Application random decision forests in astroinformatics

Andrej Paličkа[1], Petr Škoda[2]

[1] *Faculty of Information Technology, Czech Technical University, Prague, Czech Republic*

[2] *Astronomical Institute of the Academy of Sciences, Czech Republic*

Abstract

We examine the machine learning algorithm called the Random Decision Forest and its performance on massive astronomical data problems. We present overview of existing implementations and algorithms that are suitable for Big Data, such as entire sky surveys. We wrapped several implementations of the RDF into a package that can be run with a single command and that would be appropriate for a cloud environment. We performed experiments on both classification and regression problems and compared it with other data mining algorithms. We used some well known data sets from the UCI repository for some initial experiments and then concerned ourselves particularly with the classification of stars from their spectra. These experiments show promising results.

Application of Random Decision Forests in Astroinformatics

Andrej Palička

Application of Random Decision Forests in Astroinformatics

Andrej Palička
palicand@fit.cvut.cz

New challenges in astro- and environmental informatics in the Big Data

Challenges in astroinformatics

- Huge amounts of data in surveys
- It's necessary to effectively process and analyze the data
- Potential for massive paralelism

Decision trees

- A decision tree represents relationships between attributes as a tree-like structure
- Internal vertices of the tree serve as criteria tests
- Leafs contain sets of classes that can be assigned after a specimen passes through the tree
- The model clearly represents the structure of the data and is interpretable by a human
- Unfortunately the decision tree is sensitive to noise and easily overfits, losing it's generalization

Decision tree example
Detail

Figure: Excerpt of a spectra classifying tree

Decision forests

- One possible solution to the problem is a (Random) Decision Forest
- A collection of Decision Trees
- Each is built from a random subset of records and attributes
- Each tree therefore "sees" a different part of data
- The trees created this way are weaker than the original tree
- But together they create a stronger model

Experiments

Experiments boli prevedené nad:

- dátami extrahovanými z FITS tabuľky
- dátami, ktoré boli zarovnané a rozdelené do „košov" (binning)
- dátami, z ktorých sme extrahovali dôležité atribúty

Experiments
Data

- Source: archive of Ondrejov observatory
- Approximately 2000 attributes, 1500 records
- Types 1, 2 and 4 represent Be stars, 3 are for other stelar objects
- Each attribute represents an intensity on a given wavelength
- The wavelengths of different records are slightly shifted relative to each other
- Only several attributes are relevant

Experiments
Triedy

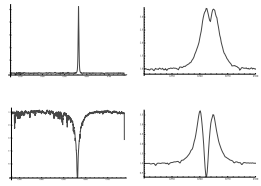

Figure: Spectra types

Experiments
After conversion from FITS

Application of Random Decision Forests in Astroinformatics
Andrej Palička

Table: Confusion matrix for a Random Forest of 2000 trees, *FITS conversion*

actual/predicted	1	2	3	4	recall
1	46	12	0	0	0.79
2	5	24	2	0	0,78
3	0	0	384	0	1.0
4	0	8	9	6	0.26
precision	0.90	0.55	0.97	1.00	

9/15

Experiments
Binning

Application of Random Decision Forests in Astroinformatics
Andrej Palička

Table: Confusion matrix for a Random Forest of 2000 trees, *binning*

actual/predicted	1	2	3	4	recall
1	52	7	0	0	0.79
2	5	63	0	0	0.92
3	0	0	384	0	1.0
4	0	0	0	15	1.0
precision	0.91	0.90	1.00	1.00	

10/15

Experiments
Feature reduction

Application of Random Decision Forests in Astroinformatics
Andrej Palička

Table: Confusion matrix for a Random Forest of 2000 trees, *feature reduction*

actual/predicted	1	2	3	4	recall
1	51	2	0	0	0.96
2	2	63	2	0	0.94
3	0	0	381	0	1.0
4	0	0	2	14	1.0
precision	0.96	0.97	0.99	1.00	

11/15

Experiments
Speed of growth

Application of Random Decision Forests in Astroinformatics
Andrej Palička

Table: Performance of scikit-learn for different forest sizes in seconds

trees/values	time
1	2.04
10	2.52
50	3.86
100	5.95
500	21.45
1000	38.19
2000	74.65

12/15

Experiments
Comparison with other algorithm

Application of Random Decision Forests in Astroinformatics
Andrej Palička

Table: Comparison with several other algorithms, accuracy

	RDF	DL	SVM	kNN
Raw Data	**96.00** %	93.16 %	75.39 %	93.73 %
Binned Data	97.71 %	**98.10** %	85.36 %	96.96 %
Extracted Data	**99.03** %	98,87 %	95.83 %	95.74 %

13/15

Package wrapper

Application of Random Decision Forests in Astroinformatics
Andrej Palička

- Wraps around interfaces of different implementations
- The input is a JSON configuration file, which includes path to data sets
- The output is a JSON file and a HTML/Javascript visualization
- Supported implementations: H2O (Java/REST API), scikit-learn (Python), CUDATree(CUDA implementation in Python)
- Supports different metrics (generalization accuracy, F1, confusion matrix), measurement of time taken for growth, preprocessing

14/15

Conclusion

Application of Random Decision Forests in Astroinformatics
Andrej Palička

- Random forests give precision that is comparable or even exceeds similar algorithms
- Time that it takes to grow the forest is directly proportional to size of the data and number of trees
- There are several implementations that work very well
- However because of huge amount of data we need reliable and scalable implementations

15/15