Towards Efficient And Reliable Data Curation for Machine Learning

Presenter: Naiqing Guan
Supervisor: Nick Koudas
“AI is akin to building a rocket ship. You need a huge engine and a lot of fuel. The rocket engine is the learning algorithms, but the fuel is the huge amounts of data we can feed to these algorithms.”

— Andrew Ng
Data Curation for ML Pipelines

**Data Creation**
Collecting and annotating datasets for training ML models

**Data Maintenance**
Integrate, update and clean datasets to maintain their value

**Data Organization**
Structuring and indexing the data for efficient queries

**Data Evaluation**
Connect data to business values or task-specific values
Data Curation for ML Pipelines

1. Collect Training Data
   Collect and manage datasets for training models

2. Train Models
   Train ML models with batches of training data

3. Deploy Models
   Push the trained model into deployment

4. Monitor Performance
   Monitor the model’s performance in deployment
Data Curation for ML Pipelines

1. Collect Training Data
   - DataSculpt [under review]
   - ActiveDP [under review]

2. Train Models
   - Train ML models with batches of training data

3. Deploy Models
   - Push the trained model into deployment

4. Monitor Performance
   - FILA [SIGMOD 2022]
DataSculpt

Automatically design label functions by prompting large language models
Programmatic Weak Supervision

Jieyu Zhang et al., A Survey on Programmatic Weak Supervision.
Programmatic Weak Supervision

- Ask human experts to design LFs
  - Require nontrivial efforts and costs

DataSculpt: Ask LLMs to design LFs
  - Will the generated LFs be accurate?
Research Questions

- RQ1: In which cases can large language models design accurate label functions?
- RQ2: How will the current prompting methods, such as chain-of-thought and self-consistency, affect the performance of label function design?
- RQ3: How do different LLMs (GPT-3.5, GPT-4, Llama-2) perform in designing label functions?
DataSculpt Overview

- Train set (unlabeled)
- Valid set
- Text
- LFs
- Train set
- Train end model
- Generate labels
- Filter LFs
- Build prompt
- LF set
DataSculpt Prompts

SYSTEM PROMPT:
You are a helpful assistant who helps users in a sentiment analysis task. In each iteration, the user will provide a movie review. Please decide whether the review is positive or negative. (0 for negative, 1 for positive)
After the user provides input, first explain your reason process step by step. Then identify a list of keywords that helps making prediction. Finally, provide the class label for the input.

USER PROMPT:
Query: dead husbands is a somewhat silly comedy about a bunch of wives conspiring to bump off each others husbands...
Explanation: the review is negative as it thinks the movie is silly.
Keywords: silly
Label: 0

Query: this movie is an extremely funny and heartwarming story about an orphanage...
Explanation: the review is positive as it describes the movie as funny and heartwarming.
Keywords: funny, heartwarming
Label: 1

Query: first the cgi in this movie was horrible i watched it during a marathon of bad movies on the scifi channel...

SYSTEM PROMPT:
You are a helpful assistant who helps users in a chemical disease relation extraction task. In each iteration, the user will provide a biomedical passage, followed by a question asking whether a chemical causes a disease. Please decide whether the chemical causes the disease based on the passage. (0 for the chemical does not cause the disease, 1 for the chemical causes the disease.)
After the user provides input, first explain your reason process step by step. Then provide a list of regular expression such that if a passage matches the regex, it is likely to have the same label with the current input. Use {{A}} to represent the first entity and {{B}} to represent the second entity occur in the user's query. Use [SEP] to separate multiple regular expressions. Finally, provide the class label for the input.

USER PROMPT:
Query: During dipyridamole-induced hyperemia, 12 of the 16 dogs with a partial coronary stenosis had a visible area of hypoperfusion... Does dipyridamole cause hyperemia?
Explanation: The claim states that dipyridamole induced hyperemia, indicating a causal relationship between them.
Regex: {{A}}-induced {{B}}
Label: 1

Query: In the present study we aimed to investigate plasma levels of CGRP during headache induced by the NO donor glyceryl trinitrate (GTN) ... Does GTN cause headache?
Experiment Setup

- 12 real-world datasets, 8 for text classification and 4 for relation classification
- Iteratively prompts 50 query instances to the LLM to design LFs

### Table 1: Datasets used in Evaluation.

<table>
<thead>
<tr>
<th>Task</th>
<th>Domain</th>
<th>Dataset</th>
<th>#Class</th>
<th>#Train</th>
<th>#Valid</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spam Cls.</td>
<td>Review</td>
<td>Youtube [1]</td>
<td>2</td>
<td>1586</td>
<td>120</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>Text Message</td>
<td>SMS [2, 4]</td>
<td>2</td>
<td>4571</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Sentiment Cls.</td>
<td>Movie</td>
<td>IMDB [27, 35]</td>
<td>2</td>
<td>20000</td>
<td>2500</td>
<td>2500</td>
</tr>
<tr>
<td></td>
<td>Review</td>
<td>Yelp [35, 47]</td>
<td>2</td>
<td>30400</td>
<td>3800</td>
<td>3800</td>
</tr>
<tr>
<td>Topic Cls.</td>
<td>News</td>
<td>Agnews [35, 47]</td>
<td>4</td>
<td>96000</td>
<td>12000</td>
<td>12000</td>
</tr>
<tr>
<td></td>
<td>Paper Abstract</td>
<td>ArxivAbs [36]</td>
<td>2</td>
<td>21367</td>
<td>4579</td>
<td>4579</td>
</tr>
<tr>
<td></td>
<td>Biomedical</td>
<td>MedAbs [37]</td>
<td>5</td>
<td>8085</td>
<td>3465</td>
<td>2888</td>
</tr>
<tr>
<td>Question Cls.</td>
<td>Web Query</td>
<td>TREC [4, 22]</td>
<td>6</td>
<td>4965</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Relation Cls.</td>
<td>News</td>
<td>Spouse [9, 31]</td>
<td>2</td>
<td>22254</td>
<td>2811</td>
<td>2701</td>
</tr>
<tr>
<td></td>
<td>Biomedical</td>
<td>CDR [10, 31]</td>
<td>2</td>
<td>8430</td>
<td>920</td>
<td>4673</td>
</tr>
<tr>
<td></td>
<td>Web Text</td>
<td>SemEval [16, 49]</td>
<td>9</td>
<td>1749</td>
<td>200</td>
<td>692</td>
</tr>
<tr>
<td></td>
<td>Chemical</td>
<td>ChemProt [20, 44]</td>
<td>10</td>
<td>12861</td>
<td>1607</td>
<td>1607</td>
</tr>
</tbody>
</table>
Prompting Methods

Downstream Model Performance (Acc/F1)

general purpose datasets

relation classification datasets

domain-specific datasets

Wrench  Few-shot  CoT  SC  SC+KATE
Pre-trained LLMs

Downstream Model Performance (Acc/F1)

Google

- Youtube
- SMS(F1)
- IMDB
- Yelp
- Agnews
- TREC
- ArxivAbs(F1)
- MedAbs
- ChemProt
- CDR(F1)
- Spouse(F1)
- SemEval

Model Performance:
- GPT-4
- GPT-3.5
- llama2-70b
- llama2-13b
- llama2-7b
Key Takeaways

- **RQ1**: In which cases can large language models design accurate label functions?
  
  The evaluated LLMs can design accurate LFs for tasks requiring general knowledge, but falls short in tasks requiring specific domain expertise, or developing pattern-based LFs for relation classification tasks.

- **RQ2**: How will the current prompting methods, such as chain-of-thought and self-consistency, affect the performance of label function design?
  
  While the prompting methods help the LLM make more accurate predictions, they do not help improve LF accuracy in general. However, combining multiple responses to create a larger candidate LF set helps improve the end-to-end performance.

- **RQ3**: How do different LLMs (GPT-3.5, GPT-4, Llama-2) perform in designing label functions?
  
  In general, GPT-4 has the best performance, and Llama-2-70b model has similar end-to-end performance with GPT-3.5. Smaller Llama-2 models (7b and 13b) have problems following the response format.
ActiveDP

Combine active learning with PWS to improve label quality
Motivation

Can we combine weak supervision with strong supervision to improve label quality?

instances with only a few weak labels
conflicting weak labels
uncovered instances
ActiveDP Overview
Label Aggregation

We design a confidence-based method for label aggregation. The threshold parameter is tuned on validation dataset to maximize predicted label accuracy.

\[
\begin{bmatrix}
0.2 \\
0.8
\end{bmatrix}
\text{confidence: 0.8}
\]

\[
\begin{bmatrix}
0.6 \\
0.4
\end{bmatrix}
\text{confidence: 0.6}
\]

\[
\begin{bmatrix}
0.2 \\
0.8
\end{bmatrix}
\text{follow AL model}
\]

\[
\begin{bmatrix}
0.6 \\
0.4
\end{bmatrix}
\text{follow label model}
\]
Active Sampler

The active sampler should select samples that are helpful for both the label model and the AL model, we thus propose a hybrid sampler to balance between these two goals

\[ x^* = \arg \max_x [\text{Entr}(f_a(x))^\alpha \times \text{Entr}(f_l(x, \Lambda))^{1-\alpha}] \]

Where \( f_a(x) \) and \( f_l(x, \Lambda) \) are the soft labels predicted by the AL model and the label model respectively, and \( \text{Entr}(p) = -\sum_j p_j \log(p_j) \) is the entropy of soft labels.
Experiments

Downstream model's accuracy on 6 evaluated datasets
Future Directions

**Specialized LLMs for annotation**
- Domain specific pre-training and finetuning

**Active learning for LLMs**
- Efficient query instance selection methods for imperfect models

**Synergize multiple paradigms**
- Combining weak supervision with instance annotations
Q&A

CREDITS: This presentation template was created by Slidesgo, and includes icons by Flaticon, and infographics & images by Freepik