Efficient automated mapping of internal source codes to OMOP CDM concepts

Automated Concept Mapping System for OMOP using Vector Representations and Cross-hospital Mapping

The searching based manual assignment of standard concepts in Observational Medical Outcomes Partnership (OMOP) to local source codes is a time-consuming and error-prone process. Although the OHDSI-developed tool USAGI was designed for this procedure, the non-English language data and OMOP integration limitations hinder efficient mapping.

Methods

We propose a method that leverages existing mappings from other hospitals, enabling the efficient scaling of a single mapping across all others. The mapping process is automated by computing vector representations of source code texts, which capture the relevant syntactic and semantic features, ensuring that similar records are grouped closely in the vector space. Consequently, assigning concept IDs becomes a matter of performing similarity searches within the vector space.

Example

<table>
<thead>
<tr>
<th>source description</th>
<th>vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creatinine mmol/dL in serum</td>
<td>[0.23, 0.45, 0.32, ... , 0.67]</td>
</tr>
<tr>
<td>Leukocytes x10E9/L (serum)</td>
<td>[0.80, 0.15, 0.20, ... , 0.80]</td>
</tr>
<tr>
<td>White blood cells x10E9/l in serum</td>
<td>[0.78, 0.21, 0.18, ... , 0.82]</td>
</tr>
</tbody>
</table>

Similar text vector representation Close in the vector space

Results

For model evaluation, a comprehensive dataset has been constructed:
- Data sourced from 11 distinct hospitals
- A total of 79,928 unique source codes

Accuracy: 72%
Top-5 Accuracy: 88%

78.07% reduction in mapping validation time achieved