Website Classifier

Realtime Machine Learning Application

Jaromír Navrátil
Laboratory of Knowledge Discovery, Faculty of Informatics, Masaryk University
In cooperation with Trusted Network Solutions
11. 6. 2013

Synopsis

- Task Reintroduced
- Web Mining
- Random Forest
- Problems
- Solutions not mentioned before
- Tests
- Performance
- Discussion

Task Reintroduced

- Create an application for classifying Czech, Slovak and English web pages.
- 61 classes.
- Multi-labeling (0 - 3 classes per example).
- Learning from pages that were classified manually.
- Realtime classification (C++, Web Content Mining, Random Forest).

Web Mining

- Web usage mining.
  - User logs - IP address, URL and access time and duration.
  - Data from web application.
- Web structure mining.
  - Extracting hyperlinks on website and mining from graphs.
  - Analyzing structure of the document.
- Web content mining.
  - Only content of website is used.
  - In this case, only HTML is parsed.

Random Forest

- Ensemble learning method.
- 50 binary decision trees.
- For each random tree:
  - Randomly select attributes to use (50 from overall number of 2800).
  - Perform bootstrapping of examples.
  - Grow tree using the ID3 algorithm.

Problems

- Encoding detection.
- Too many classes.
- Rare classes (e.g. Sects, Hacking, Sex Education).
- Classes with variable content (such as Hobbies and Chats - Blogs - Forums).
- Welcome pages containing only navigation.
- Rich internet applications (for example Portals - Search Engines).
- Computing thresholds for classifier.

Solution

Thresholds for Classifier

- Random Forest returns probability with which example belongs to class.
- We need yes / no answer.
- 75% of examples are used for growing decision trees.
- The rest is tested on grown trees.
- Thresholds are computed using maximizing F_n measure.

$$F_n \text{ measure} = \frac{((1+n^2)\cdot \text{true positive})}{((1+n^2)\cdot \text{true positive} + n^2 \cdot \text{false negative} + \text{false positive})}$$

Tests

- More tests are to be done.

First version of the program	
true positive	19970
false negative	30571

After thorough changes	
true positive	574
true negative	706132
false positive	77348
false negative	12484
F_5 measure	0.04

Performance

How fast is it?

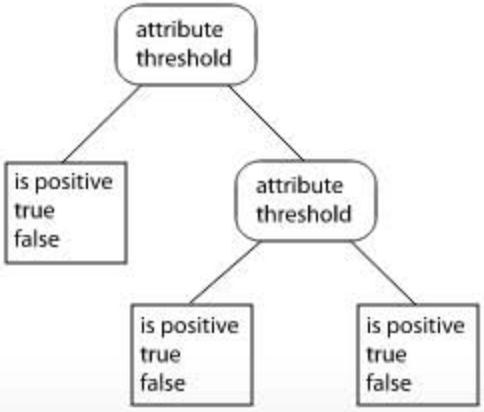
	Classifier	Learning Algorithm
Memory consumption	250 MB	less than 1 GB
Setup time	5 s (loading trees)	10 min (attribute selection)
Time processing	10 ms (already downloaded website)	20 min - 2 h (growing trees)

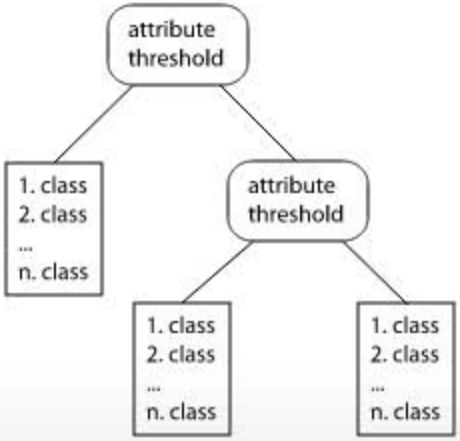
Running on 2.3GHz 64bit processor (Intel Core i5) with 1333MHz DDR3 SDRAM and SSD.

Discussion

What to do next?

- Add outlier detection.
- Compare with other classifiers (Weka's Random Forest, CRM114).
- Use the program for another task:
  - Take advantage of its speed and low memory consumption.
  - Try enormous number of trees.
- Implement different leaf nodes and measure their qualities.





Sources

- Mgr. Juraj Hreško's Thesis, Masaryk University, Faculty of Informatics, 2012
- www.saedsayad.com - website describing machine learning algorithms
- Paper on CUDA implementation of Random Forests.
- Google I/O 2012 slides template.