Author Verification: Basic Stacked Generalization Applied To Predictions from a Set of Heterogeneous Learners

Erwan Moreau [1,2], Arun Jayapal [1,2] Gerard Lynch [3] and Carl Vogel [2]

[1] Centre for Next Generation Localisation (CNGL)
[2] School of Computer Science and Statistics, Trinity College Dublin, Ireland
[3] Centre for Applied Data Analytics Research, University College Dublin, Ireland

Task

Given a set of documents written by author A and an unknown document, find whether the latter was written by A.

- Output: probability in [0, 1]
- Evaluation: product of
  - Area under the ROC curve (**AUC**),
  - **c@1** (accuracy with *"don't know"* answer)

Approach

- Supervised classification problem
- Combining multiple learners
- Genetic algorithm:
  - Training individual learners
  - Traning meta-model

Motivations

- Experience from PAN'14
  - two complementary approaches
- PAN'14 meta-classifier performance

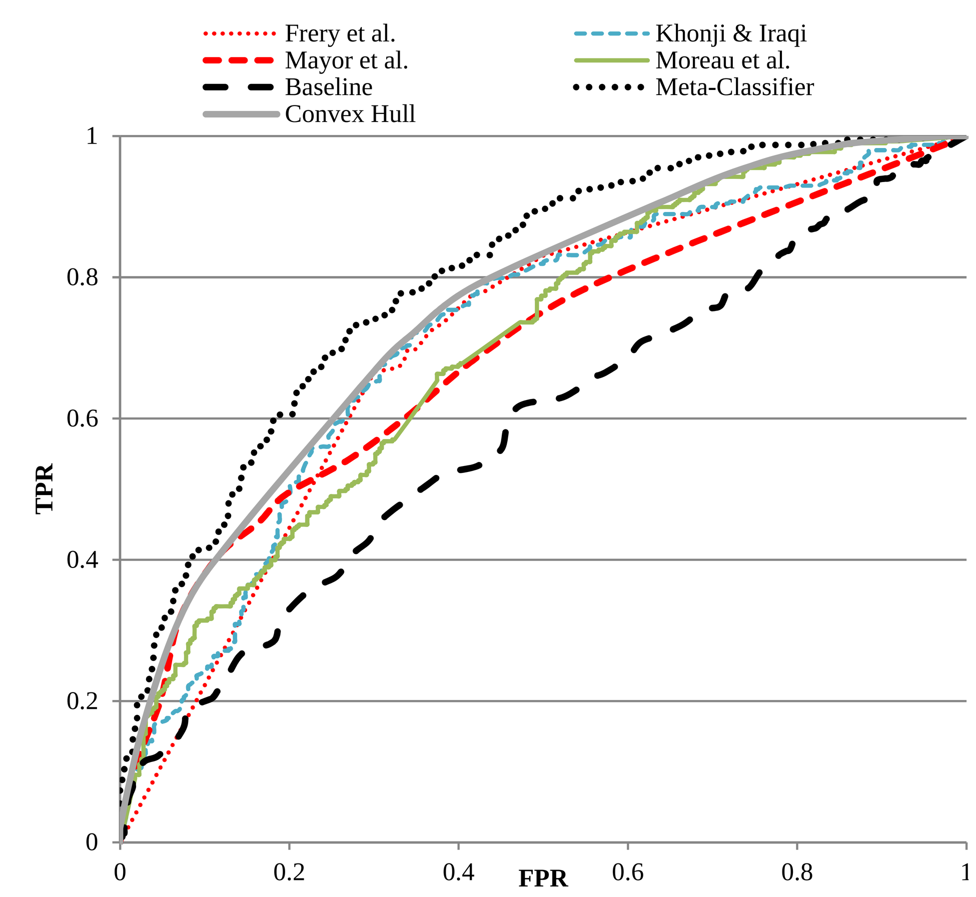


Fig. 1. ROC graphs of the best performing submissions and their convex hull, the baseline method, and the meta-classifier.

Configurations

- Representing distinct sets of parameters in a uniform way
- Set of parameter-value pairs:

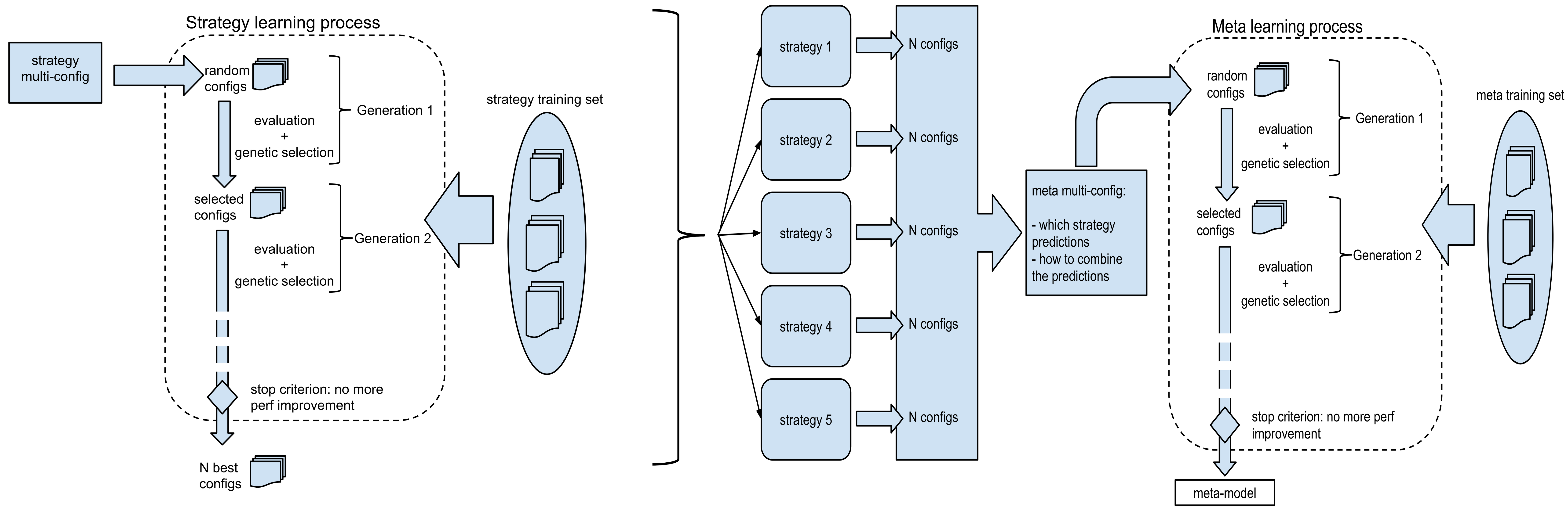
$$C = \{p_1 \mapsto v_1, \ldots, p_n \mapsto v_n\}$$

- Meta-parameters of a strategy
  - Uniquely defines how to train a model
- Very large space of possible configs

Genetic Algorithm

- Configurations = "individuals"
- Selects optimal configs for each strategy
- Parameters (at every generation):
  - Proportion of selected breeders: 10%.
  - Elite prop. :10%; Random; 5%.
  - Probability of mutation: 0.02.

Architecture



For every dataset, 5 strategies are trained individually with the genetic algorithm. Their output are 5xN *"optimal" configurations, which are then fed to the meta-training stage. In this stage, the genetic algorithm selects an optimal combination of configurations.*

Individual Strategies

1. Fine-grained strategy: many parameters, maximize performance
2. Robust strategy: basic approach, safer
3. General Impostor
   - Idea: meta-comparison against third-party documents
   - Used by best system at PAN'14
4. Topic modelling
   - Modified for *style* distinctiveness
   - Complementarity
5. Universum Inference
   - Bootstrapping method
   - Homogenity of documents snippets mixed together

ML Setting

Risk = overfitting

- Genetic process: *inner k*-fold CV
  - New *k*-partitioning at every generation
- Chained sequences with *k* increased
- Final 10 × 2 CV
  - Control the influence of *k*-partitioning

Hybrid setup

- Training set split into:
  - Strategy training: 50% instances
  - Meta-stage training: 25%
  - Meta test set: 25%
- \+ Final eval with bagging
- \+ Overall 2-fold CV

Results

- Influence of the size of the sample
  - English: only one known doc by case
  - Spanish: four known docs by case
- Similar perf on training and test set
  - no overfitting *(except with Spanish)*

Dataset	Meta test set	Full training set	Test set perf.	Test set rank
Dutch	0.710	0.722	0.635	1st
English	0.405	0.421	0.453	6th
Greek	0.656	0.761	0.693	2nd
Spanish	0.950	0.952	0.661	4th
		Macro-average	0.610	2nd

Acknowledgments

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.

Mail: [moreaue|jayapala|vogel]@cs.tcd.ie, gerard.lynch@ucd.ie WWW: http://www.cngl.ie