International Journal of Multimedia Data Engineering and Management
Volume 9 • Issue 1 • January-March 2018

Machine Learning Classification of Tree Cover Type and Application to Forest Management

Duncan MacMichael, University of Washington, Bothell, USA
Dong Si, University of Washington, Bothell, USA

ABSTRACT

This article is driven by three goals. The first is to use machine learning to predict tree cover types, helping to address current challenges faced by U.S. forest management agencies. The second is to bring previous research in the area up-to-date, owing to a lack of development over time. The third is to improve on previous results with new data analysis, higher accuracy, and higher reliability. A Deep Neural Network (DNN) was constructed and compared with three baseline traditional machine learning models: Naïve Bayes, Decision Tree, and K-Nearest Neighbor (KNN). The neural network model achieved 91.55% accuracy while the best performing traditional classifier, K-Nearest Neighbor, managed 74.61%. In addition, the neural network model performed 20.97% better than the past neural networks, which illustrates both advances in machine learning algorithms, as well as improved accuracy high enough to apply practically to forest management issues. Using the techniques outlined in this article, agencies can cost-efficiently and quickly predict tree cover type and expedite natural resource inventorying.

KEYWORDS

Decision Tree, Deep Neural Network, Forest Management, K-Nearest Neighbor, Machine Learning, MATLAB, Naïve Bayes, Tree Cover Type

INTRODUCTION

Outline

This paper begins with an introduction to the problem background and research background, followed by a description of the dataset. Next, the methodology and approach for analyzing and preprocessing the dataset are presented. Experiments that were performed are then detailed, including techniques to improve accuracy and the method used for ensemble training in parallel. Finally, results are described for the baseline classifiers and the neural network. These results are compared to past work in this area before listing challenges, future work, and conclusions.

Background and Motivation

Classifying areas of forest based on the predominant tree type can be of great use to land management agencies for natural resource inventory information, especially for federally protected and managed national forests in the United States (Blackard & Dean, 1999). This is particularly true of older-growth forests in remote areas where it may be arduous to manually measure all types of trees (White & Lloyd,

DOI: 10.4018/IJMDEM.2018010101

1994). High transportation costs, difficult terrain, and concerns of interference with natural ecosystems present many obstacles for accurate and up-to-date data gathering and analysis in such areas.

In addition to difficulty in obtaining remote forest data, several other challenges within the domain of forest conservation and management have arisen within the past ten years. Conserving forest genetic resources is one example, where a shrinking land base due to increased demand for wood and other tree-based products has strained the health and genetic diversity of the remaining forest. Efforts are in place to evaluate molecular markers and adaptive traits of the trees to inform conservation efforts, but these efforts are limited by time and resources available (Rajora & Mosseler, 2001).

Another challenge for land management agencies is working with private landowners in determining appropriate forest practices (Alavalapati, Matta, & Tanner, 2005). Properly maintaining and investing in land by following practices like maintaining un-even aged stands of trees or conducting appropriate thinning to provide open canopy can be an expensive task, one which requires constant attention and time to detail what trees grow where (Alavalapati, Matta, & Tanner, 2005).

Finally, a third major challenge is managing forests that are near human development (Butler et al., 2011). Certain species of trees, such as American beech and red pine, are more susceptible than others to human impact and thus need to be monitored more carefully (Butler et al., 2011). Additionally, humans have the capacity to introduce invasive nonnative flora and fauna to forest ecosystems, which may impact different tree types in different ways and necessitate detailed data of tree cover types (Butler et al., 2011).

In all these challenges, whether they arise from the remoteness of the forest, lack of time, lack of manpower, lack of funds, or complexity of human impact, lies one common thread: a position to benefit from faster, easier, and cost-effective prediction of tree cover types. Solving this problem is the first driving factor in this paper's research.

The second driving factor is the age of previous work in the area and lack of development since. The first pioneering paper demonstrating neural networks and machine learning applied to tree cover type classification was by Blackard and Dean in 1999, 18 years before the time of writing this paper (Blackard & Dean, 1999). In that paper, an artificial neural network (ANN) is trained on the Forest Cover Type dataset (covtype) and compared with a traditional discriminant analysis classifier to classify tree cover type as accurately as possible. This is done on data gathered from cartographic measurements, with no remote sensed data (Blackard & Dean, 1999).

Unfortunately, since 1999, the tree cover type (covtype) dataset has not been used for further development of this work. The direction and conclusions of recent papers that do reference the covtype dataset concentrate on the algorithms and techniques presented, rather than machine learning applications of covtype specifically (Guo et al., 2015; Arabmakki, Kantardzic, & Singh Sethi, 2016; Chenhan, March, Xiao, & Biros, 2016; Hsieh, Si, & Dhillon, 2014; Nguyen, Liu, Scheinberg, & Takac, 2017). These papers only use covtype as a test dataset towards their end goals.

Other research in forest species classification has been carried out in recent years, but with differently gathered data and different goals. Some newer machine learning research is related to forestry, but has broader scope (Ahmed, Franklin, Wulder, & White, 2015; Garcia-Gutierrez, Martinez-Alvarez, Troncoso, & Riquelme, 2015). These papers use remote sensed 3D LiDAR data and focus on other machine learning applications such as building environmental models or predicting forest canopy coverage and height.

The closest recent work to what Blackard and Dean did is a series of papers that use different machine learning models to predict forest species based on hyperspectral remote sensed data (Shang & Chisholm, 2014; Pu & Liu, 2011). This research is different because the data is remote sensed and looks at different machine learning models, such as Support Vector Machine (SVM), AdaBoost, Random Forest (RF). They do not use ANNs. They also rely on the spectral aspects of leaves, rather than cartographically measured attributes of the tree sites in general. Further research in this area is available in the references sections of these papers.