Improve Record Linkage Using Active Learning Techniques

CHONG FENG U4943054
SUPERVISED BY DR. QING WANG
DR. DINUSHA VATSALAN
Outline

1. Introduction
2. Motivation
3. Research Problem
4. Methodology
5. Experiments and Evaluation
6. Conclusion and Future Work
Introduction

- **Record linkage** is the process of identifying and matching records that represent the same real-world entity in a database.

<table>
<thead>
<tr>
<th>aid</th>
<th>name</th>
<th>affiliation</th>
<th>email</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Q. Wang</td>
<td></td>
<td><a href="mailto:qw@gmail.com">qw@gmail.com</a></td>
</tr>
<tr>
<td>2</td>
<td>Qing Wang</td>
<td>Curtin University</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Qing Wang</td>
<td>University of Otago</td>
<td><a href="mailto:qw@gmail.com">qw@gmail.com</a></td>
</tr>
<tr>
<td>4</td>
<td>Qing Wang</td>
<td>ANU</td>
<td><a href="mailto:qwang@anu.edu.au">qwang@anu.edu.au</a></td>
</tr>
<tr>
<td>5</td>
<td>Qingqin Wang</td>
<td>Curtin University</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Wang, Q.</td>
<td></td>
<td><a href="mailto:qing@gmail.com">qing@gmail.com</a></td>
</tr>
<tr>
<td>7</td>
<td>Q. Q. Wang</td>
<td>University of Otago</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Wang, Qing</td>
<td>University of Otago</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Q. Q. Wang</td>
<td>ANU</td>
<td><a href="mailto:qw@anu.edu.au">qw@anu.edu.au</a></td>
</tr>
</tbody>
</table>

Record linkage is based on similarity weight vectors, which represent how similar two records are.

Similarity weight vectors:

(r1,r3) – [1.0, 0.0167, 1.0, 0.667]
(r1,r4) – [0.03, 0.05, 0.0, 0.167]
In the classification process, weight vectors are classified as either **match** or **non-match** based on their similarity scores.

A clustering algorithm is used to group records into entity clusters based on their classified weight vectors.
Motivation

- **Errors** may exist in the clustering results, and these errors are normally undetectable in the linkage process.
- These errors might have been introduced into the system from the **classification** process.
- A following question arises:

  Can we leverage the errors detected by users to improve the classification results as well as the whole record linkage model?
Research Problem

• **Active learning** is a subfield in machine learning and artificial intelligence.
• The idea of active learning is to select the most **informative** training data **iteratively** based on the knowledge it gained from each learning iteration.

• Therefore, the research problem of this project is to find the most **informative** weight vectors to **refine the classification model** based on user-detected clustering errors.
  • The record linkage model should be able to handle clustering errors, locate the errors in the classification results and repair the errors.
  • The record linkage model should be able to refine the classification model as well as improve the clustering results iteratively using active learning techniques.
Methodology

- **An active learning process** is added to the record linkage model, which consists of:
  - user feedback
  - training data selection process
  - reclassification process

- This record linkage model also keeps a **weight vector black list**.
  - Outliers and unrepairable errors will be added to the black list.
  - The black list is updated by the active learning process.
  - Weight vectors stored in the black list will be manually labelled.
Methodology

1. Initial classification
2. First active learning iteration
3. Second active learning iteration
4. After third active learning iteration
Methodology

Algorithm 2 Active learning algorithm

Input:
- A list of weight vector spaces: \( V \)
- A list of training vector spaces: \( T \)
- A block list of weight vectors: \( B \)
- An oracle: \( Oracle \)
- A classifier: \( Classifier \)
- An erroneous weight vector: \( e \)

Output:
- A set of newly classified weight vectors: \( C \)

1: \( b = 0 \)
2: \( C = \{ \} \)
3: for vector space \( v \in V \) do
4: \( \text{if } e \in v \text{ then} \)
5: \( S = \text{MixSelect}(v, e) \)
6: \( \text{if } S == \emptyset \text{ then} \)
7: \( R.\text{append}(e) \)
8: \( v.\text{pop}(e) \)
9: \( \text{else} \)
10: \( b = b + |S| \)
11: \( S^m, S^n = Oracle.\text{label}(S) \)
12: \( S^m = S^m \cup T^n, S^n = S^n \cup T^m \)
13: \( Classifier.\text{train}(S^m, S^n) \)
14: \( W^m, W^n = Classifier.\text{classify}(v) \)
15: \( \text{if } |W^m| = 0 \text{ or } |W^n| = 0 \text{ then} \)
16: \( R.\text{append}(e) \)
17: \( v.\text{pop}(e) \)
18: \( \text{else} \)
19: \( V.\text{pop}(i), V.\text{append}(W^m), V.\text{append}(W^n) \)
20: \( T.\text{pop}(i), T.\text{append}(S^m), T.\text{append}(S^n) \)
21: \( C = C \cup W^m, C = C \cup W^n \)
22: \( \text{end if} \)
23: \( \text{end if} \)
24: \( \text{end if} \)
25: \( \text{end for} \)
26: return \( C \)
Experiments and Evaluation

• Datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of records</th>
<th>Number of weight vectors</th>
<th>Number of clusters</th>
<th>Pair completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCVR Non-corrupted</td>
<td>350,766</td>
<td>100,000</td>
<td>5,000</td>
<td>1</td>
</tr>
<tr>
<td>NCVR corrupted</td>
<td>314,663</td>
<td>100,000</td>
<td>5,000</td>
<td>0.5346</td>
</tr>
<tr>
<td>CORA</td>
<td>1,878</td>
<td>1,764,381</td>
<td>120</td>
<td>N/A</td>
</tr>
</tbody>
</table>

• Measures:
  • Recall, Precision and F-measure
Experiments and Evaluation

Classical measures
Non-corrupted NCVR run 1

Cluster measures
Non-corrupted NCVR run 1
Experiments and Evaluation
Experiments and Evaluation

Classification measures
CORA run 2

Cluster measures
CORA run 2
Conclusion and Future Work

- The experiments proved that the proposed record linkage model can effectively improve classification results and clustering results.
- However, the improvement is also limited by other factors.

Future works:
- Can we integrate more effective blocking strategies with the record linkage model?
- How does the selection of errors affect the quality of our models? Can we arrange the order of errors so the weight vector space is divided in the most appropriate way?
- How to optimize the classifier configuration to achieve the best result?
Thank you!

CHONG FENG U4943054@ANU.EDU.AU