COMPARATIVE STUDY OF VARIOUS SUPERVISEDCLASSIFICATION METHODSFORANALYSING DEFORESTATION FACTORS

S.P.Rajagopalan[1], C.Lalitha[2]

[1]*Professor, G.K.M College of Engineering &Technology, Chennai, TamilNadu*
[2]*Research Scholar, Vels University, Chennai, TamilNadu*

Abstract

In this paper, various supervised classification techniques are compared and the results are demonstrated. Here the classification techniques like Decision tree, Bayesian method, Neural Networks and Rule Based method are discussed with regards to the data sets given. The population, built up and agriculture are the factors that play a vital role in the development of country which directly affects economic condition. In this paper, factors such as road , population, built up development ,agriculture and industry are considered as drivers of deforestation in the study area which is located in the Erode District of TamilNadu, India.

Keywords: *Supervised Classification, Decision tree, Bayesian method, Neural Network.*

----------***----------

1. INTRODUCTION

There are two types of classification, supervised and unsupervised. Supervised methods classify the data which is known and observed by the user specifically. Unsupervised methods are classified unknowably. The results are obtained with the given data sets by using the WEKA. The aim of this paper is to compare the various supervised methods by using the factors such as demographic, built up, road and agriculture. The primary methods used in Data mining are Data selection, data reduction and filtration. Data mining examines and discovers various algorithms under several computational efficiency. It integrates machine learning, pattern recognition, statistics, databases, and visualization techniques into one so that the information can be extracted from the large databases.

The tasks of data mining are association rules mining, classification, prediction and cluster analysis. Generally speaking, association rule mining and classification rule mining are the most effective and efficient techniques in data mining. Classification rule mining is used for the prediction future objects whose class label is not known. Recently it has been determined that primary factor for the degradation of ecosystem is deforestation of forests.

Classification results are basis for interpretation, analysis and modelling for various environmental and socio-economic applications. Data mining techniques can be applied for generating the class association rules for analysing the deforestation. In this paper, we applied various supervised classification techniques with our data sets.

2. WEKA

The Waikato Environment for Knowledge Analysis (WEKA) is a tool for machine learning algorithms which can be used for classification and clustering. In this paper, we use decision tree methods like J48 and Random forest. Bayes algorithm such as Naive Bayes and Neural Network methods like Multilayer perceptron are implemented in WEKA. We divided our data set into 10 cross validation folds and all the methods are tested and compared according to the data.

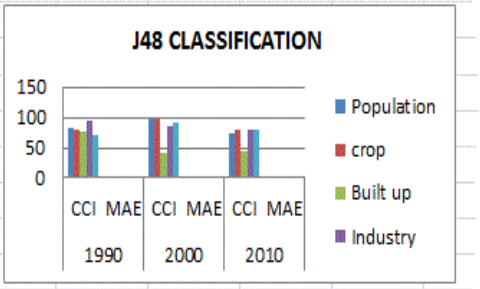
3. CLASSIFICATION METHODS

3.1 Decision Tree

Decision Tree (DT) worksaccording to the processing and deciding upon attributable data. Here attributes in DT are considered as nodes and each leaf node as a class. J48 and Random Forest were used in our experiments. It follows a recursivemethod for a given set of data.It searches the attributes as Depth-first strategy. It divides the class into several nodes and tests each node that gives the best result.It classifies the datasets invariably .This method is not suitable for finding anddid not show good results to the given datasets.Accuracy value of J48 methods is compared and shown in Table 1 and Figure1.

J48 CLASSIFICATION METHOD

	1990		2000		2010	
Forests	CCI	MAE	CCI	MAE	CCI	MAE
Population	83.95	9E-04	96.29	2E-04	74.07	0.002
crop	79.01	0.002	97.28	2E-04	80.24	0.002
Built up	78.02	0.002	41.48	0.006	45.18	0.005
Industry	94.32	0.012	84.44	0.01	78.76	0.011
Road	69.87	0.002	92.83	4E-04	78.51	0.001

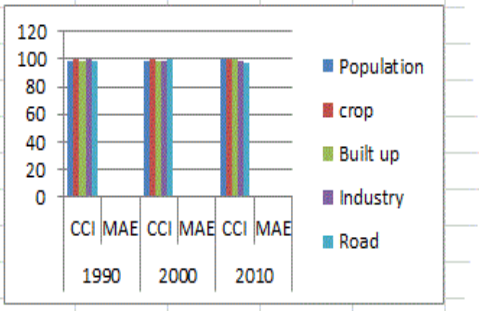


Random forest is also called regression trees that induces the data from bootstrap samples of the training data.It uses random feature selection by induction process. Itis comparatively gives better result than CART and C4.5.

It shows a better performance, after modelling the result. The disadvantages of DT are focus on continues attributes, computational efficiently with growing tree size. According to comparison provided for different classification methods in emotion recognition, Random Forest is the best classifier method on that group with 5 attributes and the results are compared and shown in Table 2 and Figure 2.

RANDOM FOREST CLASSIFICATION METHOD

	1990		2000		2010	
Forest	CCI	MAE	CCI	MAE	CCI	MAE
Population	99	0.002	99.01	0.002	99.5	0.002
crop	99.75	0.002	99.5	0.003	100	0.003
Built up	99.25	0.003	99.25	0.003	99.75	0.002
Industry	99.75	0.004	98.51	0.004	98.51	0.004
Road	99.25	0.002	99.5	0.002	97.03	0.002



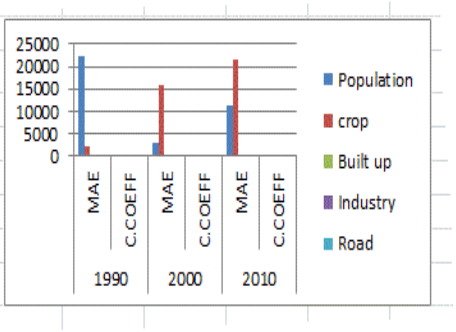
3.2 Artificial Neural Networks

3.2.1 MLP

Artificial Neural Network (ANN) is the common classification methods in data mining. Neural Network based classifiers, Multi Layer Perceptron (MLP) and Radial Base Function (RBF)were used in this work. MLP is a feed forward network that makes a model to map input data to output data. Hidden layer in MLP can include various layers between input and output. It classifies 3 factors gradually. The accuracy result of MLP is shown in Table 3 and Figure 3.

MULTILAYER PERCEPTRON CLASSIFICATION METHOD

	1990		2000		2010	
Forests	MAE	C.COEFF	MAE	C.COEFF	MAE	C.COEFF
Population	22347	0.228	2864.9	0.3141	11191	0.2381
crop	2417	0.292	15707	0.2553	21556	0.1219
Built up	68.017	0.1542	113.16	0.379	145.75	0.3932
Industry	2.21	0.1388	3.187	0.1746	8.3707	0.0617
Road	68.017	0.1542	355.67	0.1538	349.82	0.1307

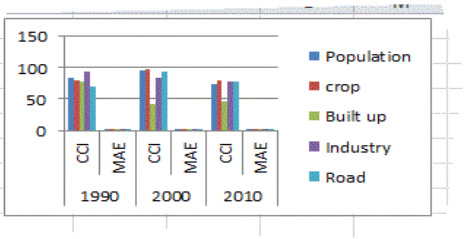


3.2.2RBF

RBF is another type on ANN. The input is linear and the output is nonlinear. Here it hides several non sequential values because of random input. At first it shows a good result for the year 1990 and 2010, but not for 2000.So that this algorithm is not suitable for the given datasets. The RBF networks are divided in two feed-forward layer is shown in Table 4 and Figure 4.

RBF CLASSIFICATION METHOD

	1990		2000		2010	
Forest	CCI	MAE	CCI	MAE	CCI	MAE
Population	83.95	9E-04	96.29	2E-04	74.07	0.002
crop	79.01	0.002	97.28	2E-04	80.24	0.002
Built up	78.02	0.002	41.48	0.006	45.18	0.005
Industry	94.32	0.012	84.44	0.01	78.76	0.011
Road	69.87	0.002	92.83	4E-04	78.51	0.001



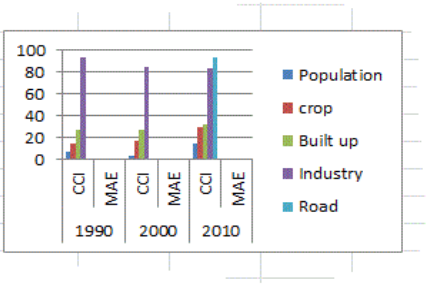
3.3 Bayesian Methods

Bayesian methods are one of the solution for the classification methods in data mining. In our work, Naive Bayesis implemented for classification. It follows an independent feature model with strong independence assumptions. This method is applicable for the statistical data.

Classification is done by appropriately to the attributes Cropand built up. It classifies and shows the better performance to 10% for all the years. This method shows good results for two attributes and is displayed in Table 5 and Figure 5.

NAIVES BAYES CLASSIFICATION METHOD

	1990		2000		2010	
Forest	CCI	MAE	CCI	MAE	CCI	MAE
Population	7.1	0.005	2.71	0.005	13.82	0.006
crop	14.32	0.007	16.54	0.007	29.13	0.006
Built up	26.66	0.006	27.4	0.006	32.59	0.005
Industry	93.82	0.021	84.69	0.012	83.7	0.006
Road	0.49	0.005	0.988	0.005	93.08	0.005



3.4 Rule Based Classification

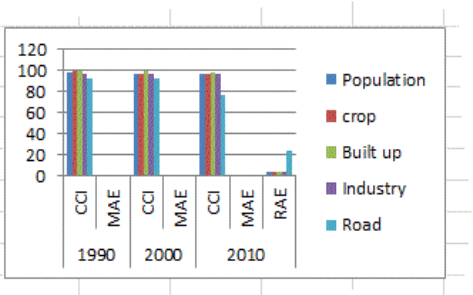
Decision trees can be translated into a set of rules by creating a separate rule for each path from the root to a leaf in the tree. However, rules can also be directly induced from training data using a variety of rule-based algorithms. Classification accuracy of rule learning algorithms can be improved by combining features using the background knowledge of the user automatic feature construction algorithms.

3.4.1 1R Algorithm

One Rule is a simple accurate, classification algorithm that generates one rule for each predictor in the data and then selects the rule with the smallest total error as its "one rule". It is the simplest rule-based classification learning algorithm for discrete attributes. It shows a gradual increase in the years 1990 and 2000 for all the attributes but suddenly no changes for the year 2010 because OneR produces rules only slightly less accurate .The result of this method is shown in Table 6 and Figure 6.

ONER CLASSIFICATION METHOD

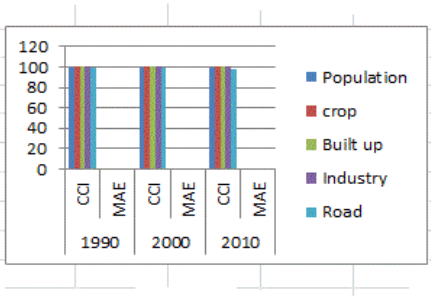
	1990		2000		2010		
Forests	CCI	MAE	CCI	MAE	CCI	MAE	RAE
Population	98.02	0.004	96.29	0	96.29	0	3.73
crop	99.75	0	96.79	0	96.79	0	3.35
Built up	100	0	99.25	0	98.51	0	3.15
Industry	97.28	0	96.79	0	96.79	0	3.2
Road	92.96	0	92.83	0	77.03	0.001	23.02



3.4.2 Prism

Prism is a greedy algorithm that finds a minimum spanning tree for a connected weighted undirected graph. This means it finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized. It can also be used to find the minimum spanning forest and reflects the same result for all the three decades , is in Table 7 and Figure 7.

PRISM CLASSIFICATION METHOD						
	1990		2000		2010	
Forest	CCI	MAE	CCI	MAE	CCI	MAE
Population	99.5	0	99.5	0	99.5	0
crop	100	0	100	0	100	0
Built up	100	0	100	0	100	0
Industry	100	0	100	0	100	0
Road	99.25	0	99.75	0	97.53	0



4 COMPARATIVE RESULTS

The Machine learning techniques such as Naïve Bayes, Bayes network, J48, Random Forest, Multi Layer Perceptron (MLP) and Radial Base Function (RBF) ,one R, PRISM were used for simulation. Here we split our original dataset of 405 samples into 66% for training purpose and remaining 34% for testing purpose. Weka incorporates k-fold cross-validation, in which the original sample is randomly partitioned into k subsamples. The cross-validation process is then repeated into several subsamples .Here ,we have used 10-fold cross validation .Kappa gives a numerical rating of -1 to 1 scale, where 1 is perfect agreement, 0 implies expected by chance, and negative values indicate agreement lesser than chance, Comparatively PRISM gives best result and shows the degradation of forest. The correctly classified Instances of various years are shown in Table8,9,10and the accuracy rates are displayed in Figure 8,9,10.

Comparative study -1990							
Forest	J48	RFOREST	MLP	RBF	NAIVES	ONER	PRISM
Population	83.95	99.00	0.228	83.95	7.1	98.02	99.5
crop	79.01	99.75	0.292	79.01	14.32	99.75	100
Built up	78.02	99.25	0.154	78.02	26.66	100	100
Industry	94.32	99.75	0.139	94.32	93.82	97.28	100
Road	69.87	99.25	0.154	69.87	0.49	92.96	99.25

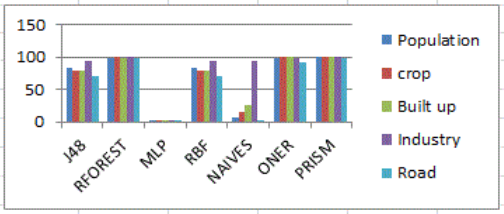


Fig. 8-1990

Comparative study-2000							
Forests	J48	RFOREST	MLP	RBF	NAIVES	ONER	PRISM
Population	96.29	99.01	0.314	96.29	2.71	96.29	99.5
crop	97.28	99.50	0.255	97.28	16.54	96.79	100
Built up	41.48	99.25	0.379	41.48	27.4	99.25	100
Industry	84.44	98.51	0.175	84.44	84.69	96.79	100
Road	92.83	99.50	0.154	92.83	0.988	92.83	99.75

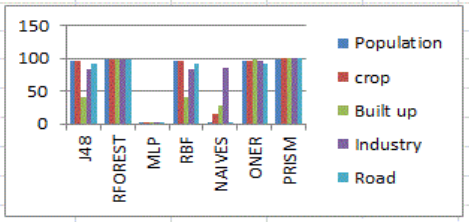


Fig.9-2000

Comparative study -2010							
Forests	J48	RFOREST	MLP	RBF	NAIVES	ONER	PRISM
Population	74.07	99.50	0.238	74.07	13.82	96.29	99.5
crop	80.24	100	0.122	80.24	29.13	96.79	100
Built up	45.18	99.75	0.393	45.18	32.59	98.51	100
Industry	78.76	98.51	0.062	78.76	83.7	96.79	100
Road	78.51	97.03	0.131	78.51	93.08	77.03	97.53

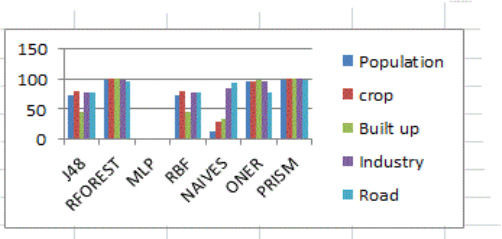


Fig. 10-2010

5. CONCLUSIONS AND FUTURE WORK

In this paper, four classifiers including Neural Network, Naïve Bayesian, Rule based, Decision tree were tested to determine the deforestation from the dataset of demographic factors. All the result were classified as 1,-1 or 0. There are many different mining and classification algorithms, and parameter settings in each algorithm. Experimental results in this paper are based on the default settings. Extensive experiments with different settings are applicable in WEKA. J48 is very simple classifier to make a decision tree, but it gives the invariable result in the experiment. Naïve Bayesian

classifier also showed good result only for two attributes(Crop and Built up), but RBF classified another two attributes(Industry and Road) properly. Rule Based classifiers such as OneR and PRISM also showed good result than compared with J48 or Naïves Bayesian classifier. From this experiment, we can find that a simple Random Forest classifier can provide best classification result for deforestation(except one attribute).It is planned to incorporate other techniques like different ways of feature selection, classification using ontology.

REFERENCES

[1] PeimanMamaniBarnaghi+,VahidAlizadehSahzabi and Azuraliza Abu Bakar "A Comparative Study for Various Methods of Classification", 2012 International Conference on Information and Computer Networks (ICICN 2012).

[2] LinlinXu, Jonathan Li, Alexander Brenning"A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery",Remote Sensing of Environment 141 (2014) 14–23.

[3] .Elaine Astrand1, Pierre Enel2, Guilhem Ibos1, Peter Ford Dominey2, Pierre Baraduc1,Suliann Ben Hamed1* "Comparison of Classifiers for Decoding Sensory and Cognitive Information from Prefrontal Neuronal Populations ",PLOS ONE | www.plosone.org.

[4] S. B. Kotsiantis "Supervised Machine Learning: A Review of Classification Techniques"Informatica31 (2007) 249-268 249.

[5] Archana Chaudhary1, Savita Kolhe2 & Raj Kamal3"Machine Learning Classification Techniques: A Comparative Study ",International Journal on Advanced Computer Theory and Engineering (IJACTE) ISSN ISSN (Print) : 2319 – 2526, Volume-2, Issue-4, 2013.

[6] Haowen You1 and George Rumbe2 "Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data" International Journal of ArtificialIntelligence and Interactive Multimedia, Vol. 1, Nº 3.

[7] PeimanMamaniBarnaghi+,VahidAlizadehSahzabi and Azuraliza Abu Bakar "A Comparative Study for Various Methods of Classification" 2012 International Conference on Information and Computer Networks (ICICN 2012) IPCSIT vol. 27 (2012) © (2012) IACSIT Press, Singapore

[8] Gaya Buddhinathand Damien Derry "A Simple Enhancement to One Rule Classification".

[9] SeongwookYoun and Dennis McLeod "A Comparative Study for Email Classification"

[10] K.Saritha1, S. Jyothi2 & K. R. Manjula3 "Class association rule mining for analyzing deforestation factors" International Journal of Civil, Structural,Environmental and Infrastructure Engineering Research and Development (IJCSEIERD) ISSN(P): 2249-6866; ISSN(E): 2249-7978 Vol. 3, Issue 5, Dec 2013, 237-248

[11] K.R.Manjula ,Prof. S. Jyothi, S.Anand Kumar Varma "Analysing the factors of deforestation using GIS "www.geospatialworld.net/paper/application/ArticleView.aspx?aid=30273