Sequential Testing in Classifier Evaluation Yields Biased Estimates of Effectiveness

William Webber
University of Maryland
College Park, USA
wew@umd.edu

Mossaab Bagdouri
University of Maryland
College Park, MD, USA
mossaab@umd.edu

David D. Lewis
David D. Lewis Consulting
Chicago, IL, USA
sigir2013pap@DavidDLewis.com

Douglas W. Oard
University of Maryland
College Park, MD, USA
oard@umd.edu

Introduction

- Goal: Economical assured effectiveness
  - Build a good classifier
  - Certify that this classifier is good
  - Use nearly minimal total annotations

- Common practice (sequential training):
  - Select a fixed “certification” test set
  - Add some training instances
  - Test whether effectiveness target reached
  - Repeat add-and-test as needed

- Key results:
  - Sequential training introduces bias
  - Sequential testing introduces bias
  - Both together introduce bias

Design

- Test Collection
  - Reuters newswire stories (RCV1-v2)
  - 29 topics with ≥ 25,000 positive examples

- Passive Learning
  - Random sampling for training and test
  - 580 randomized runs (20 per topic)

- Notation:
  - $F_1$: “True” effectiveness (on 700K documents)
  - $F_1^*$: Point estimate
  - $\theta$: Lower limit of one-sided 95% conf. int. for $F_1$
  - $\tau$: Target for $F_1$

- Confidence Level:
  - Intended: Desired % of time $\theta \geq \tau$ when we stop
  - Observed: Fraction of 580 runs that exceed $\tau$

Experiments

<table>
<thead>
<tr>
<th>Test Collection</th>
<th>1 Run</th>
<th>20 Runs x 29 Topics</th>
<th>20 Runs x 29 Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential Fixed (1067)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequential Fixed (2000)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequential (50%/50%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Stop Criterion Fail
  - $F_1 \geq \tau$ 100.0%
  - $\theta \geq \tau$ 31.55%
  - Desired 5.00%

- Test Training Sequential Fixed (1067)

Stop Criterion Fail
  - $F_1 \geq \tau$ 53.58%
  - $\theta \geq \tau$ 8.13%
  - Desired 5.00%

- Test Training Sequential Fixed (2000)

Stop Criterion Fail
  - $F_1 \geq \tau$ 68.38%
  - $\theta \geq \tau$ 9.40%
  - Desired 5.00%

- Test Training Sequential (50%/50%)

Stop Criterion Fail
  - $F_1 \geq \tau$ 53.58%
  - $\theta \geq \tau$ 8.13%
  - Desired 5.00%

- Test Training Sequential (50%/50%)

Supported in part by NSF IIS-1065250