Graded Multilabel Classification: The Ordinal Case

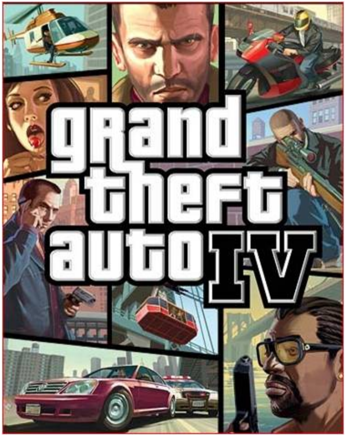
Weiwei Cheng, Krzysztof Dembczyński, and Eyke Hüllermeier

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

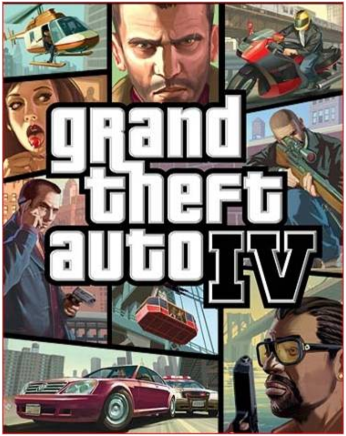
- Given a vector $\boldsymbol{x} \in \mathcal{X}$ of features, the goal is to predict a **set** of relevant labels $L_{\boldsymbol{x}} \subseteq \mathcal{L}$.

Detected objects: sky, cloud, tree, grass.

Graded Multi-Label Classification



Graded Multi-Label Classification

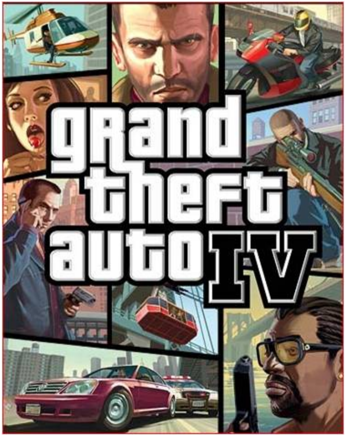


Shooting

Racing

Fighting

Role-playing



Shooting

★★★ *completely*

Racing

★★★ *almost*

Fighting

★★★ *somewhat*

Role-playing

★★★ *not at all*

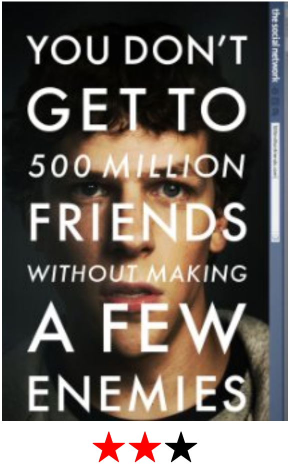
Graded Multi-Label Classification

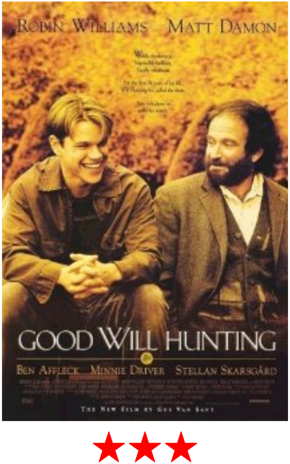
- Instance $\boldsymbol{x} \in \mathcal{X}$ can belong to each class $\lambda \in \mathcal{L}$ **to a certain degree** $\longrightarrow$ idea of graded class membership in the spirit of **fuzzy set theory**.
- Set $L_{\boldsymbol{x}}$ of relevant labels is now a **fuzzy subset** of $\mathcal{L}$ with **graded** membership degrees in $M = [0, 1]$ (instead of $\{0, 1\}$).
- A **graded** multilabel classifier is a mapping $\mathcal{X} \to \mathcal{F}(\mathcal{L})$, 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 $\mathbf{Y}$ of ordinal rates, the goal is to find matrix $\mathbf{U}$ and $\mathbf{M}$,

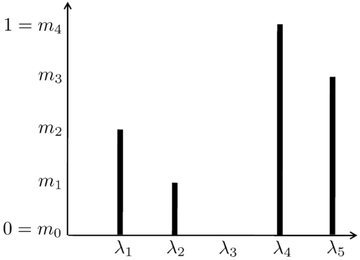
$$\hat{\mathbf{Y}} = \mathbf{U}\mathbf{M},$$

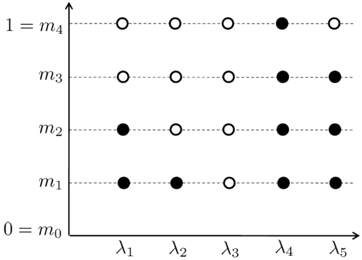
  that generalizes well over missing elements of $\mathbf{Y}$ with respect to a specific loss function $L(\mathbf{Y}, \hat{\mathbf{Y}})$ to be minimized.
- $\mathbf{U}$ can be treated as features, and $\mathbf{M}$ as models.
- **GMLC**: $\mathbf{U}$ and $\mathbf{Y}$ is given; the goal is to find $\mathbf{M}$ that for new $\mathbf{U}'$ generalizes well to predict $\mathbf{Y}'$.

- **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**:

$$\begin{array}{lcl} \text{GMLC} & \longrightarrow & \text{Ordinal Classification} \\ \text{GMLC} & \longrightarrow & \text{Multi-Label Classification} \end{array}$$





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

- Train **one** ordinal classifier,

$$h_i : \mathcal{X} \to M, \quad \boldsymbol{x} \mapsto L_{\boldsymbol{x}}(\lambda_i) \in M,$$

  for **each** label λ_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**?

Horizontal Reduction

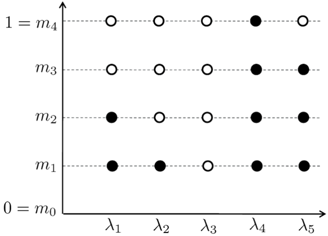
- Train **one** multi-label classifier,

$$\boldsymbol{h}^{(\alpha)} : \mathcal{X} \to 2^{\mathcal{L}}, \quad \boldsymbol{x} \mapsto [L_{\boldsymbol{x}}]_\alpha \in 2^{\mathcal{L}},$$

  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 α-cut **independently**?

- To **reconstruct** the fuzzy subset from the horizontal reduction, one has to perform:

$$L_{\boldsymbol{x}}(\lambda) = \max\{m_i \in M \mid \lambda \in [L_{\boldsymbol{x}}]_{m_i}\}.$$



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

$$\boldsymbol{h}^{(m_j)}(\boldsymbol{x}) \leq \boldsymbol{h}^{(m_{j-1})}(\boldsymbol{x})$$

- Satisfying this **monotonicity** property is a **non-trivial** problem.

Reduction

- 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.

Loss Functions for Ordinal classification

- Standard 0/1 loss:

$$\ell_{0/1}(L_{\boldsymbol{x}}(\lambda), \boldsymbol{h}(\boldsymbol{x})(\lambda)) = [\![L_{\boldsymbol{x}}(\lambda) \neq \boldsymbol{h}(\boldsymbol{x})(\lambda)]\!]$$

- Absolute error:

$$\ell_{AE}(L_{\boldsymbol{x}}(\lambda), \boldsymbol{h}(\boldsymbol{x})(\lambda)) = |L_{\boldsymbol{x}}(\lambda) - \boldsymbol{h}(\boldsymbol{x})(\lambda))|$$

- Rank loss (C-index):

$$\begin{aligned}\ell_{rank}(L_{\boldsymbol{x}}(\lambda), L_{\boldsymbol{x}'}(\lambda), \boldsymbol{h}(\boldsymbol{x})(\lambda), \boldsymbol{h}(\boldsymbol{x}')(\lambda)) = \\ (L_{\boldsymbol{x}}(\lambda) - L_{\boldsymbol{x}'}(\lambda)) \times (\boldsymbol{h}(\boldsymbol{x}')(\lambda) - \boldsymbol{h}(\boldsymbol{x})(\lambda))\end{aligned}$$

- Hamming loss:

$$L_H(L_{\boldsymbol{x}}, \boldsymbol{h}(\boldsymbol{x})) = \frac{1}{|\mathcal{L}|} \sum_{i=1}^{|\mathcal{L}|} [\![L_{\boldsymbol{x}}(\lambda_i) \neq \boldsymbol{h}(x)(\lambda_i)]\!]$$

- Rank loss:

$$L_{rank}(L_{\boldsymbol{x}}, \boldsymbol{h}(\boldsymbol{x})) = \sum_{i<j} (L_{\boldsymbol{x}}(\lambda_i) - L_{\boldsymbol{x}}(\lambda_j)) \times (\boldsymbol{h}(\boldsymbol{x})(\lambda_j) - \boldsymbol{h}(\boldsymbol{x})(\lambda_i))$$

- Jaccard distance:

$$L_J(L_{\boldsymbol{x}}, \boldsymbol{h}(\boldsymbol{x})) = \frac{\boldsymbol{h}(\boldsymbol{x}) \cap L_{\boldsymbol{x}}}{\boldsymbol{h}(\boldsymbol{x}) \cup L_{\boldsymbol{x}}}$$

- Horizontal and vertical decomposition of a loss function can be equivalent:

$$
\begin{aligned}
E_{HAE}(L_{\boldsymbol{x}}, \boldsymbol{h}(\boldsymbol{x})) &= \frac{1}{k}\sum_{i=1}^{k} L_H([L_{\boldsymbol{x}}]_{m_i}, \boldsymbol{h}^{(m_i)}(\boldsymbol{x})) \\
&= \frac{1}{|\mathcal{L}|}\sum_{i=1}^{|\mathcal{L}|} \ell_{AE}(L_{\boldsymbol{x}}(\lambda), \boldsymbol{h}(\boldsymbol{x})(\lambda))
\end{aligned}
$$

- In general, however, there does not exist an aggregation operator A such that:

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

- **Conclusion**: A choice of the loss function may imply the type of reduction.

Desired Loss Function?

- The risk minimizer of $E_{HAE}(L_{\boldsymbol{x}}, \boldsymbol{h}(\boldsymbol{x}))$ is a **marginal median**:

$$\begin{aligned} \boldsymbol{h}^*(\boldsymbol{x}) &= \arg\min_{\boldsymbol{h}} \mathbb{E}_{\mathbf{Y}\,|\,\boldsymbol{x}} E_{HAE}(L_{\boldsymbol{x}}, \boldsymbol{h}(\boldsymbol{x})) \\ &= (Med(L_{\boldsymbol{x}}(\lambda_1)), Med(L_{\boldsymbol{x}}(\lambda_2)), \ldots, Med(L_{\boldsymbol{x}}(\lambda_{|\mathcal{L}|}))) \end{aligned}$$

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

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.

graded learning	
★ ★ ★	binary test data
binary learning	YES/NO
YES/NO	

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 2	3	4 5
↓	↓	↓
non-relevant	flip a coin	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.

Experiment

Table: Performance (mean and standard error) in the case of $m = 5$ labels (above) and $m = 10$ labels (below).

	BR-LR		BR-10NN	
	binary	graded	binary	graded
Hamming/AE loss	0.210±0.029	0.186±0.031	0.220±0.051	0.213±0.052
rank loss	0.146±0.041	0.141±0.038	0.328±0.115	0.310±0.104
C-index	0.171±0.045	0.163±0.049	0.381±0.089	0.361±0.080
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.

Take-Away Message

- 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.