Motivation for Authorship Verification

- Forensic context
  - Disputed document verification
  - Author can be anyone (besides suspect)
  - From suspect several documents available
Problem Properties in Machine Learning Perspective

- Few reference samples
  - Makes modeling of intra-author variance hard
  - Makes setting of decision threshold hard

- Suitable feature representation required
  - Documents from same author have similar feature values
  - Documents from different authors have different feature values
  - Invariant for specific topic

Unsupervised Learning Approach

- Outlier detection or one-class classification
  - Model normal/reference class

- Reference class contains 1-10 documents
  - Outlier is ill-defined
Supervised Learning Approach

• Separate reference documents from constructed outlier class

• Reference class contains 1-10 documents
  – Small sample size problem

• Data collection for outlier class
  – Leads to strong class imbalance (1:100~1000)

Data Collection: Uninformed

• Virtually impossible to represent outlier class
Data Collection Procedure

- Reference documents are parts (~1000 words) of engineering text books
- Searched for similar books using substrings
- Found 70 books by 50 authors
- Preprocessed similarly to given reference documents
  - Documents of ~1000 words
  - 2-75 documents per book
Feature extraction

- Distance between documents: Compression-based Dissimilarity Method (CDM)
  \[ CD_M(x, y) = \frac{C(xy)}{C(x) + C(y)} \]

- \( C(x) \) is the length of text \( x \) after compression by the PPMd method (best available text compressor)

Submission 1 (S1)

- Straightforward compression distances

- Decision rule: if the nearest document (CDFM) is from the reference class then the documents are written by the same author, otherwise different author
Submission 2 (S2)

- Risk of overfit in S1
- Feature representation
  - distances to prototype set
  - 200 random documents

- LESS classification method
  - Sparse classifier
  - Weights both classes equally
  - Related to $L_1$-SVM

Submission 3 (S3)

- Underrepresentation of reference class in S2

- Boostrapped document samples
  - 50 documents sampled from concatenated reference documents
Results

- On collected data
  - \textbf{S1}: 0.94
  - \textbf{S2}: 0.79

- In PAN Lab evaluation
  - English task only
  - Highest score

- In \textbf{S2} and \textbf{S3} the (sparse) LESS model often uses only 2-3 features to separate reference from outlier class

\begin{table}
\centering
\begin{tabular}{|l|c|c|c|}
\hline
Submission & F$_1$ & Precision & Recall \\
\hline
zhenshi13 & 0.800 & 0.800 & 0.800 \\
serifman13 & 0.800 & 0.800 & 0.800 \\
layton13 & 0.767 & 0.767 & 0.767 \\
morava13 & 0.767 & 0.767 & 0.767 \\
jankowska13 & 0.733 & 0.733 & 0.733 \\
age13 & 0.733 & 0.733 & 0.733 \\
halvani13 & 0.700 & 0.700 & 0.700 \\
feng13 & 0.700 & 0.700 & 0.700 \\
ghemati13 & 0.691 & 0.760 & 0.633 \\
petmanson13 & 0.667 & 0.667 & 0.667 \\
boheccv13 & 0.644 & 0.655 & 0.633 \\
sorin13 & 0.633 & 0.633 & 0.633 \\
vandam13 & 0.600 & 0.600 & 0.600 \\
jayapal13 & 0.600 & 0.600 & 0.600 \\
kern13 & 0.533 & 0.533 & 0.533 \\
baseline & 0.500 & 0.500 & 0.500 \\
gillan13 & 0.500 & 0.500 & 0.500 \\
vladimir13 & 0.467 & 0.467 & 0.467 \\
green13 & 0.400 & 0.400 & 0.400 \\
\hline
\end{tabular}
\end{table}

Conclusion

- Labour intensive approach (data collection)

- Compression features simple and generic

- Robust method
  - Limited sensitivity to number of prototypes and LESS hyper parameter

- All submissions have high performance cross-validated on collected data and on PAN Lab test data
Appendix: LESS Classification Method

\[
\begin{align*}
&\min \sum_{j=1}^{p} w_j + C\left(\sum_{i=1}^{n_t} \xi_{ti} + \sum_{i=1}^{n_o} \xi_{oi}\right) \\
\text{Subject to:} & \\
&\begin{cases}
\quad x \in X_t, \sum_{j=1}^{p} w_j f(x, j) \geq 1 - \xi_{ti} \\
\quad x \in X_o, \sum_{j=1}^{p} w_j f(x, j) < -1 + \xi_{oi}
\end{cases}
\end{align*}
\]

Where \(f(x, j) = (x_j - \mu_{tj})^2 - (x_j - \mu_{oj})^2, w_j \geq 0, \xi_{ti} \geq 0, \xi_{oi} \geq 0.\)