Mining Meteorological and Forestry Data to Enhance Afforestation in Ethiopia

Hiwot Getahun
hiwotyitebitu@yahoo.com

Berhanu Borena
PhD Candidate, Addis Ababa University, Ethiopia
berhanuborena@gmail.com

Abstract

Forests are significant resources in Ethiopia both economically and ecologically as they can facilitate watershed protection, biodiversity conservation, and carbon sequestration. Forests are a critical component of the planet's ecosystem. Unfortunately, there has been significant degradation in forest cover over recent decades due to rapid population growth, extensive forest clearing for cultivation, and over-grazing, movement of political centers, and exploitation of forests for fuel wood and construction materials without replanting. Therefore, it is crucial to increase forest coverage in Ethiopia. One way to do this is through afforestation.

This paper attempts to apply data mining techniques of classification in forestry sector to enhance afforestation in Harerghe region of Ethiopia by recommending suitable tree species for that locality. The data were collected from sources such as International Center for Research in Agroforestry (ICRAF) forestry database, National Meteorology Agency, and data generated through processing. The paper aimed to predict suitable tree species in the selected 18 locations/weredas in Harerghe region using J48 and RandomTree decision tree algorithms. The result of the study has shown that the data mining techniques are valuable in the prediction of tree species using J48 decision tree algorithm.

Keywords: Afforestation; Data Mining; J48; RandomTree

1. Introduction

Destruction of the natural forests of Ethiopia results directly in the loss of unaccounted plant and animal species as well as in a shortage of fuel wood, timber, and other forest products. It also indirectly leads to more aggravated soil erosion, deterioration of the water quality, further drought and flooding, reduction of agricultural productivity, and to an ever-increasing poverty of the rural population. It is obvious that the depletion of forest resources contributes significantly to the climatic and physical environment change. To worsen the matter, the reforestation effort is not, by any means, matching with the rate of deforestation.

Plantation development is a key strategy to address the problem of deforestation and supplement the shortage of supply of woods from natural forests [1].

Forestry and forest development is receiving attention by the government of Ethiopia as it can be demonstrated by the widespread tree planting activities since the turn of the Ethiopian Millennium [2]. Recently, the contribution of forests towards mitigation of climate change is also adding to the extent of the forest development activities globally, and in the country [3].

Despite such understanding of planting trees, accumulation of traditional knowledge and widespread afforestation exercise, there is a challenge in successfully establishing and growing trees. The growth of trees like any plants depends on climate, soil, and other site factors. Among these rainfall and temperature are the predominant factors determining forest growth in the tropics, modified by local topography and soil types [4]. The most important step in forest plantation development is selecting the most suitable site for a particular tree species. Before plantation of seedlings, careful study should be undertaken about which type of tree species requires what amount of rainfall, in which altitude the tree will grow and the minimum or maximum temperature it needs. However, the conventional approach to identify suitable sites involves tedious procedure such as computing site productivity indexes, indicator species (looking at

indicator species) or characterizing the sites in terms of climate and soil through detail data collection and analysis for a particular site.

Another approach that the forestry research center experiences is trying to plant some tree species in a given area and examine its survival in that area. This task requires years to attain the result and determine the given tree species growth. The reason for these steps to plant trees is lack of adequate information about what tree type should we plant where. The above situations make it difficult the enhancement of forest cover in Ethiopia.

This problem seems to require comprehensive evaluation of options for solution In this regard, data mining can play a role in analyzing tree properties and selection of locations to plant trees under certain conditions. Therefore, the general objective of this work is to study the application of data mining techniques in the forestry sector in order to predict suitable tree species in Hararghe region, Ethiopia.

The rest of the paper is organized as follows. Section 2 presents the methodology. In Section 3, an attempt is made to reviews related work on the application of data mining techniques in forestry sector. Section 4 describes the data preprocessing steps. Section 5 presents the experiment and results. Some conclusions and recommendations are provided in Section 6.

2. Methodology

The dataset were collected from National Meteorology Agency and ICRAF Forestry database. To build up ideas and gather information for this work, various literature are reviewed. Weka 3.7.5 was used for data analysis. To generate the altitude information, ArcGIS tool called ArcMap 10 is used and the input data are Ethiopia-woreda and Ethiopia map.

The work follows CRISP-DM (Cross Industry Standard Process for Data Mining) process framework. CRISP-DM, the current process model for data mining, provides an overview of the life cycle of a data mining project. CRISP-DM embraces six phases such as business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

3. Related Work

Orlander [5] has made an attempt for tree species selection to site using the survival and growth of some tree species planted in Tigry, Wollo, Suba Menagesha, Munesa forest and Holeta regions depending on five climatic zones. The zones were classified according to their annual rainfall given in National Atlas of Ethiopia. the author obtained the data about growth and survival of some trees (experiment on the sites made by UNDP|FAO project ''Assistant to Forestry Research, Ethiopia) from the 1985 inventory.

Cortez and Morais [6], used recent real-world data (meteorological data), collected from the northeast region of Portugal, for the purpose of predicting the burned area (or size) of forest fires. They carried out several experiments by considering five data mining techniques (i.e., Multiple Regressions, Decision Tree, Random Forest, Neural Network, and Support Vector Machine) and four feature selection setups (i.e., using spatial, temporal, the Fire Weather Index system, and meteorological data). The proposed solution includes weather variables such as rain, wind, temperature and humidity in conjunction with a Support Vector Machine and it is capable of predicting the burned area of small fires, which constitute the majority of the fire occurrences. Such knowledge is particularly useful for fire management decision support.

Barry O'Sullivan *et al.* [7] applied regression techniques from the field of data mining to predict several biodiversity measures using physical attributes of the forest. They used a variety of regression techniques in Weka, specifically least median squared linear regression, linear regression, multi-layer perceptron, pace regression linear models, and regression trees.

Stojanova1 *et al*. [7] used predictive models and applied different data mining techniques to predict forest fires in three different regions in Slovenia. On the data sets from these regions, they applied logistic regression and decision trees (J48), as well as random forests, bagging and boosting of decision trees, in order to obtain predictive models of fire occurrence. According to the authors, the best results obtained in terms of predictive accuracy, precision

and kappa statistics were by bagging decision tree with the accuracy of 81.2, 86 and 84.9 in Kras, Primorska and continental of Slovenia regions, respectively.

Han *et al.* [9] proposed two statistical based predictive geo-spatial data mining methods and applied them to predict the forest fire hazardous area in the Youngdong region of Kangwon province, Republic of Korea. The proposed prediction models used in geo-spatial data mining are likelihood ratio and conditional probability methods. Using geographic maps, forestry maps and forest fire history, a spatial data mining method has been developed for analyzing the forest fire hazardous area. Then, predictive power of each model has been evaluated, after carrying out cross validation between the models. In comparison of the prediction power of the two proposed prediction models, the likelihood ratio method is more powerful than the conditional probability method.

Džeroski *et al.* [10] applied a data mining using decision trees to predict forest stand height and canopy cover from Landsat and LIDAR (Light Detection and Ranging) data in order to improve the consistency and accuracy and increase the spatial resolution of some of the supporting information to the forest monitoring system in Slovenia. They used predictive models based on multi-temporal Landsat data and calibrated it with high resolution airborne laser scanning (ALS) data.

On the first paper the five zones classified depending only on the annual rainfall given in National Atlas of Ethiopia. However, literature indicates that tree plantation not only depends on rainfall, but also temperature and elevation and other factors. This clearly shows the limitation of the prior work. To fill the limitation and improve correct tree type for certain places, additional climatic condition like temperature and altitude will be considered and analyzed using data mining techniques.

The remaining papers discussed about how to maintain the existing forest coverage area and control forests especially from forest fire and deforestation by applying different data mining techniques. In the case of Ethiopia, the forest coverage is diminishing to almost 9%. This percentage is very shocking and the government has already started giving attention. Although it is vital to control the existing forest from fire, deforestation, etc., we should give attention to plantation of forest to increase forest coverage area.

4. Data Preprocessing

While DM is a key stage in the knowledge discovery process, the data preprocessing phase often requires considerable effort and time. The purpose of the preprocessing stage is to cleanse the data as much as possible and to put it into a form that is suitable for use in latter stages. The main goal of data preprocessing is the production of the dataset used for modeling by the data mining tool of choice. In this paper the data preprocessing stages include the following:

a. Attribute Selection: the original dataset for this work consists of 39 attributes. However, based on experts' opinion of these attributes, only seven of them are considered relevant for the specific learning task to be undertaken as shown in Table 1.

Table 1: Selected Relevant Attributes

Attribute	*Description*
TreeSpecies	Name of tree species
MinAlt	Minimum altitude
MaxAlt	Maximum altitude
MinRF	Minimum rainfall
MaxRF	Maximum rainfall
MinTemp	Minimum temperature
MaxTemp	Maximum temperature

b. Data cleaning: the data is cleaned by removing a record which has no information (i.e., a record with no value in more than 3 attributes), placing some missing values with question mark.

c. Data Transformation: the required dataset are from two sources. In order to match the two data sets, some conversion has been made.

4. Data Formatting: Weka accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute-Relation File Format) file format. The Excel file is first changed into a comma delimited, Comma Separated Value (CSV) file format. After changing the dataset into a CSV format, the next

step is opening the file with the Weka DM software. Then this file is saved with ARFF file extension.

5. Results

The dataset used in the project originally contains 350 Tree species with their required altitude, rainfall, temperature, soil type, etc., a total of around 35 attributes. Out of the 350 tree species, 217 tree species and 7 attributes are selected with the help of domain experts and literature. These seven attributes are more related to the growth of a tree. The forestry data set is used as training set and the Meteorology data is used as test set. By using J48 decision tree algorithm, the predicted tree species during the experimentation show better than RandomTree algorithm. The result of the experimentation has shown the accuracy of J48 and RandomTree as 96% and 95%, respectively. Although their accuracy is almost similar, their mean error prediction has great difference. The mean error prediction of J48 and RandomTree were 0.0438 and 0.543, respectively. This implies using J48 algorithm in the prediction of tree species is statistically significant.

The tool predicts one tree species at a time for each wereda and it was necessary to repeat the experiment at least five times using each algorithm. Out of the five predicted tree species using J48 decision tree algorithm that are suitable for some weredas are given as follows in Alemaya (Gledistia triacanthos, Grevillea robusta, and Melia azedarach), AsebeTeferi (Gledistia triacanthos, Eucalyptus grandis, and Persea americana), Bedessa (Gledistia triacanthos, Grevillea robusta, and Persea americana), Deder (Eucalyptus grandis, Melia azedarach, and Persea americana), Degahabur (Cassia grandis, and Citrus bergamia), Dengego (Gledistia triacanthos, Eucalyptus grandis, Melia azedarach, and Persea americana), DireDawa (Cassia grandis, and Citrus maxima), Gelemso (Grevillea robusta, Melia azedarach and Persea americana), Girawa (Eucalyptus grandis, Melia azedarach, and Persea americana), Gode (Citrus medica, Antidesma bunius, and Chrysophium cainito).

6. Conclusions and Recommendations

The result of the study has shown that the data mining techniques are valuable in the prediction of tree species using J48 decision tree algorithm.

The result can be used in the real world by adopting the selected DM technique in tree species-site matching instead of a time consuming conventional technique. This cheaper and efficient technique is particularly relevant for Ethiopia at this time as the country is preparing itself for large scale afforestation and reforestation programs.

Based on the finding, two main points are recommended:

- According to the experts' opinion on forestry area, the growth of plants mainly depends on rainfall, temperature, altitude and soil types. But the attribute soil type couldn't be included in this work due to lack of information. Future research can be undertaken by including this attribute and more.
- Although in the study, encouraging results are obtained, growing the size of the test data and increasing the number of trials might get better result in the prediction. Furthermore, identifying tools or developing a system that can predict ranked but multiple tree species at a time can enhance such prediction efforts and its results.

References

[1] Abayneh Derero, Nedash Mamo, and Kaleb Kelemu, "Strategic Actions for Integrated Forest Development in Ethiopia".

[2] Polices to increase forest cover in Ethiopia. http://www.efdinitiative.org/research/ publications/publications-repository/ policies-to-increase-forest-cover-in-ethiopiawordsle

[3] Forestry in Ethiopia: http://en. wikipedia.org/wiki/Forestry_in_Ethiopia

[4] Julian Evans, Plantation Forestry in the Tropics, Oxford University Press, 1992.

[5] Goran Orlander, Growth of Some Forest Trees in Ethiopia and Suggestions for Species Selection in Different Climatic Zones, 1986.

[6] P. Cortez and A. Morais. "A Data Mining Approach to Predict Forest Fires using Meteorological Data", In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, Guimaraes, Portugal, pp. 512-523, 2007.

[7] Barry O'Sullivan *et al*. "Data Mining for Biodiversity Prediction in Forests", 2010.

[8] Daniela Stojanova1, P. Panče, K. Andrej, D. Sašo, and T. Katerina, "Learning to Predict Forest Fires with Different Data Mining Techniques", Slovenia Forestry Institute, 2006.

[9] Jong Gyu Han, Keun Ho Ryu, Kwang Hoon Chi, and Yeon Kwang Yeon, "Statistics Based Predictive Geo-spatial Data Mining: Forest Fire Hazardous Area Mapping Application", 2003.

[10] Sašo Džeroski *et al.*, "Using Decision Trees to Predict Forest Stand Height and Canopy Cover from LANDSAT and LIDAR Data", 2006.