Kernel Methods and String Kernels for Authorship Analysis

Marius Popescu[1] Cristian Grozea[2]

[1]University of Bucharest, Romania
popescunmarius@gmail.com

[2]Fraunhofer FOKUS, Berlin, Germany
cristian.grozea@brainsignals.de

PAN 2012 Lab

Two Problems, One Approach: Seen from Helicopter

- Character-level N-grams (the best NLP trick ever?)
- TEXT = sequence of symbols = string
- Preprocessing: whitespace seq → single space; uppercase → lowercase
- String kernels
- Kernel-based learning methods: supervised / unsupervised.

String Kernel (Embedding)

- Authorship: p-Spectrum kernel (Histogram):

$$k_p(s,t) = \sum_{v \in \Sigma^p} \mathrm{num}_v(s)\mathrm{num}_v(t)$$

$\mathrm{num}_v(s)$ = the number of occurrences of v as a substring in s.

- Sexual predators: p-grams presence bits kernel (Presence bits):

$$k_p^{0/1}(s,t) = \sum_{v \in \Sigma^p} \mathrm{in}_v(s)\mathrm{in}_v(t)$$

$\mathrm{in}_v(s) = 1$ if v occurs as a substring in s and 0 otherwise.

- Normalized versions of those kernels: self-similarity $K(x,x) = 1$.

Optimum N-gram Length, N=?

- Our (educated) guess: 5
- **Authorship attribution:** long enough to capture function words (typically short): "␣the␣", "␣to␣*", "*␣in␣" but also morphemes like suffixes: "*ing␣".
- **Sexual predator identification:** long enough to capture the ubiquitous "␣asl␣", word stems in English, and short enough to warrant frequent-enough matches between related same-stem words.
- And short enough to show reuse.

Why String Kernels?

Advantages:

- Implicit embedding of the texts in a high dimensional feature space (here the space of all character 5-grams) and the kernel-based learning algorithm aided by regularization implicitly assigns a weight to each feature, thus selecting the features that are important for the discrimination task. **For English, >10 millions features**
- Computation in the feature space is implicit, so it comes (almost) for free.
- Using them leads to language independence (TEXT=string=sequence of characters).
  Chinese? Farsi? No change of the method!
- Trad. NLP: tokenizer, parser, etc; Availability of the tools: **Romanian didn't even have a stemmer until 2007**.

Closed-Class Authorship Attribution: Model Selection

Model selection in ML = Choose your weapons!
Learning method: kernel partial least squares (PLS) regression, because:

- PLS takes directly into account the multi-class nature of the problem.
- PLS is useful when the number of explanatory variables exceeds the number of observations (it has received a great amount of attention in the field of chemometrics).

Tuning

- PLS – just 1 parameter to tune, # of latent components (iterations)
- too small: underfitting; too large: overfitting
- Just 2 samples per author ⇒ we've used the number of training examples (the rank of the training data matrix)
- Target labels encoding: -1/1 one-vs-all

Closed-Class Authorship Attribution: Why not SVM?

Problem	PLS	SVM (ova)	SVM (ovo)	Best result in the competition
A	76.92%	**84.62%**	69.23%	84.62%
B	**53.85%**	38.46%	38.46%	53.85%
C	**100.00%**	88.89%	88.89%	100.00%
D	**75.00%**	50.00%	50.00%	100.00%
E	**25.00%**	25.00%	25.00%	100.00%
F	**90.00%**	90.00%	90.00%	100.00%
G	**50.00%**	50.00%	50.00%	75.00%
H	**100.00%**	33.33%	33.33%	100.00%
I	**75.00%**	50.00%	50.00%	100.00%
J	**100.00%**	50.00%	50.00%	100.00%
K	**50.00%**	50.00%	50.00%	75.00%
L	**75.00%**	75.00%	50.00%	100.00%
M	**75.00%**	75.00%	75.00%	87.50%
Overall	**72.75%**	58.48%	55.38%	70.61%

Table: The results obtained by kernel PLS regression, one-versus-all SVM, and one-versus-one SVM on the AAAC (Juola 2006) dataset problems.

Closed-Class Authorship Attribution: Results

PLS was the right choice

Problem	PLS	SVM (ova)	SVM (ovo)
A	**100.00%**	**100.00%**	83.33%
C	**100.00%**	62.50%	50.00%
I	**92.86%**	78.57%	71.43%
Overall	**97.62%**	80.36%	68.25%

Table: The results obtained by kernel PLS regression, one-versus-all SVM and one-versus-one SVM for closed-class attribution sub-task problems

Open-Class Attribution: Class and Confidence

- We need to decide when to predict a label and when not.
- Kernel PLS regression returns a vector $\hat{Y}$ of real values.
- We have considered that what is important is the structure of $\hat{Y}$ not the actual values of $\hat{Y}$.
- If maximum of $\hat{Y}$ is far enough from the rest of the values of $\hat{Y}$ a prediction can be made, otherwise not.

Open-Class Attribution: Deciding, Results

- We have modeled "far enough" by the condition that the difference between the maximum of $\hat{Y}$ and the mean of the rest of the values of $\hat{Y}$ to be greater than a fixed threshold.
- To establish best value for this threshold we have computed the above statistic for all testing examples of the closed-class problems and have taken the value of the 20% quantile, 0.3333.
- **The results** (accuracy)
  - B: 80.0%
  - D: 76.4%
  - J: 81.2%

Authorship Clustering: Problem Statement

[18 Sept 2012, pan.webis.de]

> *Authorship clustering/intrinsic plagiarism: in this problem you are given a text (which, for simplicity, is segmented into a sequence of "paragraphs") and are asked to* **cluster** *the paragraphs* **into exactly two clusters***: one that includes paragraphs written by the "main" author of the text and another that includes all paragraphs written by anybody else. (Thus, this year the intrinsic plagiarism has been moved from the plagiarism task to the author identification track.).*

Authorship Clustering: Model Selection

Time to choose weapons again ...

- Clustering method: spectral clustering.
- Similarity between observations: p-spectrum normalized kernel of length 5 ($\hat{k}_5$).
- Similarity matrix → similarity graph: mutual k-nearest-neighbor graph with $k = 12$.

Authorship Clustering: Results

Problem	No. of paragraphs	Paragraphs correctly clustered
Etest01	30	30 (100.00%)
Ftest01	20	20 (100.00%)
Ftest02	20	19 (95.00%)
Ftest03	20	16 (80.00%)
Ftest04	20	20 (100.00%)

Table: The results obtained by spectral clustering on the problems having two clusters

Predators Identification: Fix the Rules!

Important message to the organizers:

Fix the rules!

Fix the rules!

Fix the rules!

- *in advance* **and** keep them fixed.
- indeed, it applies to the authorship clustering as well.
- and helps your teaching, if you do any.

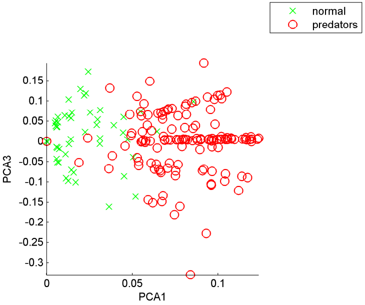
Predators Identification: Paper

Read the paper, it's not bad: two papers in one

- Sexual predator identification problem = classification problem
- chatter: their complete concatenated text labeled as predator on not (single sample).
- Kernel: character *p*-grams presence bits kernel (normalized) of length 5 ($\hat{k}_5^{0/1}$).
- Parallels network intrusion detection / malware analysis: signatures are important and difficult to hide.
- Model: Random forest on reach 8-nearest neighbours information.

Predators Identification: Paper/Results

Results: 7*th* on identifying the predators, 1*st* (thanks but ?!?) on identifying the lines.



Thank you

(and special thanks to **Marius Popescu**
for the wonderful job he did here)