Graded Multilabel Classification: The Ordinal Case

Weiwei Cheng, Krzysztof Dembczyński, and Eyke Hüllermeier

Knowledge Engineering and Bioinformatics (KEBI) Lab
Philipps-Universität Marburg
Multi-Label Classification

- Given a vector $\mathbf{x} \in \mathcal{X}$ of features, the goal is to predict a set of relevant labels $L_{\mathbf{x}} \subseteq \mathcal{L}$.

Detected objects: sky, cloud, tree, grass.
Graded Multi-Label Classification
Graded Multi-Label Classification

- Shooting
- Racing
- Fighting
- Role-playing
Graded Multi-Label Classification

- **Shooting**
  - ★★★★
  - completely

- **Racing**
  - ★★★★
  - almost

- **Fighting**
  - ★★★★
  - somewhat

- **Role-playing**
  - ★★★★
  - not at all
• Instance $x \in \mathcal{X}$ can belong to each class $\lambda \in \mathcal{L}$ to a certain degree $\rightarrow$ idea of graded class membership in the spirit of fuzzy set theory.

• Set $L_x$ of relevant labels is now a fuzzy subset of $\mathcal{L}$ with graded membership degrees in $M = [0, 1]$ (instead of $\{0, 1\}$).

• A graded multilabel classifier is a mapping $\mathcal{X} \rightarrow \mathcal{F}(\mathcal{L})$, where $\mathcal{F}(\mathcal{L})$ is a class of fuzzy subsets of $\mathcal{L}$.

• Often, an ordinal scale of membership degrees is convenient, i.e. $M = \{m_0, m_1, \ldots, m_k\}$ with

$$0 = m_0 < m_1 < \ldots < m_k = 1$$
Collaborative Filtering

- Connection between GMLC and Collaborative Filtering.
• For a given incomplete matrix \( Y \) of ordinal rates, the goal is to find matrix \( U \) and \( M \),

\[
\hat{Y} = UM,
\]

that generalizes well over missing elements of \( Y \) with respect to a specific loss function \( L(Y, \hat{Y}) \) to be minimized.

• \( U \) can be treated as features, and \( M \) as models.

• **GMLC**: \( U \) and \( Y \) is given; the goal is to find \( M \) that for new \( U' \) generalizes well to predict \( Y' \).
GMLC – Reduction

• **Reduction**: Transform complex learning problems into simpler, core problems.

• **Assumption**: Good performance on the core problems should imply good performance on the complex problem.

• **Reduction of GMLC**:

  GMLC \(\rightarrow\) Ordinal Classification

  GMLC \(\rightarrow\) Multi-Label Classification
Vertical vs. Horizontal Reduction

- **Vertical**: $L_x$ can be represented vertically, e.g., $L_x(\lambda_2) = m_1$.
- **Horizontal**: $L_x$ can be represented horizontally in terms of its **level-cuts**, e.g., \([L_x]_{m_2} = \{\lambda_1, \lambda_4, \lambda_5\} \).
Train one ordinal classifier,

\[ h_i : \mathcal{X} \rightarrow M, \quad x \mapsto L_x(\lambda_i) \in M, \]

for each label \( \lambda_i \).

\( h_i \) is solving an ordinal classification problem.

Overall, we are solving \(|\mathcal{L}|\) such problems.

The simplest approach is graded relevance.

Question: Can we solve the problem for each label independently?
• Train **one** multi-label classifier,

\[ h^{(\alpha)} : \mathcal{X} \rightarrow 2^\mathcal{L}, \quad x \mapsto [L_x]_\alpha \in 2^\mathcal{L}, \]

for each level \( \alpha \in \{m_1, m_2, \ldots, m_k\} \).

• Overall, we are solving \( k \) standard multilabel classification problems.

• **Question**: Can we **solve** the problem for each \( \alpha \)-cut **independently**?
To **reconstruct** the fuzzy subset from the horizontal reduction, one has to perform:

\[ L_x(\lambda) = \max \{ m_i \in M \mid \lambda \in [L_x]_{m_i} \} . \]

This implies that the predictions should be **consistent** in the sense that

\[ h^{(m_j)}(x) \leq h^{(m_{j-1})}(x) \]

Satisfying this **monotonicity** property is a **non-trivial** problem.
• Vertical reduction leads to ordinal classification.
• Horizontal reduction leads to multi-label classification.
• Both, ordinal classification and multi-label classification, can be reduced to binary classification.
• GMLC can be reduced to binary classification.
• What is a desired loss function for GMLC?

• GMLC loss functions in the reduction framework:
  • Ordinal classification loss functions.
  • Multilabel classification loss functions.
• Standard 0/1 loss:

\[ \ell_{0/1}(L_x(\lambda), h(x)(\lambda)) = \left[ L_x(\lambda) \neq h(x)(\lambda) \right] \]

• Absolute error:

\[ \ell_{AE}(L_x(\lambda), h(x)(\lambda)) = |L_x(\lambda) - h(x)(\lambda)| \]

• Rank loss (C-index):

\[
\ell_{rank}(L_x(\lambda), L_{x'}(\lambda), h(x)(\lambda), h(x')(\lambda)) = \\
(L_x(\lambda) - L_{x'}(\lambda)) \times (h(x')(\lambda) - h(x)(\lambda))
\]
Loss Functions for Multi-Label Classification

- **Hamming loss:**

  \[ L_H(L_x, h(x)) = \frac{1}{|L|} \sum_{i=1}^{|L|} \left[ L_x(\lambda_i) \neq h(x)(\lambda_i) \right] \]

- **Rank loss:**

  \[ L_{rank}(L_x, h(x)) = \sum_{i<j} (L_x(\lambda_i) - L_x(\lambda_j)) \times (h(x)(\lambda_j) - h(x)(\lambda_i)) \]

- **Jaccard distance:**

  \[ L_J(L_x, h(x)) = \frac{h(x) \cap L_x}{h(x) \cup L_x} \]
• Horizontal and vertical decomposition of a loss function can be equivalent:

\[
E_{HAE}(L_x, h(x)) = \frac{1}{k} \sum_{i=1}^{k} L_H([L_x]_{mi}, h^{(mi)}(x))
\]

\[
= \frac{1}{|\mathcal{L}|} \sum_{i=1}^{|\mathcal{L}|} \ell_{AE}(L_x(\lambda), h(x)(\lambda))
\]
• In general, however, there does not exist an aggregation operator $A$ such that:

$$A\left(\left\{\ell \left( h(x)(\lambda_i), L_x(\lambda_i)\right)\right\}_{i=1}^{|\mathcal{L}|}\right) = A\left(\left\{ L \left( h^{(m_i)}(x), [L_x]_{m_i}\right)\right\}_{i=1}^k\right).$$

• **Conclusion**: A choice of the loss function may imply the type of reduction.
The risk minimizer of $E_{HAE}(L_x, h(x))$ is a marginal median:

$$h^*(x) = \arg\min_h \mathbb{E}_Y |x| E_{HAE}(L_x, h(x)) = (\text{Med}(L_x(\lambda_1)), \text{Med}(L_x(\lambda_2)), \ldots, \text{Med}(L_x(\lambda_{|L|})))$$

**Question:** What would we like to estimate in GMLC?
Experiment

Showing the usefulness of the graded setting:

- We provide empirical evidence showing that labeling on graded scales offers useful extra information (binary learning VS. graded learning)
- We claim that training a learner on graded data can be useful even if only a binary prediction is actually requested.

<table>
<thead>
<tr>
<th>graded learning</th>
<th>binary test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ ★ ★</td>
<td>YES/NO</td>
</tr>
<tr>
<td>binary learning</td>
<td>YES/NO</td>
</tr>
<tr>
<td>YES/NO</td>
<td></td>
</tr>
</tbody>
</table>
BeLa-E data set (Abele & Stief, 2004):
- Degrees of importance of the future job’s different properties provided by grad students, e.g., reputation, job security, income, etc.
- Degrees are given on an ordinal scale from 5 to 1.
- 1930 instances, 50 attributes (48 job properties, 2 for sex and age).

Binarization (mimicking a person forced to decide):

1. non-relevant
2. flip a coin
3. relevant
Design of the experiment:

- A subset of features is randomly chosen as labels.
- Binary learning: the whole data is binarized.
- Graded learning: only predictions and test data are binarized.
- 10-fold cross validation with 50 randomly generated problems.
Table: Performance (mean and standard error) in the case of $m = 5$ labels (above) and $m = 10$ labels (below).

<table>
<thead>
<tr>
<th></th>
<th>BR-LR</th>
<th>BR-10NN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>binary</td>
<td>graded</td>
</tr>
<tr>
<td>Hamming/AE loss</td>
<td>0.210±0.029</td>
<td>0.186±0.031</td>
</tr>
<tr>
<td>rank loss</td>
<td>0.146±0.041</td>
<td>0.141±0.038</td>
</tr>
<tr>
<td>C-index</td>
<td>0.171±0.045</td>
<td>0.163±0.049</td>
</tr>
</tbody>
</table>

|                   | BR-LR   | BR-10NN  |
|                   | binary  | graded   | binary  | graded   |
| Hamming/AE loss   | 0.207±0.017 | 0.187±0.018 | 0.230±0.018 | 0.217±0.018 |
| rank loss         | 0.145±0.025 | 0.136±0.019 | 0.225±0.040 | 0.154±0.020 |
| C-index           | 0.175±0.011 | 0.154±0.016 | 0.237±0.011 | 0.171±0.016 |

- Graded training shows significant advantage over binary training.
• We proposed graded multilabel classification (GMLC) as an extension of conventional multilabel classification, since label relevance is often a matter of degree.
• We proposed two meta-techniques for GMLC, vertical and horizontal reduction.
• We proposed extensions of MLC loss functions and studied their usability with the two reduction schemes.
• We provided empirical evidence for the usefulness of learning from graded multilabel data.