Faculty of Computer Science

Data Science

Moscow
2023

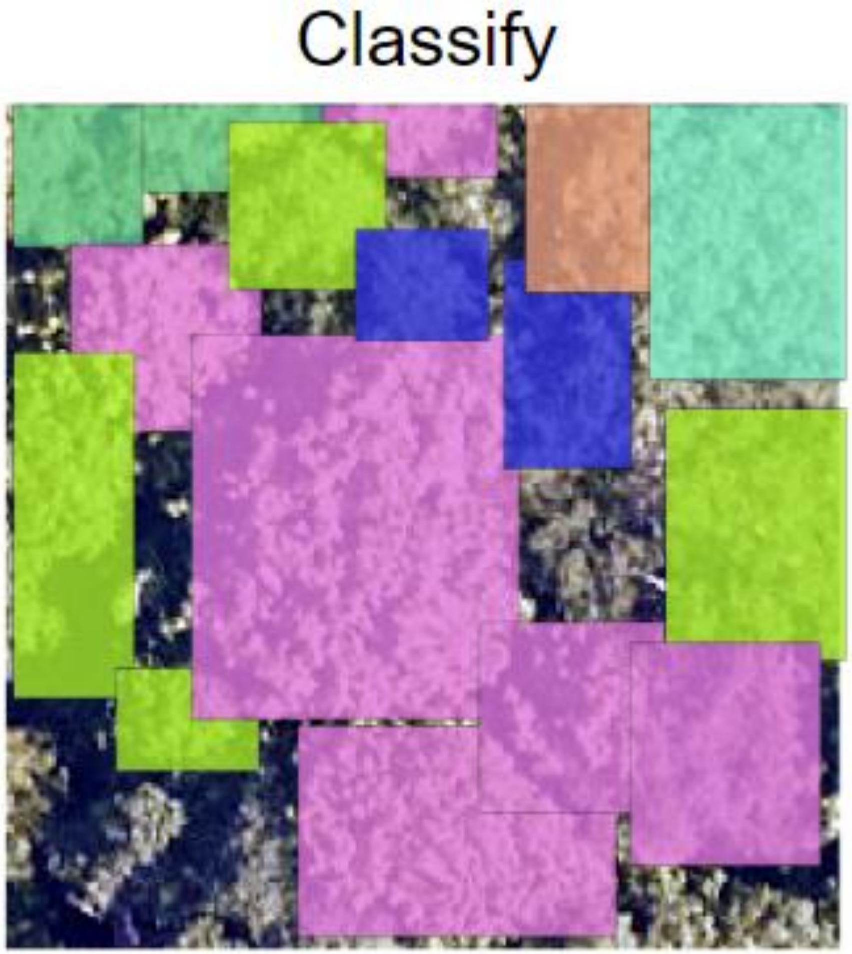
Classification of Tree Types using Earth Remote Sensing Data

Vladislav Bizin

Problem statement

IDTreeS Competition:

- Data science competition to algorithms on two tasks **(1) delineation of tree crowns** and **(2) classification of their species identity** on airborne remote sensing images
- Uses 0.1 – 1 m resolution remote sensing data (RGB, lidar and hyperspectral) from three forest sites of the National Ecological Observatory Network



Participants

Rank	Participant	Method	Cross entropy	Rank-1 accuracy
#1	StanfordCCB	Gradient boosting and random forest ensemble	0.4465	0.9194
#2	FEM	Support vector machine	0.8769	0.88
#3	GatorSense	Multiple instance adaptive cosine estimator (MI-ACE)	0.9386	0.864
#4	Conor	Ensemble of maximum likelihood classifiers based on structural and spectral features	1.2247	0.8226
#5	BRG	Ensemble of maximum likelihood classifiers based on structural and spectral features	1.4478	0.688
	Baseline	Classification based on probability distributions of species frequency in the training data	1.1306	0.6667

Data Specifics

Training Data

Sample ID	Crown ID	Canopy Reflectance (426 Spectral Bands)	Species ID
1	1	…	Type 1
2	1	…	Type 1
3	…	…	…
4	2	…	Type 5
5	…	…	…

Task:

- Predict Species of each Crown

Problem:

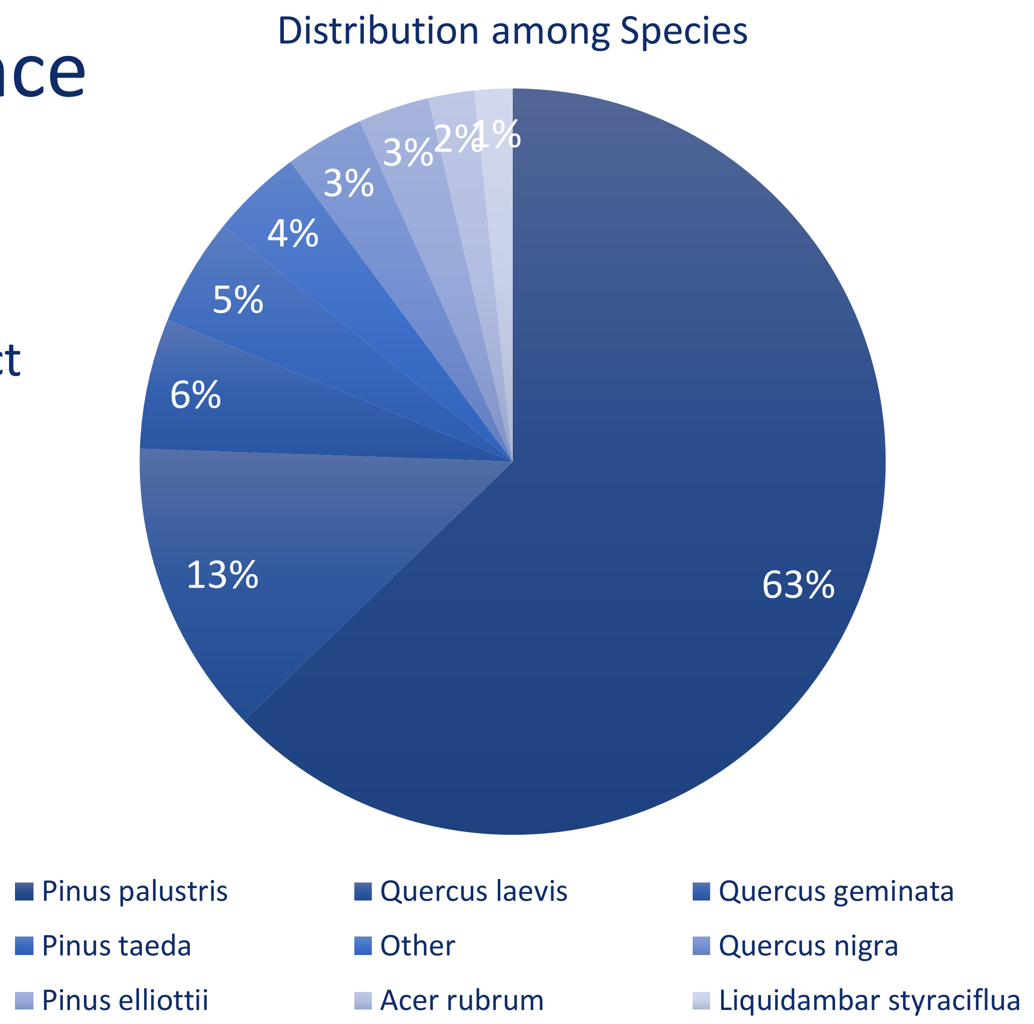
- Each row represents a pixel
- Each tree crown is represented by multiple samples

Solution:

- Predict probability of Species for each Pixel and take mean across Crown

Data Specifics - Class imbalance

Class imbalance can lead models to overpredict common classes when model performance is based on accuracy metric.



Preprocessing

Outlier removal

1. Reflectance values from the blue region of the spectrum (0.38–0.49 mm) and from noisy bands (1.35–1.43 mm, 1.80–1.96 mm, and 2.48–2.51 mm) were removed

426 bands → 1 → 345 bands

2. PCA Transformation + Centering and Scaling to 0 mean, 1 std: exclude samples with values outside of +/- three standard deviations for the first 20 principal components

6831 samples → 2 → 6591 samples

Preprocessing

Transformation
- The rest of the data was transformed via PCA

Dimensionality reduction
- First 100 of 345 principal components were kept

Standard scaling
- Transformed data was then whitened to 0 mean and 1 standard deviation

Dealing with imbalance
- Resampling

Resampling

- <400 samples – oversampled with replacement
- >400 samples - undersampled without replacement

The final training data included 400 samples for each of the nine classes:

3,600 samples total

Model Selection, Training and Calibration

Models:

- gradient boosting classifier
- forest classifier

Training:

- hyper-parameter tuning: selecting the parameters that maximized mean F1 scores in fivefold cross-validation

Tuned parameters:

- number of estimators
- maximum tree depth
- minimum number of samples required to split a node
- minimum node impurity split threshold

Calibration:

- sigmoid regression

Data Split

Scores

Accuracy: 0.919 Cross-entropy: 0.447

	Predicted									
Observed	**SpeciesID**	Acer rubrum	Liquidambar styraciflua	Other	Pinus elliottii	Pinus palustris	Pinus taeda	Quercus geminata	Quercus laevis	Quercus nigra
	Acer rubrum	**1**	0	0	0	0	0	0	0	1
	Liquidambar styraciflua	0	**1**	0	0	0	0	0	0	0
	Other	1	1	**1**	0	0	0	0	0	0
	Pinus elliottii	0	0	0	**0**	1	1	0	0	0
	Pinus palustris	0	0	0	2	**81**	0	0	1	0
	Pinus taeda	0	0	1	0	0	**4**	1	0	0
	Quercus geminata	0	0	0	0	0	0	**4**	0	0
	Quercus laevis	0	0	0	0	1	0	0	**22**	0
	Quercus nigra	0	0	0	0	0	0	0	0	**1**

Flaws

Preprocessing stage:

1. Whitening data to 0 mean, 1 standard deviation after PCA transformation
2. Arbitrary selecting number of principal components kept

Easy to fix

Training and calibration stage:

1. Splitting data into training, calibration and testing based on samples instead of crowns
2. Hyper-parameter tuning in Cross Validation split based on samples instead of crowns

Require custom CV splitter

Scores 2.0

Test Data Scores

Model	Rank-1 accuracy	Cross entropy	F1-macro
Random Forest	0.7857	1.0844	0.3044
CatBoost	0.7937	1.0706	0.4149
XGBoost	0.7381	1.4863	0.3196

Train Data CV Mean Scores

Model	Rank-1 accuracy	Cross entropy	F1-macro
Random Forest	0.8664	0.4109	0.6939
CatBoost	0.8561	0.4690	0.6645
XGBoost	0.8525	0.5862	0.6198

Possible reasons of low scores:

- Bad hyper parameter tuning
- Bad random states

Home Assignment Results and Conclusions

Results:

- One of classification approaches of the competition was studied and recreated
- Crucial flaws of the method were highlighted
- Approach was improved by correcting the flaws
- Other Classifiers and Model-Tuning methods were implemented in a pipeline

Conclusions:

- Preprocessing techniques play critical role in classification tasks
- One should be careful with specific data to not evaluate model on training data

HSE UNIVERSITY • EST 1992 •