ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

Survey On Classification Based Techniques On Non-Spatial Data

Ashish Kokare[1], Pradeep Venkatesan[2], Swapnil Tandel[3], Hemant Palivela[4]

B.E Student, Department of Computer Engineering, A.C.Patil College of Engineering, Navi Mumbai, India[1, 2, 3]

Assistant Professor, Department of Computer Engineering, A.C.Patil College of Engineering, Navi Mumbai, India[4]

Abstract— The paper marks a method that speaks on soil conservation in various areas of the country that is dry and also makes a note on the seed lots that can be used in such areas. We base our work basically on the decision tree based classification to work on categorical an numerical values and also choose on particular algorithm for application for precision agriculture for a particular data set on the basis of performance analysis and experimental results that we would obtain.

Keywords-precision agriculture, decision trees, accuracy parameters , plot

I. INTRODUCTION

Precision agriculture is a field which is reaching its pinnacle of growth where the requirement of knowledge regarding soil nutrient distribution has become indispensable for a fruitful crop yield. Farmers fall victim to inappropriate information about when and where to crop is becoming an issue which is raising concern across wide variety. of nations This situation gives this paper a chance to up forth an idea that may curb the issues related to poor farming by use of Decision Tree based classification of the collected dataset. This dataset would be collect by sensors which may be a single machine mobile enough to continuously sense the soil nutrient and create a data set with the appropriate attributes or the sensors may be deployed as a wireless sensor network. Based on this data set we now classify the tuple generated on various decision tree based classification algorithms and choose a best candidate out of these which insinuate accurate classification of that particular soil composition. The main idea behind use of only decision tree based classification was bolstered many factors. Since we are working with categorical and numerical values with a various parameters associated with each sample of data. So, decision tree stands as the best candidates to classify such kind of data as decision tree themselves are non-parametric and can be generated with ease. The technical aspects such as efficient breadth first and depth first traversal search techniques espouse decision tree based classification as the most suitable techniques of classification. So, the crux of the paper presents an idea of collection of dataset and application of various decision tree based classification algorithms to choose the best candidate for classify the dataset for future applications of precision agriculture The organization of the paper is as follows. Section 1 gives the main idea of the paper and explains the existing scenario and future prospects in a nutshell. Section 2 gives the pre-requisite and related work regarding the whole process.Section3 shows the performance analysis of each algorithm based on some accuracy based parameters and also a graph plot between the former and the latter .Section4 provides a conclusion and future scope

II. RELATED WORK

A. *J48*

J48 is java implemented version of C4.5. C4.5 is an advanced version of Ross Quinlan's ID3 algorithm.C4.5 is generally referred to as statistical classifier since the decision trees generated by this algorithm is used for classification purpose. It performs very much similarly to ID3 except using the gain ratio to determine the best target attribute. C4.5 algorithm also makes some advancements to ID3, which has the ability to handle numerical attributes by creating a threshold and splitting the data into those whose attribute value is above the threshold and those that are less than or equal to it. Also it has the ability to handle attributes with differing cost.C4.5 can also prune the decision tree after creation, which reduces the size of the tree and hence saves the memory.

B. *J48 Graft*

Decision tree grafting adds nodes to an existing decision tree with the prime aim of reducing the prediction tree. Hence, j48graft algorithm considers a set of training data for each leaf of decision tree. Here grafting is applied as a post process to an already generated decision tree. It identifies the regions of the sample space that are not occupied by training examples, and considers optional classification for those identified regions. These classifications are generated by considering alternative branches based on the predecessor nodes to the leaf contained those identified region. If an alternative classification than that assigned to the region

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

by the current tree, a new branch is grafted to that tree that replaces the old classification by the newly generated one. Complexity of this technique is comparatively lower than that of the single tree from a committee although it increases the size of the tree.

C. LAD TREE

LAD Tree addresses the alternative method to obtain regression trees by making use of LAD (Least absolute Deviation) criterion. It provides a powerful method for skewed distribution and outliers than the LS criterion previously used in standard regression trees. The actual variation in LAD tree from LS tree is the use of mean absolute deviation as error criterion instead of averages in the leaves. As a consequence of this theorem LAD trees should have medians at the leaves instead of averages like LS regression trees. The median is a much better numeric indicator of centrality than the average as far as skewed distributions are discussed.

D. NB Tree

NB tree uses the classical approach of recursive partitioning schemes. The nodes of the tree are such that each poses a functionality of presenting e each node as a naïve bayes classifiers .Entropy provides the homogeneity information of the attributes and hence, NB tree uses entropy minimization technique to set a threshold for the continuous attributes. Discretizing and cross validation techniques are used to get an estimate of how well the naive Bayesian node of the tree.

E. Random Forest

Random forest algorithm developed by Leo Breiman and Adele Cutle is an ensemble learning method for classification and regression that operate by constructing a assembly of decision trees at training time and thereby outputting the class. Random forests are basically a combination of tree predictors such that a randomly sampled vector decides the structure of the tree. Random forest is a classifier consisting of tree structured classifiers which depends upon these randomly sampled vectors. The random selection of feature to split each node produces error rates that compare favorable to Adaptive boosting, however are more powerful than with respect to noise. Random forests are used to rank the importance of variables in a classification or regression problem.

F. REP Tree

Constructing tree models was under serious criticism that a single tree or a nested sequence of trees that is generated ignores the uncertainty that gets associated with the tree structure.This problem is not as bad as it seems as hundreds of trees differ only at few nodes.So we propose a method wherein we define several distance metrics thereby summarizing a forest of trees by several representative trees and associated clusters.The added tree plot, a peculiar plot,is introduced as a means to deicde how many trees to analyse exactly.At the same time It also assist in the identifying different kinds of trees that best fit the dataset.

G. Simple Cart

Simple cart algorithm is an inference of the classification and regression tree based technique which emphasizes on the non parametric decision tree based learning which bases its classification on whether the dependent variable is categorical or numeric. This algorithm also develops decision trees based on the splitting factor classifying the tuple into their respective classes based on the class label attributes.

H. Random Tree

It is based on evolution of a random value which is called as the stochastic process. It works on structures which logically form a proper arborescence. It is encapsulates a set of tree under one classification .The types of random trees are random binary tree Random minimal spanning tree, Random recursive tree, Rapidly exploring random tree, Brownian tree, Random forest and Branching process.

I. BF Tree

It is based on evolution of a random value which is called as the stochastic process. It works on structures which logically form a proper arborescence. It is encapsulates a set of tree under one classification .The types of random trees are random binary tree Random minimal spanning tree, Random recursive tree, Rapidly exploring random tree, Brownian tree, Random forest and Branching process [9, 10].

J. Decision Stump

Decision stump is basically a one level decision tree. It is called the decision "stump" as it resembles the bottom tree portion projecting from the ground after major portion of the tree has been cut down.). It bases a decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rules algorithm. . A set of attributes were worked upon with using weka to classify the tuple with the related class labels with the following parameters.

K. Fault Tree

Fault tree is deductive failure analysis which uses Boolean logic to handle lower level events. It is proposed

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

to be widely used in the safety engineering domain which incorporates reliability engineering as an integral component. It plays an instrumental role in the process of depicting compliance with the system safety .It in highly important and of utmost importance in controlling and monitoring the safety of a complex system. A set of attributes were worked upon with using weka to classify the tuple with the related class labels with the following parameters.

III.EXPERIMENTAL RESULTS

The Decision tree process based classification done on a dataset built on weka tools was accompanied by a set of parameters such as size of the tree number of leaves generated ,time to built each tree .So a plot showing each of these parameters versus the various algorithms are shown as follows.

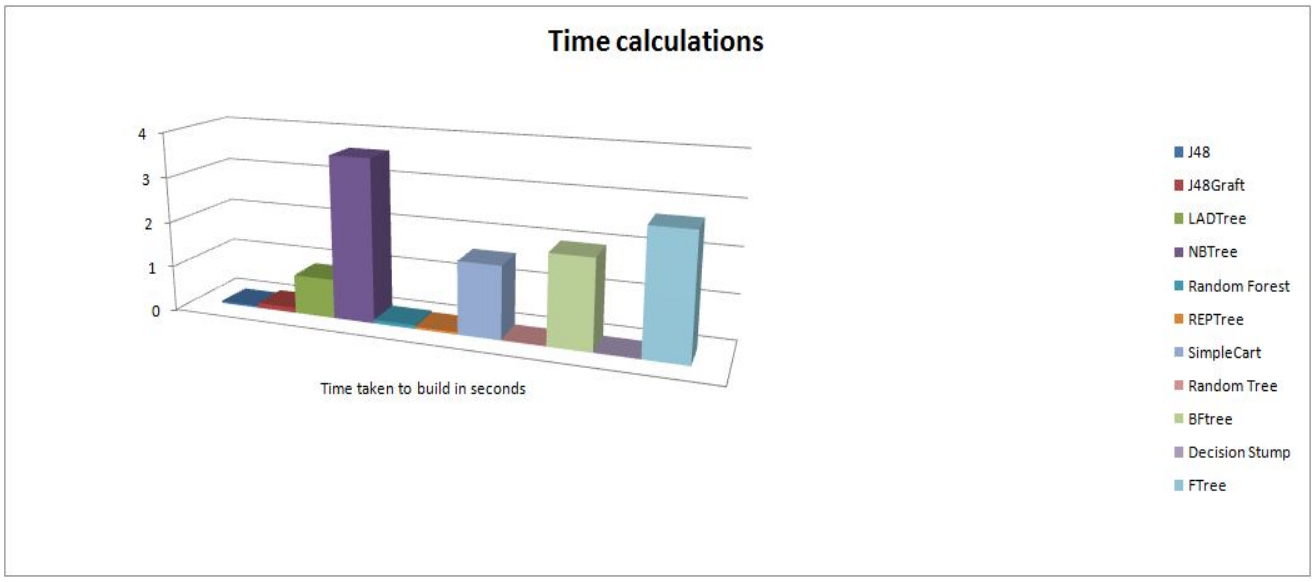


Figure 1: Time Calculations

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

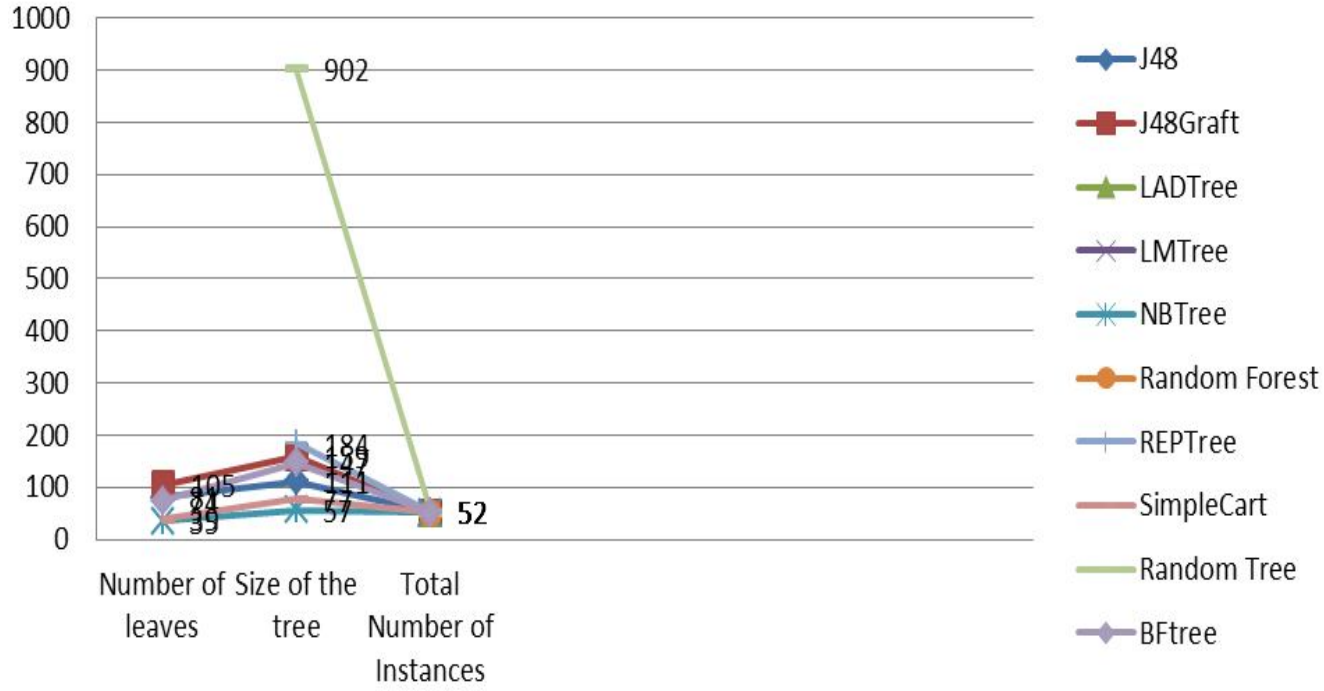


Figure 2:Storage calculations

IV.PERFORMANCE ANALYSIS

The following plot demonstrates the performance analysis of all the Decision tree algorithms under consideration with respect to parameters defining the accuracy level of each algorithm.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

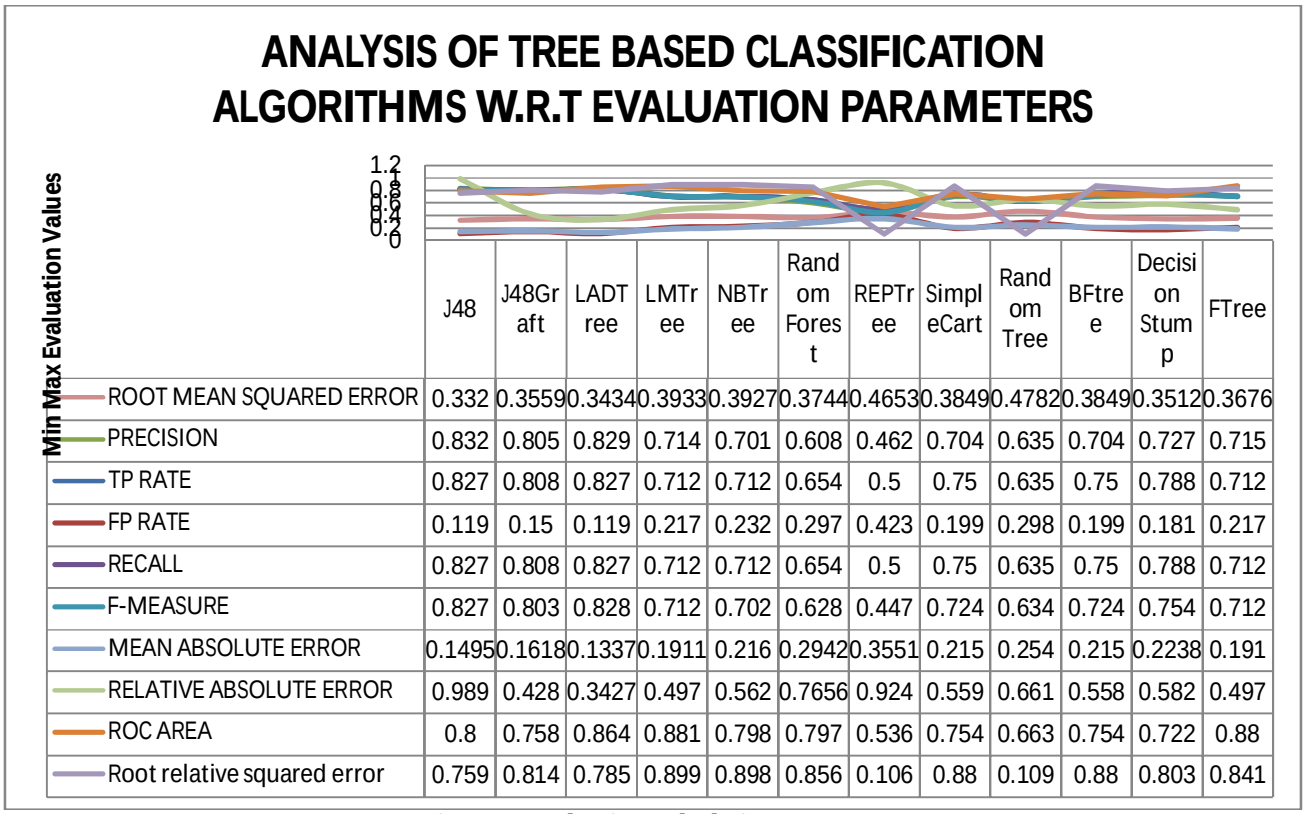
International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India



	J48	J48Graft	LADTree	LMTree	NBTree	RandomForest	REPTree	SimpleCart	RandomTree	BFtree	DecisionStump	FTree
ROOT MEAN SQUARED ERROR	0.332	0.3559	0.3434	0.3933	0.3927	0.3744	0.4653	0.3849	0.4782	0.3849	0.3512	0.3676
PRECISION	0.832	0.805	0.829	0.714	0.701	0.608	0.462	0.704	0.635	0.704	0.727	0.715
TP RATE	0.827	0.808	0.827	0.712	0.712	0.654	0.5	0.75	0.635	0.75	0.788	0.712
FP RATE	0.119	0.15	0.119	0.217	0.232	0.297	0.423	0.199	0.298	0.199	0.181	0.217
RECALL	0.827	0.808	0.827	0.712	0.712	0.654	0.5	0.75	0.635	0.75	0.788	0.712
F-MEASURE	0.827	0.803	0.828	0.712	0.702	0.628	0.447	0.724	0.634	0.724	0.754	0.712
MEAN ABSOLUTE ERROR	0.1495	0.1618	0.1337	0.1911	0.216	0.2942	0.3551	0.215	0.254	0.215	0.2238	0.191
RELATIVE ABSOLUTE ERROR	0.989	0.428	0.3427	0.497	0.562	0.7656	0.924	0.559	0.661	0.558	0.582	0.497
ROC AREA	0.8	0.758	0.864	0.881	0.798	0.797	0.536	0.754	0.663	0.754	0.722	0.88
Root relative squared error	0.759	0.814	0.785	0.899	0.898	0.856	0.106	0.88	0.109	0.88	0.803	0.841

Figure 3: Evaluation Calculations

ACKNOWLEDGMENT

We wish to thank department of computer science and engineering, and our beloved Head of the department, Prof. Nitin P Chawande of A.C.Patil College of Engineering for providing us with the right path to our research work. We very joyfully express our deep sense of gratitude towards Mr. Hemant Palivela's for supporting us and providing us stepping stones to accomplish our work perfectly.

REFERENCES

[1] Palivela, Hemant, and Pushpavathi Thotadara. "Computing Communication & Networking Technologies (ICCCNT)." 2012 Third International Conference on computing communication and network technologies, 26th-27th July. 2012.

[2] Hahsler, Michael, and Sudheer Chelluboina. "Visualizing Association Rules: Introduction to the R-extension Package arulesViz." R project module (2011).

[3] Raymer, Michael L., William F. Punch, Erik D. Goodman, and Leslie A. Kuhn. "Genetic programming for improved data mining: application to the biochemistry of protein interactions." In Proceedings of the first annual conference on genetic programming, pp. 375-380. MIT Press, 1996.

[4] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.

[5] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.

[6] Ketterlin, Alain, Pierre Gançarski, and Jerzy J. Korczak. "Conceptual Clustering in Structured Databases: A Practical Approach." In KDD, pp. 180-185

[7] Quinlan, John Ross. C4. 5: programs for machine learning. Vol. 1. Morgan kaufmann,1993.

[8] Rodriguez, Juan José, Ludmila I. Kuncheva, and Carlos J. Alonso. "Rotation forest: A new classifier ensemble method." Pattern Analysis and Machine Intelligence, IEEE Transactions on 28.10 (2006): 1619-1630.

[9] Palivela, H., Yogish, H. K., Shalini, N., & Raghavendra, S. N. (2014). Novel Approach for Finding Patterns in Product-Based Enhancement Using Labeling Technique. In Intelligent Computing, Networking, and Informatics (pp. 1249-1256). Springer India.

[10] Palivela, Hemant, H. K. Yogish, S. Vijaykumar, and Kalpana Patil. "A study of mining algorithms for finding accurate results and marking irregularities in software fault prediction." In Information Communication and Embedded Systems (ICICES), 2013 International Conference on, pp. 524-530. IEEE, 2013.