Models for dominating forest cover type prediction

D A Petrusevich
MIREA – Russian Technological University, Prospekt Vernadskogo, 78, Moscow, 119454, Russia

E-mail: petrdenis@mail.ru

Abstract. The question of the most suitable forest tree species for defined area and landscape has been investigated in the paper. A set of classifiers is constructed in order to build relations between type of soil and other features of forest area and preferable species of trees. The decision tree classifiers, ensemble methods implementing bagging and boosting over such trees are used. The machine learning methods are implemented to obtain the best suited tree species to cover given forest area. This classification task is one of very important problems of forest regeneration process. Efforts of ecologists can have better results if there are expert systems allowing to understand the best forest cover type for areas of forest fires or deforestation that takes place because of human factor. Results and conclusions of this paper can be used in processing of other forest recover tasks. The same methods can be implemented in order to get the preferable tree species for different areas if there’s enough data to solve these tasks with machine learning technique.

1. Introduction
Today deforestation is very important problem for the whole world. It’s caused with human activities in some areas, somewhere seasonable fires are caused with climate and local features. There was a lot of forest fires of giant magnitude recently: in the USA, Australia, Brazil, Siberian regions of Russia. Forest regeneration is a very important ecological problem.

Nowadays data analysis and machine learning are implemented in a lot of different domains of knowledge [1, 2]. In this research the most suitable tree species are defined with machine learning methods. This solution can accelerate process of forest regeneration. Conclusions suite the area where the data was collected in the best way [3]. But the same technique can be used to handle data of different forest areas. Of course, here data collection and dataset creation are very important problems that must be solved in different regions by ecologists [4]. Their efforts help to involve data scientists all over the world to ecological problems solution [5]. Still problems of forest regeneration after fires [6 – 13], agricultural deforestation and regeneration after logging [14 – 16] are usually researched with traditional methods. Now data science and time series analysis methods [17, 18] can be implemented in this domain of knowledge to make predictions of fires and to construct classification and clustering [19] of forest types for regeneration.

2. The dataset structure and classification quality metrics
In the original data competition [3] the main task was to predict the dominant kind of tree cover. The data analyzed in the paper are collected in the Roosevelt National Forest (Colorado, USA). Forest area was divided into cells. Its width and height are 30 m. Each row contains data about such cell. There are
parameters showing distance to the closest water source, fire points and roads. Spatial coordinates include elevation over the sea level, aspect of azimuth in degrees, slope of cell in degrees and hillshade indices. The last values show how much sunlight is obtained in the observed cell. They include values at 9 a.m., noon and 3 p.m. Categorical features describing type of soil (40 values), type of wilderness area (4 values) and type of dominating tree species (type of cover, 7 values) are handled with one-hot encoding. Cover_Type is the main parameter predicted in the data competition [3]. There are 7 types of dominating tree species introduced in the dataset: spruce (fir), lodgepole pines, ponderosa pines, willows (cottonwood), aspen, douglas-fir and krummholz. There are 581012 records in the dataset.

At the largest portion of area (85% of area observed in the dataset) Lodgepole Pine and Ponderosa Pine dominate. It means that the classes in classification problem are unbalanced. One can’t use ordinary accuracy metrics to test quality of classifiers. In this case special metrics is used. Measures of classifiers’ quality usually are precision (2), recall (3) and $F_\beta, \beta = 1$ value (4) that can be considered as their combination [20]:

\[
precision = \frac{TP}{TP + FP},
\]

\[
recall = \frac{TP}{TP + FN},
\]

\[
F_\beta = (1 + \beta^2) \frac{\text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}}.
\]

Correlation coefficients between all pairs of parameters have been considered. Hillshade indices at 9 a.m. and 3 p.m. have got negative coefficient and its value is 78%. It can be explained with daily move of the Sun in the sky. Some area gets a lot of sun emission in the morning. But in the evening illumination is lower at the same place because of the landscape specifics. Correlation between the hillshade at noon and at 3 pm indices can be treated in the same way.

Parameters describing elevation above the sea level and type of soil correlate. Type of soil depends on local climate and height above sea level is one of important factors influencing climate.

Also aspect value correlates with hillshade index measured at 3 pm (65%). It can be explained as description of cells getting oriented approximately to mean trajectory of the Sun at this time.

In pairs of parameters with high correlation one of them is removed.

Horizontal $h$ and vertical $v$ parts of distance to nearest surface water source are combined into new parameter that can be treated as Euclidean distance to that source: $d = (v^2 + h^2)^{\frac{1}{2}}$.

Parameters don’t correlate with type of forest cover. The dataset has got high quality. Linear models and classification models can be used to describe type of forest cover [20].

3. Experiments
The classification task described above has been solved with a few algorithms: “k nearest neighbours” classifier, decision tree classifier and ensemble methods (extra trees classifier, random forest classifier and gradient boosting classifier).

Levels of $F_1$ measures of constructed classifiers are shown in the table 1. Appropriate classes of classifiers in scikit-learn library are presented in the first column. Macroaveraged values of $F_1$ measure for each classifier are shown in the second column.

| Classifier          | Macroaveraged $F_1$ score |
|---------------------|---------------------------|
| KNeighboursClassifier | 90%                       |
| DecisionTreeClassifier | 89%                       |
RandomForestClassifier 94%
ExtraTreesClassifier 95%
XGBClassifier 95%

Ensemble methods unite responses of a lot of “simple” classifiers. Thus, it’s difficult to explain their decisions. At the same time decision tree classifiers operate with just one tree and their behaviour can be explained [20]. Here main parameters of classification process are elevation above the sea level, type of soil, distances to the closest fire points and roads.

The classification tree has got a lot of nodes which contain parts of the investigated examples. There’s a lot of set of conditions which define each class. So, only some simple cases are shown. All classes, except Lodgepole Pine and Ponderosa Pine (dominating at 85% of area), are combined into the third class. Bounds of classes in some cases are shown in the table 2.

Table 2. Some bounds of classes obtained with the decision tree classifier.

| Type of cover       | Conditions                                           |
|---------------------|------------------------------------------------------|
| Lodgepole Pine      | dist_roads $\leq 5338$,                            |
|                     | 2937 $< $ elevation $< 3277$,                        |
|                     | dist_fire $< 2947$,                                 |
|                     | wilderness $= 1$,                                   |
| Ponderosa Pine      | elevation $< 3087$,                                 |
|                     | dist_fire $< 2947$,                                 |
|                     | wilderness $= 1$,                                   |
| Other classes       | elevation $< 2526$ or                               |
|                     | elevation $< 2938$,                                 |
|                     | distfire $< 638$,                                  |
|                     | hillshade3pm $< 93$.                                |

Here dist_fire denotes distance to the closest firepoint, dist_roads means distance to the closest road, wilderness is a type of wilderness area (4 binary values), hillshade3pm is hillshade index at 3 p.m. and elevation shows height of cell above sea level.

Tree ensemble classifiers construct “strong” classifier with a set of “weak” ones which are decision tree classifiers. Work of a lot of classifiers can define the most appropriate subset of dataset, appropriate diapasons of parameters. The ExtraTreesClassifier is an enhanced version of the RandomForestClassifier algorithm and here its results are better. They are based on the bagging idea. The gradient boosting is supposed to be one of the best ensemble methods. It’s based on the boosting technique [20]. But the ExtraTreesClassifier shows the best result in this task.

Principal component analysis [20] has been applied to the dataset. Two components are enough to describe 97% of variance. Subset of 50000 records has been created to make plots of various types of trees in the principal components basis containing two components PC1 and PC2. As it was mentioned above Lodgepole Pine and Ponderosa Pine dominate at the largest portion of forest area (85%). So, the first plot contains information only about these types of trees. The other ones are shown at the second plot.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)},$$

$$x' = \frac{x - \mu}{\sigma}. \quad (5)$$
The components have been scaled. Results of scaling with minimum and maximum values of components are shown at the figures 1 and 2 (scaled according to expression (4)). Implementation of scaling by means of standard deviation and mean value (according expression (5)) delivers results that look the same.

**Figure 1.** Plot of dependence between the first and second principal components of the first 50000 records in the dataset (Lodgepole Pines are marked with white, Ponderosa Pines are black points).

**Figure 2.** Plot of dependence between the first and third principal components of the first 50000 records in the dataset except Lodgepole Pines and Ponderosa Pines (each type is marked with its own color).

4. Conclusion

The Forest cover type dataset has been investigated in this paper. It includes information about tree species of the Roosevelt National Forest (USA). The data competition [3] was aimed to construct cover type classifiers. The most appropriate tree species for given type of landscape need to be found.

Forest recover task is very important ecological work. Forests area decreases steadily because of fires and human activity. There’s need in strong efforts to recover forests all over the world at high speed.

A set of learning models have been used to classify the dataset [3]. As it’s shown at the figures 1 and 2 types of trees are mixed in the dataset. So, linear regression or linear classification models haven’t got appropriate results.

Here the decision trees classifiers, bagging and boosting methods using trees are implemented. Their $F_1$ measures are greater or equal to 89%.

Way of classes distinguishing with decision trees can be explained. Main parameters of classification process are elevation above the sea level, type of soil, distances to the closest fire points and roads. Some bounds of classes obtained with the decision tree are shown in the table 2.

Solutions of such classification tasks can improve efforts of ecologists aimed to recover forests.

References

[1] Petruzevich D A 2020 Implementation of machine learning algorithms in the Sloan Digital Sky Survey DR14 analysis *IOP Conference Series: Materials Science and Engineering* **862** 042005 doi: 10.1088/1757-899x/862/4/042005

[2] Sarkisov S S, Lomonosova N V, Zolkina A V and Sarkisov T S 2020 Integration of digital technology in mining and metallurgy industries *Tsvetnye Metally* (In Russ) **3** 7-14 doi: 10.17580/tsm.2020.03.01

[3] Cortez P and Morais A 2007 A Data Mining Approach to Predict Forest Fires using Meteorological Data *Artificial Intelligence, Proceedings of the 13th EPIA 2007 – Portuguese Conference on Artificial Intelligence* 512-
23 ISBN-13 978-989-95618-0-9

[4] Mikhaïlova T A, Kalugina O V, Shergina O V and Afanasieva L V 2020 Creation of Databases of the Condition of Pine Forests in the Baikal Region Siberian J. of Forest Science (In Russ) 3 3-11

[5] Gao J, Shalini K, Gaur N and Guan X 2017 Data-driven forest fires analysis 2017 Proc. Of IEEE Int. Conf. on Smart City and Innovation doi: 10.1109/UIC-ATC.2017.8397558

[6] Filicetti A T and Nielsen S E 2018 Fire and forest recovery on seismic lines in sandy upland jack pine (Pinus banksiana) forests Forest Ecology and Management 421 32-9 doi: 10.1016/j.foreco.2018.01.027

[7] Bartels S F, Chen H Y H, Wulder M A and White J C 2016 Trends in post-disturbance recovery rates of Canada’s forests following wildfire and harvest Forest Ecology and Management 361 194-207 doi: 10.1016/j.foreco.2015.11.015

[8] Ahmad F and Goparaju L 2019 An analysis of forest fire and climatic parameters’ trend using geospatial technology: a case study in the state of Chhattisgarh, India Siberian J. of Forest Science 2 74-87

[9] Tsytovsk P A and Kudinov E N 2019 Estimation of Fire Resistance of Partially Logged Pine Stands in the Krasnoyarsk Forest-Steppe Siberian J. of Forest Science (In Russ) 5 pp 54-60

[10] Sannikova N S, Sannikov S N, Kochubei A A and Petrova I V 2019 Natural Pine Regeneration on Burns in Forest-Steppe of Western Siberia Siberian J. of Forest Science (In Russ) 5 pp 22-9

[11] Sheshukov M A, Kovalev A P, Orlov A M and Pozdnyakova V V 2020 Problems and Prospects of Protecting Forests from Fires Siberian J. of Forest Science (In Russ) 2 14-20

[12] Arefyev S P and Kazantskeva M N 2020 Periodicity of Fires and Natural Regeneration of Light-Coniferous and Sparse Forests in Nadym District of the Yamal-Nenets Autonomous Okrug Siberian J. of Forest Science (In Russ) 1 3-15

[13] Morozov A S, Ivanova G A, Baksheeva E O and Ivanov V A 2020 Fire Hazard of Young Pine Stands on Unused Agricultural Lands Siberian J. of Forest Science (In Russ) 3 26-36

[14] Taeroe T, de Konig J H C, Løf M, Tolvanen A, Heiðarsson L and Raulund-Rasmussen K 2019 Recovery of temperate and boreal forests after windthrow and the impacts of salvage logging. A quantitative review Forest Ecology and Management 446 304-16 doi: 10.1016/j.foreco.2019.03.048

[15] Flores M P, Pastur G M, Cellini J M and Lencinas M V 2019 Recovery of understory assemblage along 50 years after shelterwood cut harvesting in Nothofagus pumilio Southern Patagonian forests Forest Ecology and Management 450 117494 doi: 10.1016/j.foreco.2019.117494

[16] Semenyakin D A 2019 Natural regeneration after partial logging in pine forests of Krasnoyarsk forest steppe Siberian J. of Forest Science (In Russ) 6 63-71

[17] Andrianova E G, Golovin S A, Zykov S V, Lesko S A and Chukalina E R 2020 Review of modern models and methods of analysis of time series of dynamics of processes in social, economic and socio-technical systems Russ. Technological J. (In Russ) 8(4) 7-45 doi: 10.32362/2500-316X-2020-8-4-7-45

[18] Petrulevich D A 2019 Analysis of mathematical models used for econometrical time series forecasting Russ. Technological J. (In Russ) 7(2) 61-73 doi:10.32362/2500-316X-2019-7-2-61-73

[19] Anfyorov M A 2019 Genetic clustering algorithm Russ. Technological J. (In Russ) 7(6) 134-50 doi: 10.32362/2500-316X-2019-7-6-134-150

[20] James G, Witten D, Hastie T and Tibshirani R 2015 An introduction to statistical learning with applications in R (Springer-Verlag New York) 426 p. doi: 10.1007/978-1-4614-7138-7