LaBiD: Automating Weak Supervision to Find Missing Labels for Big Data

Mona Nashaat¹, Aindrila Basak¹, James Miller¹, Shaikh Quader²

¹University of Alberta, Edmonton, Canada, ²IBM Canada

Motivation

Obtaining Accurate Labels is Expensive!

About 70% of complex analytical tasks today are related to data preparation. There have to be people who are labeling data for machines to understand. Here’s a situation in which human labor automation driven by ML creates new job opportunities.

Guru Banavar, IBM data scientist

Overall Architecture

1. Unlabeled Dataset
2. Automatically create heuristics (LFs)
3. Apply LFs to the unlabeled dataset
4. Detect LFs Disagreements
5. Use Generative Model to generate Labels
6. Use the labeled dataset to train a classifier

Proposed Solution: LaBiD

LaBid: Labeler for Big Data

- Data Programming (DP)
- Automating Weak Supervision
- Meta Active Learning (Meta-AL)

Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of instances</th>
<th># of attributes</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higgs</td>
<td>11,000,000</td>
<td>28</td>
<td>Physical</td>
</tr>
<tr>
<td>Renewal Sales</td>
<td>1,354,704</td>
<td>11</td>
<td>Business</td>
</tr>
<tr>
<td>Rain Prediction</td>
<td>142,000</td>
<td>24</td>
<td>Business</td>
</tr>
<tr>
<td>Travel Insurance</td>
<td>63,300</td>
<td>11</td>
<td>Business</td>
</tr>
<tr>
<td>Bank</td>
<td>45,211</td>
<td>17</td>
<td>Business</td>
</tr>
<tr>
<td>News</td>
<td>39,797</td>
<td>61</td>
<td>Social</td>
</tr>
<tr>
<td>Credit Card</td>
<td>30,000</td>
<td>24</td>
<td>Business</td>
</tr>
<tr>
<td>Occupancy Detection</td>
<td>20,560</td>
<td>7</td>
<td>Computer</td>
</tr>
<tr>
<td>Magic</td>
<td>19,020</td>
<td>12</td>
<td>Physical</td>
</tr>
</tbody>
</table>

Results

Labelling Accuracy

- Magic
- Occupancy Detection
- News
- Bank
- Renewal Sales

Classification Performance (F1 measure)

- Magic
- Occupancy Detection
- News
- Bank
- Renewal Sales

Existing Approaches

- Data Programming
- Automated Weak Supervision
- Traditional AL
- Meta-AL

2019 CASTLE