Author Verification: Exploring a Large Set of Parameters using a Genetic Algorithm

Erwan Moreau, Arun Jayapal and Carl Vogel

CNGL & Trinity College Dublin

moreaue@cs.tcd.ie, jayapala@cs.tcd.ie, vogel@cs.tcd.ie

This research is supported by Science Foundation Ireland (Grant 12/CE/I2267) as part of the Centre for Next Generation Localisation (www.cngl.ie) funding at Trinity College, University of Dublin.

PAN 2014

Approach

- Regression problem (at the dataset level)
  - one instance = one problem (known docs + unknown doc)
  - optimize **AUC** × **c@1**
- **Robust strategy:** simple, reliable but not optimized
- **Fine-grained strategy:** maximize performance
  - large set of parameters (10^{19} combinations)
  - risk of overfitting
  - **Genetic learning**
  - **Reference corpus**
    - all documents in the dataset
    - assumption: variability among authors

The robust strategy

- **Only four features**
- A simple similarity measure
  - based on Jaccard similarity
  - characters 4-grams
- A simple consistency measure
  - Difference of the relative frequencies
  - Mean at document level

$$J_1 = \frac{(p+q)}{(p+q+r)} \quad J_2 = \frac{(p+r)}{(p+q+r)}$$

with:

p = n-grams in both X and Y

q = n-grams in X but not in Y

r = n-grams in Y but not in X

The fine-grained strategy

- Algorithm = step-by-step process controlled by parameters
- Goal: find an optimal **configuration**
  - set of parameter/value pairs
  - defines the features, methods, thresholds, ML options...
- The configuration is **generic**:
  - represents **how** to capture an author's style
  - Example: using words bigrams? $\neq$ specific words bigrams
- Regression model

Observations types

- *n*-grams
  - words (3), characters (3), POS tags (4)
  - Combinations with skip-grams (8)
    - e.g. "`<token> ___ <POS tag>`"
- stop-words *n*-grams (3)
  - *n*-grams, only most frequent words
    - e.g. "`the ___ ___ is ___`"
- word length (1), Token-Type Ratio (1)

- Thresholds
  - min. frequency in a document
  - min. proportion of documents which contain the observation
    - known docs
    - reference corpus

Abstract indicators

- **Consistency**
  - how constant is the observation accross known documents?
  - requires at least two known documents
  - standard deviation, min-max range, ...
- **Divergence**
  - how specific is the observation to the author?
  - against the reference corpus
  - mean/median difference, Bhattacharrya, ...
- **Confidence**
  - is this observation a good indicator?
  - uses consistency and divergence
- **Distance**
  - compare known vs. unknown doc
  - Cosine, Jaccard, normal distribution-based measures

Scoring stage

- From abstract indicators to features
  - different methods
  - observation level → document level
  - independent values, merge, ignore,...
- Regression model: **SVM, decision trees** (+ variants)
- Optional **score confidence estimation**
  - idea: assign 0.5 (*"don't know"*) to ambiguous cases to optimize c@1

Meta-configuration file (excerpt)

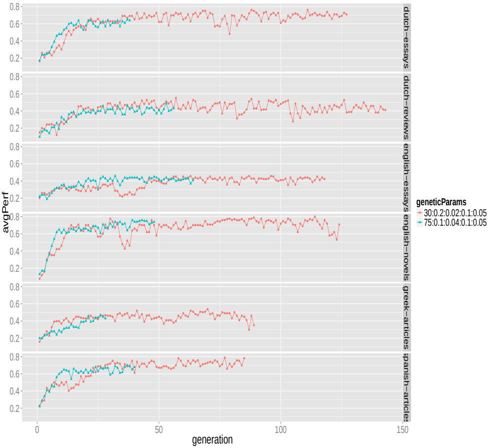
```
obsTypeActive.tokens=0 1
obsTypeActive.w2=0 1
obsTypeActive.w3=0 1
obsTypeActive.STOP3=0 1
obsTypeActive.STOP4=0 1
obsTypeActive.STOP5=0 1
minKnownDocsByObs=0.51 0.3
minRefDocsByObs=0.1 0.25 0.5
minFreqObsIndiv=2 3 5
consistencyValue=stdDev rangeQ1Q3 rangeMinMax stdDevRelMean ratioQ1Q3 [...]
consistencyUseRefIfOnlyOneKnownDoc=0 1
distinctivenessValue=areaCommonDistrib bhattaDiscCoeffDistrib [...]
confidenceMethod=onlyDistinct product mean geomMean constLogDistinct [...]
confidenceFromRanks=0 1
distMethod=euclid cosine jaccard area areaNorma CDF PDF PDFstd
distWithConfidence=no mult multLog multLogInv multSqrt
distMeanType=arithm geom harmo
featuresConfidenceFilterProp=0.05 0.1 0.2 0.5
featuresByObservMaxObserv=5 10 20
featuresIndicatorsMaxObserv=10 25 50 100
featuresIndicatorsMerge=global byObservType
learnMethod=M5P-M4 M5P-M8 SMO-C1-N0 SMO-C1-N1 SMO-C1-N0-RBF SMO-C1-N1-RBF
wekaFeatures=indicators distances all
```

Genetic learning: approach

- Basic algorithm:
  - **population = configurations**
- For each generation
  - measure performance by cross-validation on the training set
  - **rank the configurations by their performance**
  - top configs more likely to be selected as **breeders**
  - next generation generated by crossing over
- Mutations + variants
  - **elitism**: always keep the best configurations
  - **random**: generate new random configurations

Genetic learning: observations

- **Fast convergence**
- Small population sufficient
  - more stable if larger population
- 14,000 to 28,000 configurations evaluated (among 10^{19})
- main training: 3-fold CV
- final stage: 20-fold CV



Best configurations found

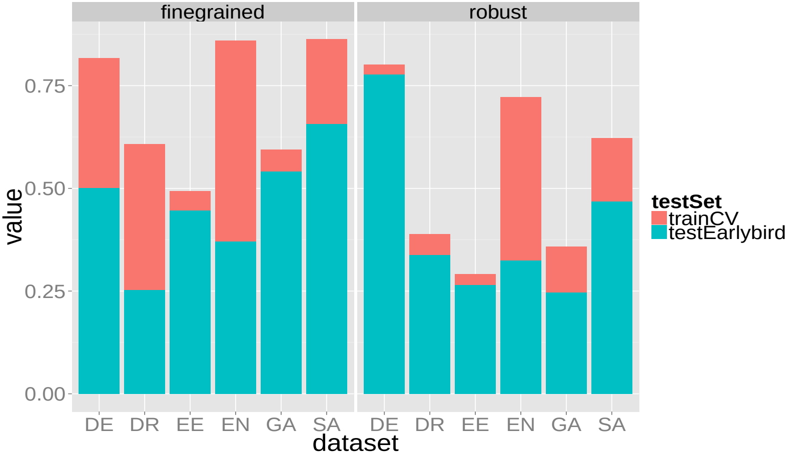
- Few observations types selected: **3 to 11** (among 24)
  1. Words *n*-grams, POS tags
  2. Word length, TTR, stop-words *n*-grams
  3. Unused: characters *n*-grams

- Methods
  - **Consistency unused** in most cases
  - Divergence: **Bhattacharrya coefficient** (more than 1 known doc)
  - Simple distance measures: mean difference, cosine, euclidean
    - frequency weighted with confidence score

- Learning stage
  - Decision trees selected most of the time
  - Confidence estimation model used only once

Final model selection (1)

- Both strategies evaluated on the "earlybird" corpus
  - thanks to the "Tira" system



Final model selection (2)

- Perf. loss lower for robust strategy
  - fine-grained strategy: overfitting probable
  - especially where most cases have only one known document
  - **correlation perf. drop / mean known docs = 0.77**

- English Essays, Greek, Spanish
  - Median known docs / case ≥ 3

  $\Rightarrow$ fine-grained
- Dutch and English Novels
  - Median known docs / case $= 1$

  $\Rightarrow$ robust

Results

Dataset	Final test set			
	robust	fine-grained	final	rank
Dutch essays	0.755	0.563	0.777	4
Dutch reviews	0.375	0.350	0.375	3
English essays	0.325	0.372	0.372	3
English novels	0.313	0.352	0.313	8
Greek articles	0.436	0.565	0.565	3
Spanish articles	0.335	0.634	0.634	2
Macro-average	0.423	0.473	0.502	3
Micro-average			0.451	4

- Selecting strategy by dataset better than any of the two strategies alone
- Hypothesis correlation known docs/performance not confirmed

Results

Dataset	Final test set			
	robust	fine-grained	final	rank
Dutch essays	0.755	0.563	0.777	4
Dutch reviews	0.375	0.350	0.375	3
English essays	0.325	0.372	0.372	3
English novels	0.313	0.352	0.313	8
Greek articles	0.436	0.565	0.565	3
Spanish articles	0.335	0.634	0.634	2
Macro-average	0.423	0.473	0.502	3
Micro-average			0.451	4

- Selecting strategy by dataset better than any of the two strategies alone
- Hypothesis correlation known docs/performance not confirmed

Results

Dataset	Final test set			
	robust	fine-grained	final	rank
Dutch essays	0.755	0.563	0.777	4
Dutch reviews	0.375	0.350	0.375	3
English essays	0.325	0.372	0.372	3
English novels	0.313	0.352	0.313	8
Greek articles	0.436	0.565	0.565	3
Spanish articles	0.335	0.634	0.634	2
Macro-average	0.423	0.473	0.502	3
Micro-average			0.451	4

- Selecting strategy by dataset better than any of the two strategies alone
- Hypothesis correlation known docs/performance not confirmed

Conclusion and future work

- Good results with the genetic learning approach
  - meta-parameters optimized at a reasonable cost
- Benefits from combining the two strategies
  - multiple runs on the Earlybird corpus
  - chance or real specificity in the data?
- Investigate the performance loss with single known document
  - no appropriate method?
- Improve the approach
  - methods and features
  - genetic algorithm