ENSEMBLES OF MULTI-OBJECTIVE REGRESSION TREES: A CASE STUDY FOR PREDICTING THE CONDITION OF REMNANT INDIGENOUS VEGETATION

Dragi Kocev[1], *Sašo Džeroski*[1], *Matt D. White*[2], *Graeme R. Newell*[2], *Peter Griffioen*[3]

[1]Jozef Stefan Institute, Department of Knowledge Technologies, Jamova cesta 39, 1000 Ljubljana, Slovenia. {Dragi.Kocev, Saso.Dzeroski}@ijs.si
[2]Arthur Rylah Institute for Environmental Research, Department of Sustainability and Environment, 123 Brown Street, Heidelberg, Victoria 3084, Australia. {Matt.White, Graeme.Newell}@dse.vic.gov.au
[3]Acromap, Pty. Ltd.,37 Gloucester Drive, Heidelberg, Victoria 3084, Australia. pgriffioen@acromap.com

ABSTRACT

In this paper, we show the application of Multi-Objective Decision Trees (MODTs) and Ensembles of MODTs to environmental modelling. MODTs have ability to make simultaneous prediction of several target attributes. One essential component of ecological studies and planning processes is the assessment of quality, condition or status of stands of the native vegetation or habitat. Recently, ‘Habitat Hectares’ was proposed as an approach for vegetation quality assessment. Habitat Hectares method includes assessments of the retention of characteristics within a site (site condition components) and the nature of the landscape surrounding the site (landscape context components). The data for this study consists of 16967 ‘homogenous’ sites that are described with a total of 40 variables (GIS and remote-sensed data) that include biophysical and spectral data. The data were analyzed using multi-objective regression trees (MORTs) and ensembles of MORTs. The results show that ensembles have better predictive performance than the single MORT or single-objective regression tree (SORTs). Ensembles of MORTs and ensembles of SORTs have approximately equal performance, but ensembles of MORTs are faster to learn. Additionally, we learned predictive models (pruned trees) that can be used to better understand the resilience of indigenous vegetation and landscapes.

1 INTRODUCTION

Multi-Objective Regression Trees (MORTs) are decision trees capable of predicting several target attributes simultaneously [1]. The main advantages of this approach (over building a separate model for each target attribute) are: (1) a multi-objective model is smaller than the total size of the individual models for all target attributes, and (2) such a multi-objective model explicates dependencies between the different target attributes.

Ensembles of MORTs can be used to lift the predictive performance of the MORTs [2]. Ensemble methods construct a set of classifiers for a given prediction task and classify new data instances by taking a vote over their predictions. Ensemble methods improve the predictive performance of their base classifier when used in a single target setting (learn an ensemble for each target attribute separately) [3]. In [2], it is shown that this applies also for the multi-target setting (learn one ensemble for all target attributes). In addition, the ensembles for multi-target predictions should be preferred because they are faster to learn.

In this paper, we apply MORTs and ensembles of MORTs on environmental modelling dataset – modelling of remnant indigenous vegetation.

Governments and other agencies (within Australia) are required to demonstrate their compliance with the policies and legislation that are related to remnant indigenous vegetation [4]. These policies may extend the requisite knowledge base and representation of vegetation beyond just ‘extent’ and ‘type’, to incorporate the notion of ‘condition’ or ‘quality’. Concepts of vegetation condition are typically idiosyncratic and/or context-specific. Recent attempts have been made to clarify these concepts, and develop general and widely applicable metrics and indices for assessing vegetation condition. Recently the habitat hectares approach, a rapid assessment technique, was proposed [5]. The ‘vegetation quality’ in the ‘habitat hectares’ approach is defined as the degree to which the current vegetation differs from a ‘benchmark’ that represents the average characteristics of a mature and long-undisturbed stand of the same plant community. Therefore dissimilar community assemblages such as rainforests and savannah can be compared by employing the same general index. The overall ‘habitat hectares’ index comprises 10 components. Seven of these are related to site characteristics (including structural, compositional and other ecological features). The remaining three components are related to the landscape characteristics (patch size, neighborhood and distance to core area).

Employing the ‘habitat hectares’ approach, 16967 ‘homogenous’ sites within the study area were sampled. Each sampling point was described with a total of 40 variables (GIS and remote-sensed data) that include biophysical and spectral data.

For building predictive models from the data (the descriptions and ‘habitat hectares’ scores for the sampling sites) we used, both MORTs and ensembles of MORTs. MORTs were used in order to obtain models that can explain the problem at hand, while ensembles of MORTs to obtain models that have better predictive performance.

The development of predictive models of condition will contribute towards an understanding of the resilience of indigenous vegetation types and landscapes and the relative importance of biophysical and landscape attributes that influence observed condition states. In addition, spatially explicit models of condition, could when used in conjunction with other data, inform natural resource investment decisions, statutory protection and reserve design, while providing a basis for new forms of environmental accounting.

2 METHODOLOGY

2.1 Multi-Objective Regression Trees

Multi-Objective Regression Trees (MORTs) [1] are regression trees that can predict several numeric target variables at once (Figure 1 depicts a MORT). MORTs are a special instantiation of predictive clustering trees (PCTs) [6]. In the PCTs framework, the tree is viewed as a hierarchy of clusters: the top-node corresponds to one cluster containing all data, which is recursively partitioned into smaller clusters while moving down the tree. MORTs are constructed with a standard top-down induction algorithm. This algorithm uses heuristic that minimizes the intra-cluster variation to select an attribute test in the internal nodes. The heuristic score is calculated as sum over the subsets that are induces by the test. Minimization of the intra-cluster variation results in homogeneous leaves, which in turn results in accurate predictions. The predicted vector (that contains predictions for each target attribute) is the vector mean of the target vectors of the training examples belonging to it. More detailed explanations for MORTs can be found in [1,6].

2.2 Ensemble Methods

Ensemble methods are learning algorithms that construct a set of classifiers (called ensembles) [7]. Each new data instance is classified by combining the prediction of each classifier from the ensemble. For regression tasks, the predictions can be combined using average, while for classification tasks using majority vote. Also, more complex combinations of the predictions can be used [8,9].

A condition for an ensemble to be more accurate than any of its individual members is that the individual classifiers are accurate and diverse [10]. An accurate classifier is one that does better than random guessing on new examples. Two classifiers are diverse if they make different errors on new examples. The diversity can be introduced in several ways: by manipulating the training set (changing the weight of examples [3,11] or changing the weight of attributes[12,13]) or by manipulating the learning algorithm itself [11].

Bagging [3] is an ensemble method that constructs the different classifiers by making bootstrap replicates of the training set that are used to construct individual classifiers. Each bootstrap sample is obtained by randomly sampling training instances, with replacement, from the original training set. The bootstrap sample and the training set have an equal number of instances. Bagging can give substantial gains in predictive performance, when applied to an unstable learner (i.e., a learner for which small changes in the training set result in large changes in the predictions), such as classification and regression tree learners [3].

Random Forest [11] is an ensemble method for trees, where the diversity among the individual classifiers is obtained from two sources: (1) by using bagging and (2) changing the feature set during learning. At each node in the decision tree, a random subset of the input features is taken and the best split is selected from this subset. The size of the random subset is given by a function *f* of the number of descriptive attributes *x* (e.g. $f(x)=1, f(x)=\sqrt{x}, f(x)=\lfloor \log_2 x+1 \rfloor,$ $f(x)=\frac{x}{2}\ldots$). If $f(x)=x$, then random forests are equal to bagging.

The diversity between the individual classifiers, when using **Random Subspaces** [12] method, is obtained with random sampling of the feature space (each individual classifier is learned over randomly chosen feature subspace). The number of retained features is given by the function *f* of the number of descriptive attributes *x* as given above.

Recently, combination of Bagging and Random Subspaces (**SubBag** algorithm) was proposed [13]. This method takes bootstrap replicates of the training set and randomly selects feature subspaces. The difference between this approach and random forests is that here feature subspace is used to learn the whole model (while in Random Forests feature subset is selected at each node). In addition, this method can use variety of learning algorithms as individual classifiers, and Random Forests can be constructed only with trees.

The ensemble methods for multi-objective regression trees are obtained using MORT as a individual classifier. More detailed description can be found in [2].

3 DATA DESCRIPTION

The dataset contains 16967 samples. Each sample is described with a total of 40 independent variables (GIS and remote-sensed data) and 7 dependant variables (the ‘habitat hectares’ score). The ‘habitat hectares’ score was represented with the following components: Large Trees, Tree (canopy) cover, Understorey (non-tree) strata, Lack of weeds, Recruitment, Organic litter and Logs. Each score was calculated comparing the current status of the vegetation with the benchmark (average characteristics of a mature and long-undisturbed stand of the same vegetation community).

The large tree score represents the number of large trees (both living and dead) that are present at the measuring site. Tree canopy score assesses the projective foliage cover of canopy trees in the stand, while the understorey score assesses the abundance of various shrubs and forb/herb strata of a community. The understorey assessment includes only indigenous plant species. The lack of (indigenous) weeds score is calculated from the coverage of non-indigenous and native weed plant species. The recruitment score gives the potential for the recruitment of plant species (that is essential part of the long-term site viability). Litter represents both fine and coarse plant debris less than 10 cm diameter, while logs represent the fallen timber or branches of trees that are substantially detached from the parent tree. More detailed description of the 'habitat hectares' scores can be found in [5].

4 EXPERIMENT SETUP

Two sets of experiments were performed. With the first set of experiments we opted for interpretability of the models, so we learned highly pruned MORT (for prediction of all target variables simultaneously) and regression trees (for prediction of each target attribute separately). The pruning was controlled with setting the parameter minimum instances in a leaf to 2048. The second set of experiments consists of un-pruned MORTs, ensembles of MORTs, regression trees and ensembles of regression trees. With this experimental setting the goal was to obtain models that are as accurate as possible, so later on can be used for drawing maps of the quality of remnant indigenous vegetation.

For combination of the predictions output (voting scheme) of the base classifiers from the ensemble, average was used. The ensembles consisted of 100 un-pruned trees. For building Random Forests, Random Subspaces and SubBag the parameter $f(x)$ was set to $f(x) = \lfloor \log_2 x + 1 \rfloor$ as suggested in [11].

The obtained models were validated using 10-fold cross-validation. For assessing the predictive performance of the obtained models we report the correlation coefficient and root mean squared error (RMSE) in the next section.

5 RESULTS AND DISCUSSION

Table 1 shows the predictive performance of the pruned models. From these results we can note that the MORT has comparable performance with the regression trees for each target attribute. Note that, MORT is smaller than all models together and is faster to learn. The size (sum of internal nodes and leafs) of obtained MORT was 11, while the size of each SORT was 11(except the SORT for LargeTreeScore that had size 13).

Table 1: *Correlation coefficient and RMSE (MORT – Multi-Objective Regression Tree, SORT – Single-Objective Regression Tree)*

Target	**Correlation**		**RMSE**	
	MORT	**SORT**	**MORT**	**SORT**
LargeTreeScore	0.502	0.520	2.905	2.871
TreeCanopyScore	0.671	0.677	1.665	1.652
UnderstoreyScore	0.702	0.707	5.103	5.064
LitterScore	0.715	0.699	1.428	1.461
LogsScore	0.698	0.712	1.491	1.461
WeedsScore	0.784	0.789	3.811	3.773
RecruitmentScore	0.607	0.614	2.592	2.574

Figure 1 depicts the pruned MORT. The predictions for the target attributes are the vectors at each leaf. The ordering of the target attributes in the vector of predictions is given in Table 1.

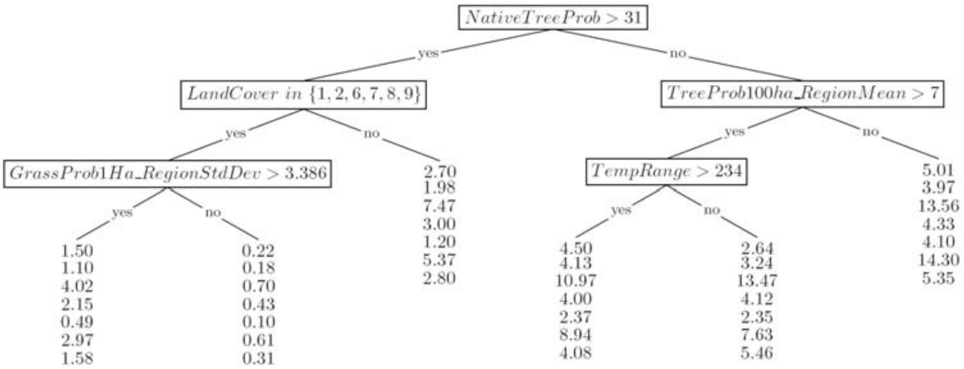

Figure 1: *Pruned MORT*

Tables 2 and 3 show the correlation coefficients and RMSEs of the obtained models. The ensemble methods show better predictive performance than the MORT and SORTs. All ensemble methods have comparable predictive performance, but Random Forests are fastest to learn. Additionally, Random Forests of Multi-Objective Regression Trees are faster to learn, compared to the time needed for learning ensemble for each target attribute separately.

6 CONCLUSIONS

Traditionally most people have thought of native vegetation in terms of extent and type, with relatively few considering its condition or quality. However, the new Government policies (in Australia) increasingly require environmental managers to also consider native vegetation condition. As a result of these policies new ways of measuring the condition of native vegetation have been developed (e.g. Habitat Hectares). These metrics provide simple methods to obtain a score for a site, and to enable comparisons of condition between locations in different vegetation communities.

While this point-based data provides a useful tool for land managers, maps of the condition of native vegetation are an obvious extension to assist identifying the priority areas for restoration and conservation activities. With the modeling approach it is possible to extrapolate point-based native vegetation condition data into a map of remnant native vegetation condition.

Different modeling techniques were used, and their performances were compared. In terms of predictive performances, the ensembles produced approximately equal results. But, Random Forests with multi-objective regression trees should be preferred because they are faster to learn.

Also, in this study interpretable models were learned (pruned trees). These models will be used to further understand the resilience of indigenous vegetation and landscapes.

Table 2. *Correlation Coefficients of the obtained models (MO – Multi-Objective, SO – Single-Objective; RT – Regression Trees, Bag – Bagging, RF – Random Forests, RSub – Random Subspaces, BSub – Bagging and Random Subspaces – SubBag)*

Target	**MORT**	**SORT**	**MOBag**	**SOBag**	**MORF**	**SORF**	**MORSub**	**SORSub**	**MOBSub**	**SOBSub**
LargeTreeScore	0.627	0.601	0.685	0.684	0.690	0.690	0.669	0.667	0.686	0.684
TreeCanopyScore	0.754	0.728	0.798	0.802	0.802	0.803	0.788	0.788	0.798	0.801
UnderstoreyScore	0.779	0.765	0.827	0.826	0.827	0.828	0.812	0.812	0.827	0.826
LitterScore	0.768	0.753	0.812	0.815	0.816	0.815	0.802	0.801	0.812	0.814
LogsScore	0.765	0.744	0.802	0.798	0.800	0.801	0.788	0.786	0.801	0.798
WeedsScore	0.830	0.824	0.872	0.871	0.872	0.873	0.860	0.861	0.872	0.871
RecruitmentScore	0.692	0.677	0.743	0.744	0.744	0.748	0.728	0.728	0.743	0.745

Table 3. *RMSEs of the obtained models (MO – Multi-Objective, SO – Single-Objective; RT – Regression Trees, Bag – Bagging, RF – Random Forests, RSub – Random Subspaces, BSub – Bagging and Random Subspaces – SubBag)*

Target	**MORT**	**SORT**	**MOBag**	**SOBag**	**MORF**	**SORF**	**MORSub**	**SORSub**	**MOBSub**	**SOBSub**
LargeTreeScore	2.618	2.718	2.448	2.451	2.439	2.437	2.527	2.530	2.445	2.451
TreeCanopyScore	1.476	1.563	1.355	1.343	1.344	1.342	1.405	1.407	1.355	1.344
UnderstoreyScore	4.492	4.649	4.034	4.040	4.040	4.023	4.255	4.257	4.033	4.034
LitterScore	1.310	1.352	1.195	1.185	1.186	1.186	1.242	1.244	1.194	1.188
LogsScore	1.340	1.399	1.245	1.256	1.249	1.247	1.290	1.294	1.247	1.255
WeedsScore	3.426	3.506	3.011	3.015	3.013	2.999	3.196	3.181	3.009	3.017
RecruitmentScore	2.357	2.423	2.184	2.180	2.183	2.170	2.262	2.262	2.184	2.176

References

[1] J. Struyf, and S. Dzeroski. Constraint based induction of multi-objective regression trees, Knowledge Discovery in Inductive Databases, 4th International Workshop, KDID'05, LNCS vol. 3933, pp. 222-233, 2006

[2] D. Kocev, C. Vens, J. Struyf, and S. Dzeroski. Ensembles of multi-objective decision trees, Proceedings of 18th ECML, Warsaw, Poland, 2007, LNAI vol.4701, pp. 624-631, 2007 (to appear)

[3] L. Breiman. Bagging predictors, Machine Learning **24** (2), 123–140, 1996

[4] D. Parkes, and P. Lyon. Towards a national approach to vegetation condition assessment that meets government investors' needs: A policy perspective. Ecological Management and Restoration **7**, S3-S5, 2006

[5] D. Parkes, G. Newell, and D. Cheal. Assessing the quality of native vegetation: The 'habitat hectares' approach. Ecological Management and Restoration **4**, S29-S38, 2003

[6] H. Blockeel, L. De Raedt, and J. Ramon. Top-down induction of clustering trees, In Proceedings of the 15th ICML, p. 55–63, 1998

[7] T. Dietterich. Ensemble methods in machine learning, In: J. Kittler, F. Roli (Eds.) MCS 2000. LNCS vol. 1857, p.1-15. Springer, Heidelberg, 2000

[8] T. Ho, J. Hull, and S. Srihari. Decision combination in multiple classifier systems, IEEE Trans. on Pattern Anal. and Mach. Intell. **16**(1), p.66–75, 1994

[9] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers, IEEE Trans. on Pattern Anal. and Mach. Intell. **20**(3), 226–239, 1998

[10] L. Hansen, and P. Salamon. Neural Networks ensembles. IEEE Trans. on Pattern Anal. and Mach. Intell. **12**, 993–1001, 1990

[11] L. Breiman. Random forests, Machine Learning **45**(1), 5–32, 2001

[12] T. Ho. The random subspace method for constructing decision forests, IEEE Trans. on Pattern Anal. and Mach. Intell. **20**(8), 832–844, 1998

[13] P. Panov, and S. Dzeroski. Combining bagging and random subspaces to create better ensembles, In: Advances in Intelligent Data Analysis VII - 7th Int'l Symposium on Intelligent Data Analysis, IDA 2007. Ljubljana, Slovenia. Proceedings. LNCS Vol. 4723. Springer