Feature Bagging for Author Attribution

PAN - CLEF 2012

François-Marie Giraud / Thierry Artières
LIP6 – University Paris 6 - France
Motivation

• From the littérature on author attribution
  – Hard to beat a simple and efficient system
    
    Linear SVM on bag of features

• Hypothetical explanations
  – Intrinsic difficulty to define relevant stylistic features
    • Stylistic individual features are embedded and hidden in a large amount of features
    • Stylistic features depend on the writer
  – Optimization concern
    • Undertraining phenomenon [McCallum et al., CIIR 2005]
Motivation

• Undertraining phenomenon

Training Document set: Bag of features (words sorted most to less frequent)
Motivation

• Undertraining phenomenon

Training Document set: Bag of features (words sorted most to less frequent)

- Red subset of feature alone allows perfect training set discrimination
- Blue subset of feature alone allows either
- Green subset is useless

Linear SVM

Discrimination based on red features only
Motivation

• Undertraining phenomenon

Test Document containing no RED features.

Linear SVM

Near random prediction
Undertraining investigation

Document: Bag of 2500 features (words sorted most to less frequent)

Training accuracy

Validation accuracy

F-M. Giraud - T. Artières  
PAN/CLEF 2012
Undertraining investigation

Document: Bag of 2500 features (words sorted most to less frequent)

Training accuracy

Validation accuracy

F-M. Giraud - T. Artières  PAN/CLEF 2012
Undertraining investigation

Document: Bag of 2500 features
(words sorted most -> less frequent)

Random X

Training accuracy

Validation accuracy

F-M. Giraud - T. Artières
PAN/CLEF 2012
Principle of feature bagging

- Document
  - Bag of ~ 3000 most frequent words
  - Bag of words
  - Random selection of 50 to 200 features

K base classifiers learned on random subsets of features

- Base classifier 1
- Base classifier 2
- ... (n classifiers)
- Base classifier K

Majority vote

Base classifiers results aggregation

Prediction

Author

F-M. Giraud - T. Artières
PAN/CLEF 2012
Preliminary results

English public available blog corpus

<table>
<thead>
<tr>
<th># features</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Valid</td>
<td>Test</td>
</tr>
<tr>
<td>100</td>
<td>99.8</td>
<td>33.2</td>
<td>32.2</td>
</tr>
<tr>
<td>225</td>
<td>99.8</td>
<td>50</td>
<td>46.1</td>
</tr>
<tr>
<td>600</td>
<td>99.8</td>
<td>55</td>
<td>48.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># features</th>
<th>Train</th>
<th>Valid</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>99.9</td>
<td>82.2</td>
<td>79.4</td>
</tr>
<tr>
<td>225</td>
<td>100</td>
<td>83.9</td>
<td>76.7</td>
</tr>
<tr>
<td>600</td>
<td>100</td>
<td>83.9</td>
<td>76.1</td>
</tr>
<tr>
<td>Single SVM with all 3000 features</td>
<td>100</td>
<td>79.4</td>
<td>71.6</td>
</tr>
</tbody>
</table>
Experimental methodology for PAN

<table>
<thead>
<tr>
<th>Learning stage</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Train</th>
<th>valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>A2</td>
</tr>
<tr>
<td>B1</td>
<td>B2</td>
</tr>
<tr>
<td>C1</td>
<td>C2</td>
</tr>
</tbody>
</table>

Diagram showing the flow from training to validation and further stages.
Experimental methodology for PAN Learning stage
Comments on PAN results

- Less random features works better.
- Better ranks on closed tasks
- Reject method have to be improved
- Interest to use several training/validation split
Perspective: A two Stage Approach

• Motivation
  – The way the classifier behaves when removing features depends on the author [Koppel 2007]

• Investigate mixing
  – this result
  with
  – our feature bagging approach

Author profiles for unmasking method, [Koppel 2007]
Two Stage Approach

1. Bagging Approach
   Learn multiple base classifiers exploiting random selected subsets of features.

2. Building new data (called profile) for each pair (document, author)

3. (Optional) sort all vectors of the new dataset according to.

4. Learn a binary classifier to say if a profile is correct or not

Profile vector for document d and author a

$j^{th}$ component stand for the percentage of base classifier that exploit feature j and predict a
Two Stage Approach

1. Bagging Approach
   Learn multiple base classifiers exploiting random selected subsets of features.

2. Building new data (called profile) for each pair (document, author)

3. (Optional) sort all vectors of the new dataset according to.

4. Learn a binary classifier to say if a profile is correct or not

Similar results as Bagging approach
Conclusion and further works

• Feature bagging approach to enforce exploiting all features
  ⇒ Outperforms the SVM baseline
  ⇒ Should be improved for handling open problems (cf PAN results)

• Similar results of the second approach
  • While different representation
  ⇒ Should be combined
ANY QUESTION ?
## Additional results on PAN

<table>
<thead>
<tr>
<th>TASK</th>
<th>Run Name</th>
<th>K (# splits)</th>
<th>N (# Models / split)</th>
<th># Models overall</th>
<th>Type of feature</th>
<th># Random features</th>
<th>Open/Closed task</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Lip6 1</td>
<td>8</td>
<td>100</td>
<td>800</td>
<td>WORDS-1500</td>
<td>200</td>
<td>closed</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>Lip6 1</td>
<td>8</td>
<td>100</td>
<td>800</td>
<td>WORDS-1500</td>
<td>200</td>
<td>open</td>
<td>70</td>
</tr>
<tr>
<td>A</td>
<td>Lip6 2</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td>3CHAR-3500</td>
<td>3500</td>
<td>closed</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>Lip6 2</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td>3CHAR-3500</td>
<td>3500</td>
<td>open</td>
<td>60</td>
</tr>
<tr>
<td>A</td>
<td>Lip6 3</td>
<td>8</td>
<td>100</td>
<td>800</td>
<td>WORDS-1500</td>
<td>300</td>
<td>closed</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>Lip6 3</td>
<td>8</td>
<td>100</td>
<td>800</td>
<td>WORDS-1500</td>
<td>300</td>
<td>open</td>
<td>70</td>
</tr>
<tr>
<td>C</td>
<td>Lip6 1</td>
<td>10</td>
<td>100</td>
<td>1000</td>
<td>WORDS-1500</td>
<td>400</td>
<td>closed</td>
<td>100</td>
</tr>
<tr>
<td>D</td>
<td>Lip6 1</td>
<td>10</td>
<td>100</td>
<td>1000</td>
<td>WORDS-1500</td>
<td>400</td>
<td>open</td>
<td>41.18</td>
</tr>
<tr>
<td>C</td>
<td>Lip6 2</td>
<td>10</td>
<td>300</td>
<td>3000</td>
<td>3CHAR-3500</td>
<td>1000</td>
<td>closed</td>
<td>75</td>
</tr>
<tr>
<td>D</td>
<td>Lip6 2</td>
<td>10</td>
<td>300</td>
<td>3000</td>
<td>3CHAR-3500</td>
<td>1000</td>
<td>open</td>
<td>52.94</td>
</tr>
<tr>
<td>C</td>
<td>Lip6 3</td>
<td>10</td>
<td>300</td>
<td>3000</td>
<td>3CHAR-3500</td>
<td>1250</td>
<td>closed</td>
<td>62.5</td>
</tr>
<tr>
<td>D</td>
<td>Lip6 3</td>
<td>10</td>
<td>300</td>
<td>3000</td>
<td>3CHAR-3500</td>
<td>1250</td>
<td>open</td>
<td>35.29</td>
</tr>
<tr>
<td>I</td>
<td>Lip6 1</td>
<td>12</td>
<td>1</td>
<td>12</td>
<td>WORDS-1500</td>
<td>1500</td>
<td>closed</td>
<td>85.71</td>
</tr>
<tr>
<td>J</td>
<td>Lip6 1</td>
<td>12</td>
<td>1</td>
<td>12</td>
<td>WORDS-1500</td>
<td>1500</td>
<td>open</td>
<td>81.25</td>
</tr>
<tr>
<td>I</td>
<td>Lip6 2</td>
<td>12</td>
<td>1</td>
<td>12</td>
<td>WORDS-2000</td>
<td>2000</td>
<td>closed</td>
<td>78.57</td>
</tr>
<tr>
<td>J</td>
<td>Lip6 2</td>
<td>12</td>
<td>1</td>
<td>12</td>
<td>WORDS-2000</td>
<td>2000</td>
<td>open</td>
<td>68.75</td>
</tr>
<tr>
<td>I</td>
<td>Lip6 3</td>
<td>12</td>
<td>1</td>
<td>12</td>
<td>WORDS-2500</td>
<td>2500</td>
<td>closed</td>
<td>78.57</td>
</tr>
<tr>
<td>J</td>
<td>Lip6 3</td>
<td>12</td>
<td>1</td>
<td>12</td>
<td>WORDS-2500</td>
<td>2500</td>
<td>open</td>
<td>75</td>
</tr>
</tbody>
</table>