The Label Complexity of Mixed-Initiative Classifier Training

Jerry Zhu / University of Wisconsin-Madison
Jina Suh, Saleema Amershi / Microsoft Research

June 22, 2016

http://aka.ms/mixedinitiative
MOTIVATION

- Train a classifier
- Start with no labeled data
- Use human annotators
- Main message: Do not run active learning. We have a better procedure
THEORETICAL LABEL COMPLEXITY
Interactive classifier training

- (Example) train this 1D threshold classifier:

- Human oracle, starting from no labeled data
- Cost = label complexity
- Shall we do active learning?

- Training is computer-initiated: computer picks $x$ to query, no idea where $\theta^*$ is
- Active learning label complexity (realizable, Kulkarni et al. 1993)

$$AL \geq \log(1 - \delta) + \max \left( VC, \log \frac{1}{\epsilon} \right)$$
But an optimal oracle can do better!

- Human knows $\theta^*$, provides optimal teaching set:

- Teaching dimension $TD$ = smallest training set size to teach a concept

- $TD \leq AL$ always [Cakmak & Thomaz 2011; Angluin 2004; Goldman & Kearns 1995]

- $TD \ll AL$ often

- Training is human-initiated: human must pick the teaching set

<table>
<thead>
<tr>
<th></th>
<th>optimal oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>computer-initiated</td>
<td>$AL$</td>
</tr>
<tr>
<td>human-initiated</td>
<td>$TD$</td>
</tr>
</tbody>
</table>
But humans are not always optimal oracles

- Naive oracle: can be arbitrarily bad in picking $x$ (but always gives correct labels $y$)

<table>
<thead>
<tr>
<th></th>
<th>optimal oracle</th>
<th>naive oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>computer-initiated</td>
<td>$AL$</td>
<td>$AL$</td>
</tr>
<tr>
<td>human-initiated</td>
<td>$TD$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

- Best of the two worlds: mixed-initiative training
The mixed-initiative algorithm

1: Data $D = \emptyset$
2: for $i = 1$ to $TD$ do
3:  if human no longer wants to lead then
4:    break;
5:  else
6:    human chooses $(x_i, y_i)$
7:    append $(x_i, y_i)$ to $D$
8:  end if
9: end for
10: run active learning starting from $D$ until completion

<table>
<thead>
<tr>
<th></th>
<th>optimal oracle</th>
<th>naive oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>computer-initiated</td>
<td>$AL$</td>
<td>$AL$</td>
</tr>
<tr>
<td>human-initiated</td>
<td>$TD$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>mixed-initiated</td>
<td>$TD$</td>
<td>$TD + AL$</td>
</tr>
</tbody>
</table>
"Neither optimal nor naive" oracle

Seed oracle: provides one point per positive region

<table>
<thead>
<tr>
<th></th>
<th>optimal</th>
<th>seed</th>
<th>naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>computer-initiated</td>
<td>$AL$</td>
<td>$AL$</td>
<td>$AL$</td>
</tr>
<tr>
<td>human-initiated</td>
<td>$TD$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>mixed-initiated</td>
<td>$TD$</td>
<td>$TD + AL - AL_B$</td>
<td>$TD + AL$</td>
</tr>
</tbody>
</table>
Teacher education

- Goal: naive or seed $\rightarrow$ optimal oracle
- Show analogues: “To teach $\theta'$ you could have used $D'$ (optimal teaching set)”
- Show expert-written explanation
EMPIRICAL LABEL COMPLEXITY
Experiment Setup

Mechanical Turk, between-subjects

481 participants

Integer 1D threshold and interval classifier

Goal: Teach a robot assistant acceptable car prices
1D Threshold Classifier Task

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Input range</th>
<th>AL complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>[10000, 30000]</td>
<td>14</td>
</tr>
<tr>
<td>$\Theta^*$</td>
<td>19000, inclusive</td>
<td></td>
</tr>
</tbody>
</table>

"If your price threshold was $20000 or below, you could show your robot these 2 examples: $20000 is acceptable, $20001 is unacceptable"
1D Interval Classifier Task

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Mixed</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Ed.</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Analogue</td>
<td>4</td>
<td>5</td>
<td>x</td>
</tr>
<tr>
<td>Explanation</td>
<td>6</td>
<td>7</td>
<td>x</td>
</tr>
</tbody>
</table>

Conditions

a*, b* 1260, 1360; inclusive

Input range [500, 1500]

AL complexity 26

Step-by-Step Hints

Would it be necessary to provide one more example of ‘$950 is acceptable’ to train the robot?

- Yes
- No

That’s unnecessary because I already know that all prices between $900 and $1000 are acceptable. Please select the correct answer:
### Procedure

<table>
<thead>
<tr>
<th>Instructions</th>
<th>Teacher Education</th>
<th>Training Task</th>
<th>Post-Task Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cover story</td>
<td>Step-by-step tutorial with quizzes</td>
<td>Human-initiated Computer-initiated</td>
<td>Demographics</td>
</tr>
<tr>
<td>Target concept</td>
<td>Hints</td>
<td>Mixed-initiative</td>
<td>Difficulty</td>
</tr>
<tr>
<td>Instructions</td>
<td></td>
<td></td>
<td>Confidence</td>
</tr>
</tbody>
</table>

**Teacher Education**

- Human-initiated Computer-initiated Mixed-initiative

**Post-Task Survey**

- Demographics
- Difficulty
- Confidence
- Teaching experience
- Attention
- Numeracy
- Teaching strategy
Empirical Label Complexity

**TD 1D Threshold Classifier Binary Search**

- **Human**
  - Label Count:
    - 2: 0, 3: 0, 4: 0, 5: 0, 6: 0, 7: 0, 8: 0, 9: 0, 10: 0, 11: 0, 12: 0, 13: 0, 14: 0, 15: 0, 16: 0, 17: 0, NC: 0

- **Computer**
  - Label Count:
    - 2: 0, 3: 0, 4: 0, 5: 0, 6: 0, 7: 0, 8: 0, 9: 0, 10: 0, 11: 0, 12: 0, 13: 0, 14: 0, 15: 0, 16: 0, 17: 0, NC: 0

- **Mixed**
  - Label Count:
    - 2: 0, 3: 0, 4: 0, 5: 0, 6: 0, 7: 0, 8: 0, 9: 0, 10: 0, 11: 0, 12: 0, 13: 0, 14: 0, 15: 0, 16: 0, 17: 0, NC: 0

**1D Interval Classifier**

- **Human**
  - Label Count:
    - 4: 0, 6: 0, 8: 0, 10: 0, 12: 0, 14: 0, 16: 0, 18: 0, 20: 0, 22: 0, 24: 0, 26: 1, NC: 0

- **Computer**
  - Label Count:
    - 4: 0, 6: 0, 8: 0, 10: 0, 12: 0, 14: 0, 16: 0, 18: 0, 20: 0, 22: 0, 24: 0, 26: 1, NC: 0

- **Mixed**
  - Label Count:
    - 4: 0, 6: 0, 8: 0, 10: 0, 12: 0, 14: 0, 16: 0, 18: 0, 20: 0, 22: 0, 24: 0, 26: 1, NC: 0

**Average**

- **Human**
  - Average: 6.6

- **Computer**
  - Average: 13.1
Optimal Teachers

1D Threshold Classifier

1D Interval Classifier
Seed Teachers

1D Threshold Classifier

<table>
<thead>
<tr>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>∞</td>
</tr>
<tr>
<td>Computer</td>
<td>AL</td>
</tr>
<tr>
<td>Mixed</td>
<td>TD + (AL − AL₂)</td>
</tr>
</tbody>
</table>

1D Interval Classifier

<table>
<thead>
<tr>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td></td>
</tr>
<tr>
<td>Computer</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td></td>
</tr>
</tbody>
</table>
Naïve Teachers

1D Threshold Classifier

<table>
<thead>
<tr>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>∞</td>
</tr>
<tr>
<td>Computer</td>
<td>AL</td>
</tr>
<tr>
<td>Mixed</td>
<td>TD + AL</td>
</tr>
</tbody>
</table>

1D Interval Classifier

<table>
<thead>
<tr>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>14</td>
</tr>
<tr>
<td>Computer</td>
<td>26</td>
</tr>
<tr>
<td>Mixed</td>
<td>30</td>
</tr>
</tbody>
</table>
Effects of Training Paradigms on Optimal Teachers

1D Threshold Classifier

1D Interval Classifier
Benefits of Mixed-Initiative Training

Enables optimal teaching
Prevents over-teaching
Eliminates not-completed (NC) participants
Removes blind search complexity
Effects of Teacher Education

1D Threshold Classifier

- Human
- Human + Analogues
- Mixed
- Mixed + Analogues

1D Interval Classifier

- Human
- Human + Analogues
- Human + Explanation
- Mixed
- Mixed + Analogues
- Mixed + Explanation
BRIDGING THEORY AND HUMANS
Humans alone are inefficient

Humans can provide more than the necessary TD training items. (29% in threshold, 8.1% in interval)

“I taught robot all acceptable price ranges.”

→ Support for mixed-initiative training
Humans are noisy

Human teachers provided wrong labels 3.5% of the time.

Nearly half (19/39) of the participants in computer-initiated, interval

→ Allow for attentive labeling or correcting mislabels
Humans have incorrect mental models

Several participants did not understand the robot

“My teaching strategy was to provide the lowest and highest acceptable prices, then provide some acceptable prices in between the range.”

→ Educate humans how to interact with ML algorithms
Labeling effort = label complexity?

Manually selecting an example requires more cognitive effort than providing a label for a given example.

Computer-initiated: 17.3 labels/min
Human-initiated: 2.8 labels/min

→ Help humans explore data or generate examples efficiently
Future research
Teacher education strategy
Interaction techniques or translation layer
Efficient exploration
Other mixed approach
SUMMARY

• Formal justification of mixed-initiative classifier training
• Label complexity analysis and empirical verification
• Benefits of a mixed-initiative training and teacher education
• Limitations and design implications
• Future research directions
• **Main message:** Mixed-initiative training is a better procedure
Thank you!

Jerry Zhu / University of Wisconsin-Madison
Jina Suh, Saleema Amershi / Microsoft Research

June 22, 2016

http://aka.ms/mixedinitiative