in Section V.

II. RELATED WORK

Timely rainfall forecast is a big challenge and requirement for a country like India. Various rainfall forecast models have been created in past. Numerical Weather forecasting model is used to generate Short range and Medium range forecasts which uses partial differential equations to conduct multiple numerical predictions. Ensemble forecasting is a numerical weather prediction model that uses different forecasts model with slightly different input to gather better forecasts.

A Bayesian approach based model for rainfall prediction is created by author [1] where posterior probabilities are used to calculate likelihood of each class label for input data instance and the one with maximum likelihood is considered resulting output.

The author [6] used regression based statistical model for rainfall prediction. Five years data is taken as input to compute rainfall using Karl person coefficient which is then compared with rainfall statistics for future years predicted using multiple regression technique.

C. E. Brodleyf and M. A. Friedl [7] have used Decision tree classification algorithm to classify land cover for different vegetations using remotely sensed data. They have described three different types of decision trees. Univariate trees tests single attribute at any particular node of tree whereas multivariate tree uses more than one attribute for testing condition at branch while splitting. Hybrid trees are heterogeneous trees as they use more than one algorithm to build the tree.

John Mingers [8] has provided a comparison of various pruning methods available for improving decision tree classification. Error complexity method by Breiman for CART trees gives a set of pruned trees and one with lowest misclassification rate is considered final tree. Pessimistic Pruning by Quinlan is used to prune trees build by C4.5 algorithm.

Martin Fodslette Moller [12] in his original contribution paper has given a faster supervised learning algorithm for

978-1-4673-7231-2/15/$31.00 ©2015 IEEE

training the network. He has described other algorithms proposed by other authors for the same purpose. Martin has finally described his algorithm Scaled Conjugate Gradient algorithm, its advantages over past related algorithms and show results of algorithms on parity problem. Standard backpropagation algorithm BP is based on user dependent parameters learning rate and momentum and uses constant step size which makes algorithm less robust as there is no basis to decide them. Conjugate gradient algorithm performs line search to find step size that is how much to move in a particular direction to find next weight vector. Another way to find step size is to compute Hessian matrix that stores second derivative of error function and it requires $O(N^2)$ memory and $O(N^3)$ calculations for this purpose. Also it becomes indefinite at some points in weight space. Another past algorithm is BFGS which also provides better performance than standard BP and works on the same approach of conjugate directions as CG but the way to find direction involves certain constants like Scaling factor. SCG is independent of parameters like learning rate and momentum and also does not perform line search as CG and BFGS. SCG is a combination of conjugate gradient with Levenberg's model trust approach. Scalars sigma and lambda are used to compute step size and weight is updated to vector when gradient that is derivative of error function becomes zero.

The authors [14] have used multilayer neural networks to model disease of HIV in South Africa. They have worked on 8000 samples from antennal clinic to classify whether a patient is HIV positive or negative. They have discussed and show results for by varying number of neurons in hidden layer and number of epochs for training dataset. Also they have performed sensitivity analysis in order to specify that which input factor have how much effect on modelling the output. Increasing the number of neurons in hidden layer gives more accurate prediction results but up to a certain amount that is 10 is decided as appropriate value for this.

The author [15] has used ensemble of Backpropagation neural network, Radial basis function neural network and General regression neural network to forecast rainfall in Sri Lanka. They have worked on dataset of 41 years using 26 predictor variables as input. The performance of all the three types of networks has been compared with the performance of the network. BPN architecture contains 26, 10 and 1 neuron in input, hidden and output layer respectively. Sigmoid function is used as activation function for hidden and output layer. RBF neural network has been designed with 26, 73 and 1 nodes in three layers. The third network that is GRNN is designed with 26 nodes in input layer, hidden layer contains nodes one for each training sample, summation layer contains numerator and denominator layer, while output layer contains in node to divide output of numerator and denominator node. Weighted average method is used to decide number of each type of networks in ensemble. Individually GRNN gives best results for the performance metrics mean absolute error and root meansquare error, while Ensemble of eight BPNN, 1 GRNN and 2 radial basis neural networks gives best prediction results.

III. PREMISE

The process of classification is used to analyze and classify historical and current data to predict future data trends. Classification is concerned with problems in which resulting class labels are categorical that is discrete, unordered. The process of classification starts with applying learning algorithm to the training dataset which results into classification rules. These rules are then applied to a test data to compute the accuracy of the algorithm. The accuracy of a classifier (or learning algorithm) is percentage of samples or instances that are correctly classified. The data classification method is explained using Fig. 1.

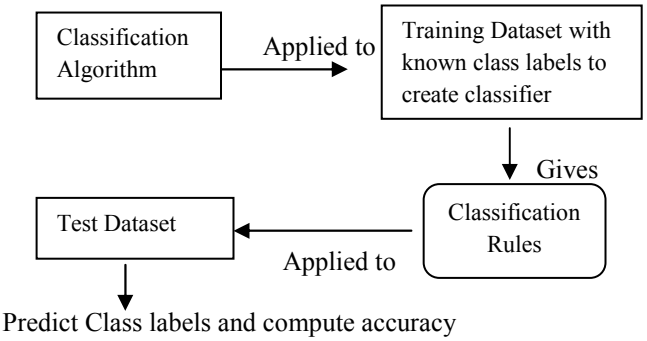

Fig. 1: Classification Process

A. Classification and Regression Trees

Decision tree classification is a non parametric technique which uses a tree like structure to classify the input dataset instances. Each sample is assigned a class label by moving in a top down manner from root and testing the condition at the branch. It is based on greedy approach as best split is used at each step to build the tree. Decision trees are easy to work with and even with low domain knowledge. Decision trees can be drawn using ID3 algorithm, C4.5 and CART. ID3 and C4.5 were given by Quinlan where C4.5 is a successor of ID3. C4.5 adds some improvements to ID3 like pruning of tree to avoid overfitting of data, handling of missing values. ID3 uses Information gain to find best split while C4.5 uses Gain ratio. Classification and Regression trees (CART), an algorithm by Breiman uses Gini index as selection measure to find best attribute for splitting and constructing binary decision tree.[5] Classification trees are built when resulting class label is categorical like y or n whereas regression trees are used for numerical class labels like income of a person, rainfall in millimetres. CART works well with continuous values. CART uses Gini index to measure the impurity in a training dataset S as:

$$Gini(S) = 1 - \sum_{i=1}^{m} p_I \qquad [1]$$

Where m is the number of class labels. For continuous valued attributes as in this case, first the values for each of the features is sorted in increasing order. Then Gini index is

calculated at midpoint of each of the two adjacent values of the feature using equation 2.

$$Gini_A(S) = |S_1|/|S|Gini(S_1) + |S_2|/|S|Gini(S_2) \quad [2]$$

The one with the minimum impurity measure is considered best option. The process is repeated for all the attributes. The attribute with the maximum $\Delta Gini(A)$ is considered best splitting attribute where

$$\Delta Gini(A) = Gini(S) - Gini_A(S) \quad [3]$$

The classification tree keeps on growing until all the samples belong to same class label or all the attributes are used for splitting. Now next is to prune the tree to handle overfitting. Overfitting occurs when classifier is specific to training dataset and makes many classification errors on another dataset. CART uses cost complexity method by Breiman to decide the final decision tree from among the series of trees produced by the method. Firstly compute the resubstitution and cross validation errors for all the subsets of the tree. Then the tree having misclassification error within one standard deviation of minimum cost is considered final tree.

B. Naive Bayes Approach

Bayesian approach for classification is a statistical and linear classifier which predicts class label for data instance on the basis of distribution of attribute values. This is a parametric classification where the size of classifier remains fixed. Distribution can be normal (Gaussian), kernel, multivariate or multi nominal. Assuming normal distribution for weather data, Bayesian classifiers use Bayes theorem to find posterior probabilities of occurrence of input data instance in all classes. Class label having maximum conditional probability is assigned to data instance. Naive Bayes assumes that attributes have no effect on each other that is they have independent distribution of values. The algorithm [11] starts with computing prior probability, $P(C_i)$ for each class. Then for an input data instance q, compute posterior probabilities for each class as given in equation 4.

$$P(q|C_i) = \prod_{j=1}^{n} P(F_j|C_i) \quad [4]$$

Where F_j are attributes for input sample q. Finally compute $P(C_i|\, q)$for each class using:

$$P(C_i|q) = P(q|C_i) * P(C_i)/\sum_{i=1}^{m} P(C_i) \quad [5]$$

The class C_i having maximum probability $P(C_i|q)$ is the predicted class label for input sample q.

C. KNN

K nearest neighbours are lazy learners where learning is based on analogy. The algorithm cannot be used until the sample for which neighbours are to be computed is available. K nearest neighbours are computed for given instance t for which class label is to be predicted. The attributes or features are preferred to be numeric in nature as neighbours are computed using the distance metric. Euclidean distance is used here for analysis as distance metric which is calculated as equation 6:

$$d(Y,Z) = \sqrt{\sum_{i=1}^{n}(y_i - z_i)^2} \quad [6]$$

Where Y is an n dimensional observation from training dataset and Z is an n dimensional observation from test dataset for which class is to be predicted. Majority voting is then used to take decision for the instance. The class to which maximum neighbours belong is considered for the input sample. The value of K that is number of neighbours depends on the size of training dataset.

D. 5-10-1 PRNN

An artificial neural network is a set of neurons arranged according to a specific architecture. Neurons are processing elements that transform input given to network into an output. Each of the connections between the neurons of one layer to another layer is assigned a particular weight which signifies its importance. A bias is available for each unit in the hidden layer and the output layer which acts as a threshold to vary the activity of neurons in the network. Pattern Recognition NN (PRNN) used here for analysis is feed forward two layer network which is used for classification to classify samples into one of the target class labels. The number of neurons in hidden layer depends on number of input and output. One neuron in output is used to represent two target class labels in training data in form of 0(n) and 1(y).

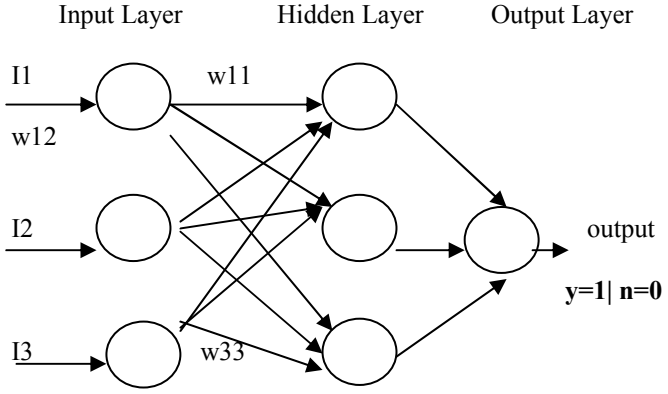

Fig. 2: Two Layer Neural Network

The inputs are fed to the units in the input layer. The output from input layer is same as input fed into network. Then for each unit k in the hidden layer, the input is sum of the product of weights incident on the unit and output from input layer which is computed as in equation 7:

$$I(k) = \sum_j w\,(jk)O(j) + \theta(k) \quad [7]$$

And $\theta(k)$is the bias for unit k. Log Sigmoid is used as an activation function which is applied on the neurons in the hidden and the output layer to compute their output as:

$$O(k) = 1/(1 + e^{-I(k)}) \quad [8]$$

Squashing (activation) function is used to limit the output from connectionist network. It is a differentiable and nonlinear function. [9] After class labels are available, backpropagation algorithm is used to learn the network. Backpropagation algorithm is a neural network training algorithm which works in reverse direction to adjust connection weights and unit biases in each of the iteration to create an optimized network. Scaled Conjugate Gradient (SCG) algorithm has been used to train the network. SCG, given by M.F Moller [12] is faster than standard BP algorithm.. A number of epochs or iterations are consumed to find optimized network each time minimizing the Mean Square Error. Mean square error is the difference of actual values and predicted values computed over all samples as:

$$mse = 1/N \sum_{i=1}^{N} (av_i - pv_i)^2 \quad [9]$$

Neural networks are difficult to interpret and accurate classifications using networks largely depends on architecture of network. Deciding number of layers and number of units in each layer is trial and error process. They work well even with noisy data.

IV. PROCESS AND OUTCOMES

For creating the various classifiers, a dataset with 2245 samples of New Delhi for rainfall period (June to September) for 19 years has been used [13]. The five factors which affect rainfall and required for forecasting are taken as input and mentioned in TABLE I.

TABLE I : INPUT FEATURES

Feature	Type	Units
Mean Temperature	Numerical	In Degree Celsius
DewPoint Temperature	Numerical	In Degree Celsius
Humidity	Numerical	In Percentage
Sea Level Pressure	Numerical	In hecto Pascal
Wind Speed	Numerical	In km/hr

For all the techniques except PRNN, number of records in training dataset and test data set are in the ratio of 70: 10. While for PRNN, dataset of 2245 records is divided randomly into training set, validation set and test set with 70%, 15% and 15% samples respectively to be used while learning. The accuracy of classifier is tested on 280 samples. Training dataset is used to train the network, validation set to avoid overfitting and test samples to test the network for accuracy. For the classification tree various pruning levels, cost or misclassification error at those levels, its standard deviation and number of terminal nodes in pruned tree is mentioned in TABLE II.

TABLE II: COST AT DIFFERENT PRUNING LEVELS

Pruning Level	Cost	Standard Deviation	Number of Terminal Nodes In Pruned Tree
0	0.2102	0.0085	162
1	0.2020	0.0081	144
2	0.2010	0.0081	139
3	0.2015	0.0081	135
4	0.2015	0.0081	129
5	0.2000	0.0080	116
6	0.1964	0.0078	114
7	0.1944	0.0077	105
8	0.1939	0.0076	101
9	0.1924	0.0076	96
10	0.1919	0.0076	80
11	0.1919	0.0076	70
12	0.1913	0.0076	55
13	0.1913	0.0075	45
14	0.1954	0.0074	39
15	0.1859	0.0074	21
16	0.1852	0.0068	13
17	0.1847	0.0069	11
18	0.1868	0.0069	6
19	0.1842	0.0068	4
20	0.1832	0.0070	2
21	0.2224	0	1

A plot of misclassification error against number of terminal nodes is represented in Figure 3.The dotted line in figure represents the cut off. The cost and number of terminal nodes of final optimized tree will lie within this cut off and marked by black circle. Cut off is decided by summing minimum cost and standard deviation of it. So cut off is calculated as .1902. Trees at pruning level 15 to 20 as given in above Table II has cost below the cut off. Hence the best pruning level is 20 and the pruned tree has 2 terminal nodes and 0.1832 as cost.

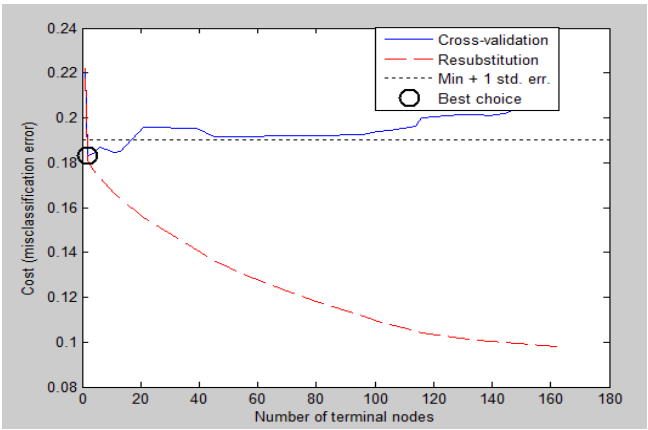

Fig.3: Plot showing best choice for number of terminal nodes

The final pruned tree is shown in Fig. 4. The condition at the branches of the tree shows the splitting condition and Relative Humidity(x3) is used as root node.

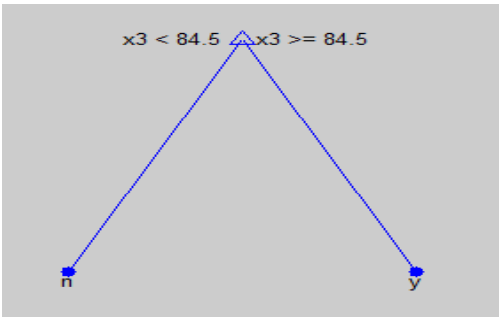

Fig.4: Pruned Decision tree (x3=Humidity)

Extracted Rules from Decision tree for classification

1. If x3<84.5, then class=n
2. If x3>=84.5, then class=y

The K nearest neighbours are computed using Euclidean distance. Two values for K have been used to find nearest neighbours. For K=22 and sample = [28, 26, 94, 1003, 5],nearest neighbours are marked using circles and sample is represented by cross sign in red colour in a plot of Relative Humidity against Temperature given in Fig. 5. The value of K is taken as $\sqrt{N}/2$ where N is number of training samples. As majority of neighbours to sample belong to class y, so sample is assigned y as class label.

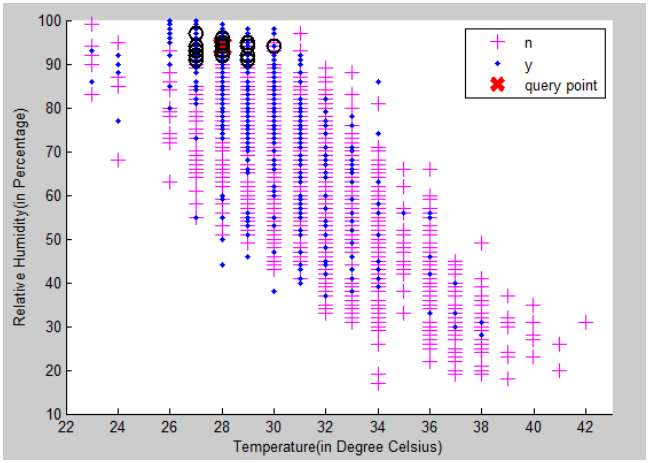

Fig.5: KNN plot for Test sample = [28, 26, 94, 1003, 5] and K=22

For rainfall prediction using PRNN, network has been implemented with two layers, 1, 2, 5 and 10 neurons in hidden layer and one neuron in output layer. One neuron is enough to model two class labels, 1(for label y) and 0(for label n). The network is trained using Scaled Gradient Descent algorithm where the connection weights and unit biases are adjusted to optimize the network and minimize the Mean Square Error. Performance is measured in terms of Mean square error. It takes 24 epochs to train the network with 10 neurons in hidden layer and best performance for validation set is 0.11261 at epoch 17.The time complexity of the network depends on the number of weights and biases.

The performance plot of Mean Square Error (MSE) against each of the epochs for training, validation and test dataset is shown in Fig. 6. The minimum MSE for validation set and its corresponding epoch number is shown using dotted line. Best Validation performance will be at point where validation curve intersects with dotted line.

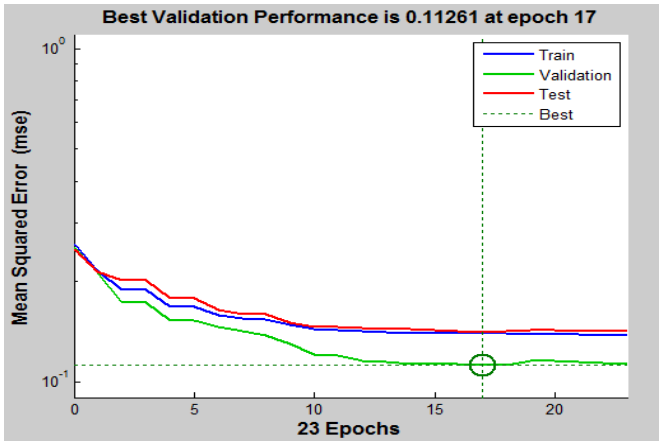

Fig.6: Performance plot for Datasets

Confusion matrix for all the algorithms has been presented in Fig. 7. Confusion Matrix is the n x n matrix where n is number of outputs or class labels. It shows the number of correctly classified and misclassified samples for all the class labels. The values in dark coloured boxes represent the number and percentage of correctly classified samples.

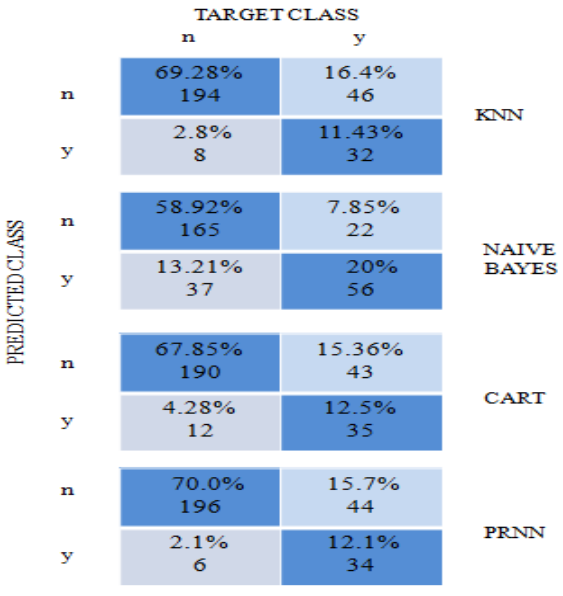

Fig. 7: Confusion Matrices

Size of training and test dataset along with final analysis results for all the algorithms has been summarized in TABLE III. It can be easily seen that PRNN gives best results among all.

TABLE III: SUMMARIZED RESULTS FOR ALL ALGORITHMS

Algorithm	Number of Training Samples (N)	Number of test samples	Number of Correctly Classified Samples	%
KNN, K=44 $(\sqrt{N})$	1965	280	224	80
KNN, K=22 $(\sqrt{N/2})$	1965	280	226	**80.7**
Naive Bayes	1965	280	221	**78.9**
CART(before pruning)	1965	280	200	71.4
CART(after pruning)	1965	280	225	**80.3**
PRNN(1 neuron in HL)	1571	280	225	80.3
PRNN(2 neuron in HL)	1571	280	227	81.07
PRNN(5 neuron in HL)	1571	280	230	82.1
PRNN(10 neuron in HL)	1571	280	230	**82.1**

Forecasting Results are plotted in Fig.8 using Bar Chart.

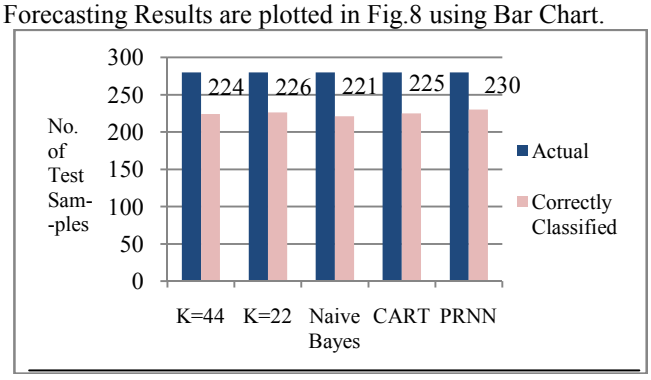

Fig. 8: Plot of Forecasting Results

V. CONCLUSION AND FUTURE WORK

Classification as part of data mining is very useful for finding unknown patterns like forecasting the future trends. Value for K in K Nearest Neighbour technique is difficult to determine. A prediction result for two different values of K has been presented in this paper. More accurate forecasts are made for $\sqrt{N/2}$ than $\sqrt{N}$ (as discussed by author [10]) where N is number of records in training set. Naive Bayes and Decision trees are easy to understand and work with but pruning the tree using cost complexity method requires intensive calculations and is time consuming. Neural networks provide better results than any of the other discussed algorithms. Best predictions are found with 10 neurons in hidden layer. It can handle noisy and continuous values better than CART. Despite of its advantages, deciding number of neurons and training network multiple times is time consuming and a difficult task. As all algorithms discussed have their advantages and limitations, it is difficult to conclude for best algorithm. Ensemble forecasting where algorithms are combined in a single model will be used to provide more accurate forecast in future.

REFERENCES

[1] B.B Meshram, Valmik B Nikam, *Modelling Rainfall Prediction using Data Mining Method, A Bayesian Approach*,Fifth International Conference on Computational Intelligence, Modelling and Simulation,2013

[2] Thair Nu Phyu, *Survey Of Classification Techniques in Data Mining*, Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong 2009 Vol I IMECS 2009, March 18 - 20, 2009

[3] Angus Ng, Bing Liu, Dan Steinberg, David J. Hand, Geoffrey J. McLachlan, Hiroshi Motoda, J. Ross Quinlan, Joydeep Ghosh, Michael Steinbach, Philip S. Yu,Qiang Yang, Vipin Kumar , Xindong Wu , Zhi-Hua Zhou, *Survey Paper-Top 10 algorithms in data mining*, Published online: 4 December 2007 © Springer-Verlag London Limited 2007

[4] Rajesh Kumar, Decision Tree for the Weather Forecasting, International Journal of Computer Applications, Volume 76-No.2, August 2013

[5] Jiawei Han, Micheline Kamber , *Data Mining- Concepts and Techniques*, Second Edition

[6] M. Kannan, P. Ramachandran, S. Prabhakaran, *Rainfall Forecasting Using Data Mining Technique*, International Journal of Engineering and Technology Vol.2 (6), 397-401,2010

[7] C. E. Brodleyf, M. A. Friedl, *Decision Tree Classification of Land Cover from Remotely Sensed Data*, Remote Sens. Environ. 61:399-409(1997) @ Elsevier Science Inc. 1997

[8] John Mingers, *An Empirical Comparison of Pruning Methods for Decision Tree Induction*, Machine Learning, 4, 227-243 (1989), Kluwer Academic Publishers, Boston © 1989

[9] Simon Haykins, *Neural Networks, A Comprehensive Foundation*, Second Edition

[10] David G. Stork, Peter E. Hart , Richard O. Duda, *Pattern Classification*, Second Edition

[11] Kotagiri Ramamohanarao, *Pattern Based Classifiers*, World Wide Web archive , Volume 10 Issue 1 , March 2007

[12] Martin Fodslette Moller , *A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning*, Neural Networks,Vol 6 , 13 November 1991

[13] http://www.wunderground.com

[14] Philip Pretorius, Wilbert Sibanda, *Novel Application of Multi-Layer Perceptrons (MLP) Neural Networks to Model HIV in South Africa using Seroprevalence Data from Antenatal Clinics*, International Journal of Computer Applications (0975 – 8887) Volume 35– No.5, December 2011

[15] Harshani R.K. Nagahamulla, Uditha R. Ratnayake, Asanga Ratnaweera, *An Ensemble of Artificial Neural Networks in Rainfall Forecasting*, The International Conference on Advances in ICT for Emerging Regions - ICTer 2012: 176-181