Using Data Mining Techniques for Predicting Individual Tree Mortality in Tropical Rain Forest : with Optimized Parameters in Logistic Regression and Decision Trees Approach

Mr.Shamsundar K Bairagi[1]

[1] *PG Student, Department of Computer Engineering, SND College of Engineering and Research Centre, Yeola, Nashik, Maharashtra, India*

ABSTRACT

Forest ecosystem dynamics uses tree mortality. It is one of the most common facts due to species rich in tropical rain forests. Individual tree mortality model was developed for predicting the probability of mortality in dipterocarpaceae tree family group in Koh Kong province, Cambodia. It is a big challenge of finding appropriate methods for modeling mortality. There two data mining methods here; Logistic regression and decision trees. To chose for decision trees method, Chi squared Automatic Interaction Detector method was selected which always chooses independent variables. With increasing individual tree basal area, the probability of mortality gets decreased. Using calibration and discrimination the performance was compared of each model. The study presented that logistic regression outperformed decision trees for both calibration and discrimination. To improve the accuracy of the stand forecast the model developed.

.

Keywords: *Logistic regression, decision trees, tree mortality, model performance.*

1. INTRODUCTION

To find unsuspected relationship and to summarize the data in models and patterns that are both understandable and useful to the data owner it is analysis of data observation. A typical data mining process includes data acquisition, data integration, data exploration, model building, and model validation. On using logistic regression and decision trees in data mining techniques to predict individual tree mortality. The dynamics and structure of forest tree populations and the succession and composition of forest communities; Mortality is one of the key factors that influenced. Mortality has been studied extensively in tropical forests, especially in modeling individual tree mortality. To describe and modeling tree mortality at different scales, A variety of empirical approaches have been used. Stand level mortality models predict stand density changes over time and they often estimate mortality in terms of volume, basal area or number of trees per hectare, while individual tree mortality models predict the probability of survival or death for each individual tree involved in growth projections.

Stand averages in terms of mortality will be less useful because of the high degree of variability, but the scale selection also depends on the other components of the growth model. As individual tree models are often preferred from a management perspective, because they provide a more detailed description of the stand structure and its dynamics. Tree level models enable a more detailed description of the stand structure and its dynamic than stand level models, when individual tree information for a stand is available.

Logistic function, generalized logistic function, two step approaches and three step approaches and neural network have been used several statically methods. To estimate mortality rate for tropical forest mathematical formulation was used. Very few studies had been done for predicting tree mortality using decision trees method.

Logistic regression appears to be the best method for individual tree mortality modeling and has been widely applied. Vanclay believes the logistic function fitted to individual-tree data may offer the best way to model mortality in tropical forest which cover a spectrum of different species mixtures and age structures, precluding the possibility of using either stand age or site index as predictor variables.

The greatest method to predict tree mortality is logistic regression. We decided to use decision tress as another alternative method to compare with logistic regression. The aim of this paper is to model individual tree mortality for dipterocarpaceae tree group in Koh Kong province, Cambodia by using two of data mining techniques: logistic regression and decision trees. Regular mortality was modeled here. The result should be to contribute the effective and efficient methods to predict individual tree mortality for this tree group in any tropical rain forest.

2. LITERATURE SURVEY

There are lists of data mining techniques such as frequent pattern mining, classification, regression, clustering, association rule mining and many more, but out of these classification is frequently used at the most. In classification, the model is trained which describes and differentiate different data classes to predict the classes whose labels are not known. The classification can be performed with different algorithms such as neural networks, decision trees, regression, etc. Due to the significance important of decision tree for large data sets, in this research decision tree approach has been used. Generally the classification is the sequence of the operations as follows:

1. Prepare the training data set using pre-processing on the raw data.
2. Class attribute and classes are identified.
3. Identify useful attributes for classification.
4. Learn a model using training examples in Training set. 5. Use the model to classify the unknown data samples.

3. PROBLEM STATEMENT

Tropical rain forests are known of comprising of a huge number of species; however, most of them are represented by only a few trees. It is almost impossible to develop mortality model for each species. Mortality in tropical rain forest having major impact of ecosystem. In this paper we are going to minimize the parameter to find out mortality of tree using decision tree & logistic regression.

4. PROPOSED SYSTEM AND ARCHITECTURE

Thus the problem clearly shows that there is need to optimize parameter of trees to find mortality in tropical rain forest. Initially some parameter are not having influence into actual result but still existing system taking that parameter into consideration. So we can optimized that parameter are cut out efforts, save time, also performance improvement.

Objectives of the proposed method

1. Optimized parameters using logistic regression and decision tree.
2. Save time to take unnecessary parameter readings.
3. Performance improvement.

Data and Method

4.1 Data Base

The twenty permanent sample plots in koh kong province were used from data for all dipterocarpaceae. Koh Kong province is located in western part of the country, between 12011'-13026'N and 104012'-1050 44E, and has a total land area of 12,116km2, about 6.67% of the country area (Forest Administration, 2006). 50 by 50 meters of plots were plot (quarter hectare) in which all trees with diameter at breast height (DBH) of 30 cm and greater were numbered and measured for diameter in this area. While all trees with DBH 7.5 cm to less than 30 cm were measured for their diameters in the sub plots of 20 by 20 meter. The counting of saplings and seedlings were done in the sub plots of 5 by 5 meters and 2 by 2 meters respectively. Data were collected over 12 years from 1998 to 2010 with the interval of two to seven years. There are four measurements taken place which was in year 1998, 2000, 2007 and 2010. DBH was re-measured; mortality trees and recruitment trees were also recorded. Regular and irregular mortality can be separated from natural mortality. Regular mortality, or self-thinning, is due to competition for light, water and soil nutrients within a stand. Dead trees caused by human disturbance such as illegal cutting was not considered as natural mortality.

Tropical rain forests are known of comprising of a huge number of species; however, most of them are represented by only a few trees. It is almost impossible to develop mortality model for each species.

Using criteria which appropriately reflect the intended use of model being build; species need to be group. According to tree families data were divided into two major groups. However, only one group (dipterocarpaceae) was used for analysis. There were 29 families exist in this forest with dipterocarpaceae as the major family. Dipterocarpaceae is well-known trees of the Asian rain forest which consist over 500 species [6]. In this study, the dipterocarpaceae family consists of only four species which are *Hopea pierrei* Hance, *Vatica odorata* Griff Sym, *Shorea siamensis* Miq and *Shorea farinosa* CEC Fisch while other families were grouped in non dipterocarpaceae. The non dipterocarpaceae tree family in this data set consists of more than 30 different species.

4.2 Selection of Variables

Selection of appropriate predictor variables should not only be based on test statistics, but also on a basic understanding of how forest ecosystems function and how factors contributing to mortality are expressed.
As the age can't be calculated of every tree, site index and age were excluded intentionally. To characterize tree size individual diameter is used. The independent variables evaluated in this analysis were tree diameter at breast height (DBH), tree diameter squared (Dsq), basal area (BA) and total basal area of trees larger than subject tree (BAL). The present analysis focuses on a major tree family group of dipterocarpaceae in evergreen forest at Koh Kong province, Cambodia. Logistic regression was performed to generate model using training data set. Decision trees were then used to generate models based on the significant variables found in the logistic regression by using the same data set. To generate the logistic regression and decision trees models we used SPSS software.

4.3 Model Development and Validation

Discrete event, either alive or dead is a tree mortality. In this study, the dependent variable was coded as 0 for trees that were alive at both ends of a measurement interval, and 1 for trees that were alive at the beginning of a measurement but dead at the last measurement. Data for this group was split into two subsets; 70% for training data set and 30% for testing data set. As logistic regression be the best method and most widely used in modeling tree mortality [30, 19, 33, 32], in this study the following logistic model (survival model) was hypothesized:

$$PS = \left[1 + e^{-(\beta_0 + \beta_1 * DBH + \beta_2 * Dsq + \beta_3 * BAL + \beta_4 * BA + \beta_5 * DI)}\right]^{-1}$$

where DBH is individual tree diameter at breast height, Dsq is DBH square, BAL is basal area for all trees larger than subject trees, BA is the basal area, DI is the annual diameter increment, □ *s* are the coefficients to be estimated and *e* is the base of the natural logarithm.

Architecture

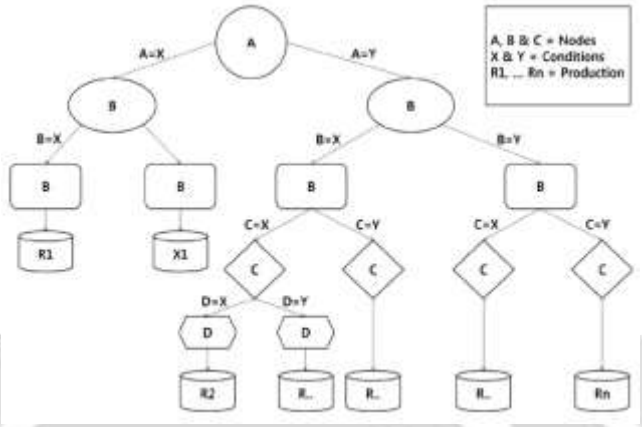


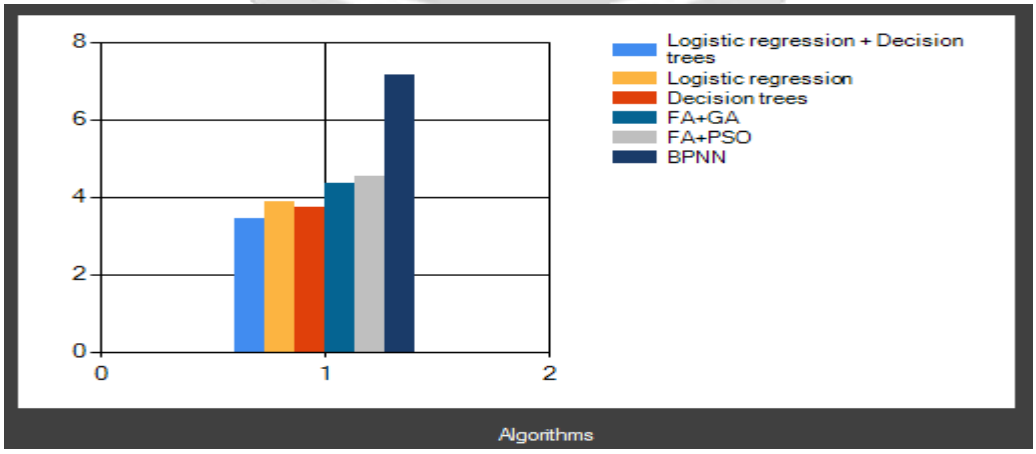
Fig. Architecture of decision tree model.

5. ALGORITHM USER FOR IMPLEMENTATION

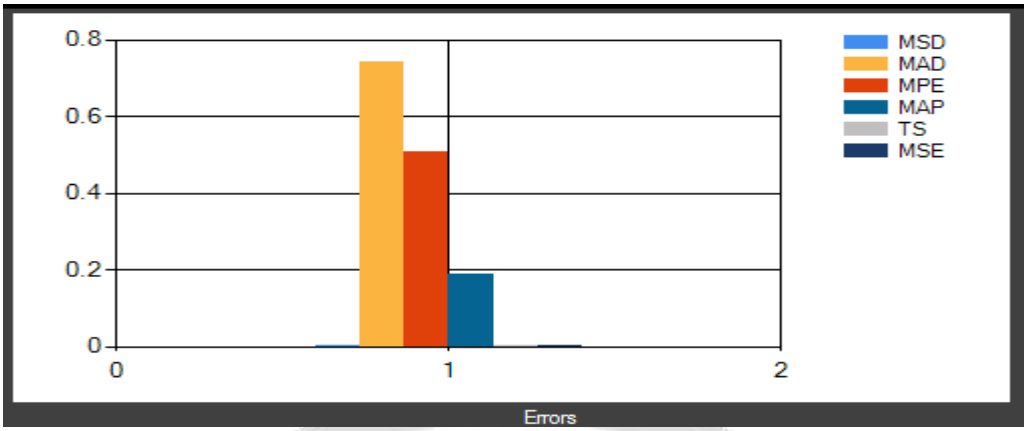
To generate decision tree on training data set we follow ID3 (Iterative Dichotomiser 3) decision algorithm. It having top down approach with greedy search through the given sets to test each attribute at every tree node. In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree. Thus, we can select the attribute with the highest information gain (entropy reduction in the level of maximum) as the test attribute of current node. In this way, the information needed to classify the training sample subset obtained from later on partitioning will be the smallest. That is to say, the use of this property to partition the sample set contained in current node will make the mixture degree of different types for all generated sample subsets reduce to a minimum. This algorithm does not produce more accurate outcomes if the noise is present or some attribute values are missing. To overcome this problem we deleted that rows which having missing or null values.

6. RESULT

While taking reading we consider following parameters Basal Area (BA), spcode, species, cover, utme, utmn, elev, tci, streamdist, disturb, beers etc. After applying algorithm & all process some parameters are eliminated.
Result is show graph



Error graph



7. CONCLUSION

Logistic regression using training data set produced only one significant variable which was basal area (BA) with $P < 0.05$. Most of the tested predictors were rejected from the final model. The probability of survival for dipterocarpaceae group increased when basal area increased. In other word, the predicted probability of individual-tree mortality of dipterocarpaceae decreased with increasing tree basal area. It is well established that small plants have higher probabilities of dying than large plants. When trees are very small, they are exposed to various mortality agents such as severe weather conditions and competing vegetation. Mortality rate at this stage are high and the rates start to decrease with increasing tree size.

8. REFERENCES

[1] Adame P, Rio M D, Canellas I. 2010. Modeling individual tree mortality in Pyrenean oak (*Quercus pyrenaica* Willd.) stands. *Ann. For. Sci. 67 810.*

[2] Álvarez González J.G., Castedo Dorado F., Ruíz González A.D., López Sánchez C.A., and Von Gadow K. 2004. A two-step mortality model for even-aged stands of *Pinus radiata* D. Don in Galicia (Northwestern Spain). *Ann. For. Sci. 61: 439–448.*

[3] Badriyah T, Briggs J S and Prytherch D R. 2012. Decision Trees for Predicting Risk of Mortality using Routinely Collected Data. *World Academy of Science, Engineering and Technology 62 2012.*

[4] Chen H. Y. H., Fu S., Monserud R. A. and Gillies I. C. 2008. Relative size and stand age determine Pinus banksiana mortality. Forest Ecology and Management 255, 3980 – 3984.

[5] Condit R, Hubbell SP & Foster RB. 1994. Density dependence in two understorey tree species in a neotropical forest. *Ecology 75: 671–680.*

[6] Dayanandan S, Peter S. Ashton, Scott M. Williams, and Richard B. Primack. 1999. Phylogeny of the Tropical Tree Family Dipterocarpaceae Based On Nucleotide Sequences Of The Chloroplast Rbcl GenE. American Journal of Botany 86(8): 1182–1190.

[7] Diéguez-Aranda U., Castedo Dorado F., Álvarez González J.G., and Rodríguez-Soalleiro R. 2005. Modelling mortality of Scots pine (Pinus sylvestris L.) plantations in the northwest of Spain. Eur. J. Forest. Res. 124: 143–153.

[8] Eid Tron and Erik Tuhus. 2001. Models for individual tree mortality in Norway. Forest Ecology and Management, Volume 154, Issues 1–2, Pages 69–84

[9] Forest Administration (FA). 2006. Cambodia: Forestry Statistics 2005.

[10]Fridman J. and Stahl G. 2001. A three-step approach for modelling tree mortality in Swedish Forests. Scand. J. For. Res. 16: 455–466.

Mr. Bairagi Shamsundar Keshavdas Completed BE in Information Technology in 2012 and pursuing ME in Computer Engg. From SND COE & RC Yeola, (MH) India