ICD Coding

Zhenghui Wang

Apex Data & Knowledge Management Lab
Shanghai Jiao Tong University
ICD

- ICD: The International Statistical Classification of Diseases and Related Health Problems
- Hierarchical architecture
## Data

### MIMIC III

<table>
<thead>
<tr>
<th>HADMID</th>
<th>189797</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Texts of Discharge Summary</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>DISCHARGE DIAGNOSIS:</td>
</tr>
<tr>
<td></td>
<td>1. Prematurity at 35 4/7 weeks gestation</td>
</tr>
<tr>
<td></td>
<td>2. Twin number two of twin gestation</td>
</tr>
<tr>
<td></td>
<td>3. Respiratory distress secondary to transient tachypnea of the newborn</td>
</tr>
<tr>
<td></td>
<td>4. Suspicion for sepsis ruled out</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Extracted Diagnosis Descriptions</td>
<td>1. Prematurity at 35 4/7 weeks gestation</td>
</tr>
<tr>
<td></td>
<td>2. Twin number two of twin gestation</td>
</tr>
<tr>
<td></td>
<td>3. Respiratory distress secondary to transient tachypnea of the newborn</td>
</tr>
<tr>
<td></td>
<td>4. Suspicion for sepsis ruled out</td>
</tr>
</tbody>
</table>

![Graphs](image.png)

(a) Distribution of diagnosis description count  
(b) Distribution of ICD code frequency  
(c) Distribution of ICD code count
ICD-10 Coding Contest

![Average Accuracy Scores]

“The code entry method was very easy and user-friendly.”
“This is really good for students. And to further their education.”

[https://www.centrallearning.com/codingcontest/]
Solutions

• String to string comparison

绒癌史 → Z85.406，绒毛膜癌个人史

• Multi-label text classification

Q21.100，房间隔缺损

Q21.203，部分性房室隔缺损

I37.000，肺动脉瓣狭窄
Str2Str comparison method

• Longest common subsequence
• Semantic similarity with HowNet
**Longest common subsequence**

- **New LCS**

\[
C[i][j] = \begin{cases} 
0 & (i = 0 \text{ or } j = 0) \\
C[i-1][j-1] + 1 & (i, j > 0, \text{sim}[i-1][j-1] > \varepsilon) \\
\max\{C[i-1][j], C[i][j-1]\} & (i, j > 0, a_i \neq b_j, \text{sim}[i-1][j-1] \leq \varepsilon)
\end{cases}
\] (2)

- **New similarity**

\[
sim(A, B) = \frac{LCSL}{\max\{L(A), L(B)\}}
\]

\[
sim(A, B) = \frac{2 \times LCSL}{L(A) + L(B)} \quad \text{Sim}(A, B) = \frac{(LCSL + 1) \times LCSL}{L(A) \times LCSL + L(B)} \quad L(A) \leq L(B)
\]

Semantic similarity with HowNet

• Sentence similarity:

\[
sim(T_1, T_2) = \frac{1}{2} \left( \frac{\sum_{w \in S(T_1, T_2)} (\maxSim(w, T_2) \cdot idf(w))}{\sum_{w \in \{T_1\}} idf(w)} + \frac{\sum_{w \in (T_2, T_1)} (\maxSim(w, T_1) \cdot idf(w))}{\sum_{w \in \{T_2\}} idf(w)} \right),
\]

• Word similarity:
  • Both in HowNet: \[sim(s_1, s_2) = \frac{\alpha}{\text{distance}(s_1, s_2) + \alpha}\]
  • Otherwise: \[sim(w_1, w_2) = \frac{\text{len}(\text{LCS}(w_1, w_2))}{\text{len}(w_1) + \text{len}(w_2) - \text{len}(\text{LCS}(w_1, w_2))}\]

Semantic similarity with HowNet

• Predict in a hierarchical way

Multi-label text classification

- Multi-label classification
  - Algorithms
  - Evaluation metrics

- Multi-label text classification
  - Binary Relevance
  - Label correlation
  - Label specific text representation
  - Label embedding
  - Others
Multi-label classification

<table>
<thead>
<tr>
<th>Multi-Label Problem:</th>
<th>Output vector:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instance</td>
<td>Classes</td>
</tr>
<tr>
<td>1</td>
<td>A, B</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>A, B</td>
</tr>
<tr>
<td>4</td>
<td>C</td>
</tr>
</tbody>
</table>
Multi-label classification algorithms

Problem transformation

Transform to binary classification

Binary Relevance [Subsection III-B.1]
Classifier Chains [Subsection III-B.2]
Calibrated Label Ranking [Subsection III-B.3]
Random k-labelsets [Subsection III-B.4]

Transform to label ranking

Transform to multi-class classification

Lazy learning

ML-kNN [Subsection III-C.1]
ML-DT [Subsection III-C.2]
Rank-SVM [Subsection III-C.3]
CML [Subsection III-C.4]

Algorithm adaptation

Decision tree

Kernel learning

Information-theoretic

Evaluation Metrics

Evaluation Metrics for ICD Coding

- Over coding & under coding problems

**Figure 1** Quantities used in novel evaluation metrics for evaluation of automated ICD9 coding for different cases (left: prediction path diverges from the gold-standard path; middle: prediction is on the correct path but is too granular; and right: prediction is on the correct path, but is not granular enough).

Normalized divergent path to gold standard \((g-c)/g\)

Normalized divergent path to predicted \((p-c)/p\)

Multi-label text classification

• Binary Relevance
• Label correlation
• Label specific text representation
• Label embedding
• Others
Multi-label text classification

• Binary Relevance
• Label correlation
• Label specific text representation
• Label embedding
• Others
• Binary Cross-entropy objective

$$\min_{\Theta} -\frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{L} [y_{il} \log(\sigma(f_{il})) + (1 - y_{il}) \log(1 - \sigma(f_{il}))]$$


Figure 2: Hierarchical Attention Network.
CNN+D2V

Multi-label text classification

• Binary Relevance
• Label correlation
• Label specific text representation
• Label embedding
• Others
Modeling Label co-occurrence

\[ P(C_i|C_j) = \frac{\exp(w_0 + \sum_{k=1}^{K} w_k \cdot F_k(C_i, C_j))}{1 + \exp(w_0 + \sum_{k=1}^{K} w_k \cdot F_k(C_i, C_j))} \]

where \( \exp(\cdot) \) is the natural exponent, \( F_k(C_i, C_j) \) are a set of \( K \) feature functions tracking various aspects of the codes \( C_i \) and \( C_j \), as explained below, and \( w_k \) are the model weights estimated during the training phase.

This model is intended to capture solely the trends of code co-occurrence, leaving prediction of individual codes from the document to the primary auto-coder. Therefore, it does not use features that

Modeling Label co-occurrence

1. Input:
2. \( D_1 \ldots D_M \): a set of \( M \) documents with manually assigned codes
3. \( MAN(D_1) \ldots MAN(D_M) \): sets of manually assigned codes
4. \( GEN(D_1) \ldots GEN(D_M) \): top-scoring outputs of primary auto-coder
5. For each \( D_i \) in \( D_1 \ldots D_M \):
   6. For each \( C_j^{man} \) in \( MAN(D_i) \):
      7. For each \( C_k^{pred} \) in \( GEN(D_i) \cup MAN(D_i) \):
         8. Extract features for estimate \( P(C_k^{pred} | C_j^{man}) \)
         9. If \( C_k^{pred} \in MAN(D_i) \):
            Generate positive training instance
         10. else:
             Generate negative training instance

Modeling Label co-occurrence

1. Input:
2. \(GEN(D_1) \ldots GEN(D_M)\): top-scoring outputs of primary auto-coder
3. Data structures:
4. \(CURRENT\): map of codes to current scores
5. \(FINAL\): map of codes to final scores
6. \(QUEUE\): priority queue of scored codes
7. For each \(D_i\) in \(D_1 \ldots D_M\):
8. Initialize \(CURRENT\) with \(GEN(D_i)\) using primary auto-coder scores
9. Initialize \(QUEUE\) with \(GEN(D_i)\) using primary auto-coder scores
10. Initialize \(FINAL\) to be empty
11. For \(i\) from 1 to depth of exploration \(d\):
12. Pop \(C_{top}\) from \(QUEUE\)
13. \(FINAL(C_{top}) \leftarrow CURRENT(C_{top})\)
14. For each \(C_k\) in \(QUEUE\):
15. \(CURRENT(C_k) \leftarrow CURRENT(C_k) \times P(C_k|C_{top})\)
16. Update \(QUEUE\) with \(CURRENT\)
17. For each \(C_k\) in \(QUEUE\):
18. \(FINAL(C_k) \leftarrow CURRENT(C_k)\)
19. Output \(FINAL\)

Exploiting Associations between Class Labels

• Association rule extraction
  • E.g., (laptop → wireless mouse) [support: 20%, confidence: 80%]

• Types
  • Positive relationship: $y_1 y_2 \rightarrow y_5$
  • Negative relationship: ~$y_6 \rightarrow ~y_3$
  • Hybrid relationship: $y_3 \rightarrow ~y_6$ or $y_3 ~y_2 \rightarrow y_1$

• Algorithm
  • Filter the extracted rules and keep the high quality rules
  • Apply the final rules in the prediction phase in order to correct the errors (where possible) to improve the classification results.

[Mirzamomen Z, Ghafooripour K. Exploiting Associations between Class Labels in Multi-label Classification[J]. Journal of AI and Data Mining, 2018.]
Better Weight Initialization

Multi-label text classification

• Binary Relevance
• Label correlation
• Label specific text representation
• Label embedding
• Others
Labels Information Based Feature Mapping

label specific features

label-specific features and local pairwise label correlation

Multi-label text classification

• Binary Relevance
• Label correlation
• Label specific text representation
• Label embedding
• Others
Label Embedding with Graph

Label space

sample1  1, 3, 5
sample2  2, 4
sample3  1, 2
sample4  4, 5

Label graph

[2 3]
[1 4 5]

Multi-label text classification

• Binary Relevance
• Label correlation
• Label specific text representation
• Label embedding

• Others
  • Classifier Chain
  • Code embedding
CNN-RNN Model

RNN Model

(a) PCC

(b) RNN

(c) RNN$^m$

(d) EncDec

label Decomposition

- Fix ‘There are some combination class labels which are associated with records less frequently than others in training datasets.’

![Table]

<table>
<thead>
<tr>
<th>Physical Records</th>
<th>Disease Labels (Combination)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>{A, B, C}</td>
</tr>
</tbody>
</table>

![Table]

<table>
<thead>
<tr>
<th>Physical Records</th>
<th>Disease Decompose Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>{A, B}</td>
</tr>
<tr>
<td>R</td>
<td>{A, C}</td>
</tr>
</tbody>
</table>

Hierarchical Embedding

\[
\min_{U,V} J(U, V) = \sum_{l=1}^{L} \sum_{i \in S} h(y_{il}(x_i U V_l^T)) + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2)
\]

**Algorithm 1: MLC-HMF \((\mathcal{X}, \mathcal{Y}, k, \mathcal{T}, h)\).**

**input**: Data Matrix: \(\mathcal{X}\), Label Matrix: \(\mathcal{Y}\), Size of Reduced Dimension Space: \(k\), Threshold: \(\mathcal{T}\), Depth of the Hierarchy: \(h\)

**output**: Tree with Mapping \(U\) and Label Feature Matrix \(V\) at Each Node

Divide \(\mathcal{X}\) into \(\mathcal{X}^1\) and \(\mathcal{X}^2\) using \textit{kmeans} clustering

\[\text{for } i \in \{1,2\} \text{ do}\]

\[\text{if } |\mathcal{X}^i| \text{ is small or depth is exceed } h \text{ then}
\]

\[\text{Let its corresponding node as leaf node}
\]

\[\text{return}\]

\[\text{Learn the mapping } U \text{ and label feature matrix } V \text{ for } \mathcal{X}^i \text{ using Eq. (4).}\]

\[\text{Let } \hat{\mathcal{X}} \subseteq \mathcal{X}^i \text{ is the set of instances whose hamming loss is less than the threshold } \mathcal{T} \text{ and } \hat{\mathcal{Y}} \text{ is their corresponding label matrix}\]

\[\text{Maintain } U, V \text{ and } \hat{\mathcal{X}} \text{ at the current node}\]

\[\text{MLC-HMF } (\mathcal{X}^i \setminus \hat{\mathcal{X}}, \mathcal{Y}_i \setminus \hat{\mathcal{Y}}, k, \mathcal{T}, h)\]

\[\text{end}\]

Attention

Figure 2. Model Architecture.

Attention

\[ a_{i,j} = \sum_{k=1}^{d} u_{i,k} h_{j,k} \]

\[ p_i = \text{sigmoid}(\max_{j=1,2,\ldots,m} (a_{i,j})) \]

\[ \tilde{a}_{i,j} = \frac{\exp(a_{i,j})}{\sum_{j=1}^{m}(\exp(a_{i,j}))} \]

\[ \tilde{u}_i = \sum_{j=1}^{m} \tilde{a}_{i,j} * h_j \]

\[ s_i = \sum_{k=1}^{d} w_{i,k} \tilde{u}_{i,k} \]

\[ p_i = \text{sigmoid}(s_i) \]

Problems

• Performance
• Ontology (MeSH, SNOMED)
• Global and Local Label Correlation
• Cross & Multi Specialty
Thanks!