J. Ind. Soc. Agril. Statist.
61(2), 2007 : 192-198

Prediction of Forest Cover using Decision Trees

B. Chandra and Pallath Paul V.[1]
Indian Institute of Technology, New Delhi

SUMMARY

Information regarding forest land is highly required for developing ecosystem management strategies which will facilitate the decision-making process. It is often difficult to get the relevant data for forest land that are outside the immediate jurisdiction of the concerned authorities. One of the approaches for obtaining this information is through the use of predictive models like Decision Trees and Neural Networks. (Blackard *et al.* 2000) have shown that Neural Network approach outperforms the traditional discriminant analysis method in predicting forest cover types. The accuracy achieved by Neural Network was 70.58%. Decision Trees algorithms have been proposed in the past for classification of numeric as well as categorical attributes. SLIQ algorithm was proposed (Mehta *et al.* 1996) as an improvement over ID3 and C4.5 algorithms (Quinlan 1993). Robust algorithm for Decision Tree Classification was proposed (Chandra *et al.* 2006) as improvement over SLIQ where the Decision Tree is built by examining reduced number of split points and maintaining the same classification accuracy. Prediction of forest cover types using Decision Trees is discussed in this paper. Maximum accuracy of about 84% is achieved using Decision Trees.

Key words : Classification, Prediction, Information gain, Gain Ratio, Gini Index.

1. INTRODUCTION

Several decision tree algorithms have been developed for classification. ID3 algorithm (Quinlan 1981) for classification uses information gain as a measure to select the best splitting attribute. The attribute with the highest information gain is selected as the splitting attribute. One of the main drawbacks of ID3 is that the measure Gain used tends to favor attributes with a large number of distinct values. This drawback was overcome to some extent in C4.5 (Quinlan 1993) by introducing a new measure called Gain Ratio.

SLIQ (Metha *et al.* 1996) is a decision tree classifier developed by the Quest team to handle both numeric and categorical attributes. SPRINT (Shaefer *et al.* 1996) was also developed by the Quest team that basically aims at parallelizing SLIQ. In SLIQ algorithm, while evaluating the best split for each numeric attribute having n values, the Gini index has number of split points to be evaluated at a node with m attributes is m*(n-1), m being the number of attributes. This makes SLIQ computationally complex for numeric attributes. The PUBLIC algorithm proposed by Rastogi *et al.* (1998) for tree generation is the same as SPRINT but Entropy is used as a measure for checking the goodness of split. Robust C4.5 algorithm (Zheng *et al.* 2005) is an improvement over C4.5. Elegant Decision Tree Algorithm (EDTA) developed by Chandra *et al.* (2002) was proposed as an improvement over SLIQ where the Gini index is computed not for every successive pair of values of an attribute but over different ranges of attribute values. This reduces the number of split points as well as the number of computations. It was shown that the classification accuracy was much better than SLIQ. In EDTA the number of split points evaluated is n / k (where n is the total number of different values the attribute can take and k is the total number of intervals or group size). In this paper an improvement over EDTA has been suggested to reduce the computational complexity. Robust Algorithm for Classification using Decision Trees (RDTA) developed by Chandra *et al.* (2006) is a further improvement over EDTA where Gini Index is evaluated for each attribute at displacements of sigma (standard

[1] *Institute of Technology, Nirma University, Ahmedabad*