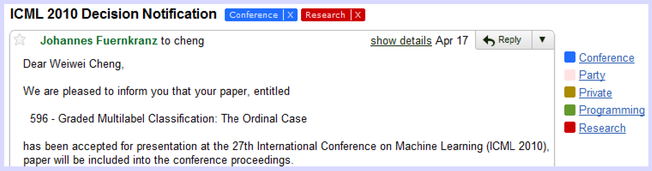
GRADED MULTILABEL CLASSIFICATION: THE ORDINAL CASE

Weiwei Cheng, Krzysztof Dembczyński, and Eyke Hüllermeier
Knowledge Engineering & Bioinformatics Lab
Department of Mathematics and Computer Science
University of Marburg, Germany

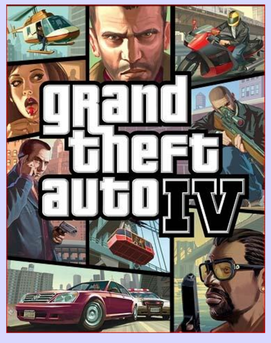
Graded Multilabel Classification

Conventional Setting



Each instance belongs to a subset of the classes.

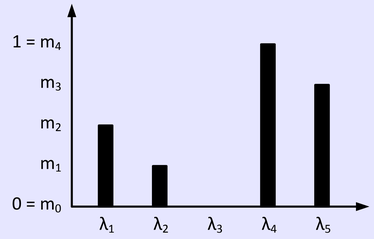
Graded Setting



Shooting ★★★ *completely*
Racing ★★☆ *almost*
Fighting ★☆☆ *somewhat*
Role-playing ☆☆☆ *not at all*

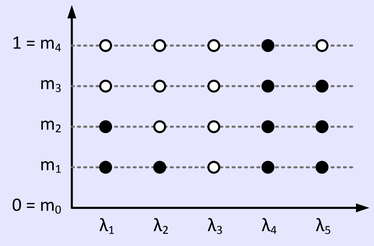
– An instance $x \in \mathcal{X}$ can belong to each class $\lambda \in \mathcal{L}$ to a certain degree; the set of relevant labels is a fuzzy subset of the label set;
– A graded multilabel classifier is a mapping $\mathcal{X} \longrightarrow M^{\mathcal{L}}$, where $M \subseteq [0,1]$ (instead of $M = \{0,1\}$) is the set of graded membership degrees;
– An ordinal scale of membership degrees is often preferred, i.e., $M = \{m_0, m_1, \ldots, m_k\}$ with $0 = m_0 < m_1 < \ldots < m_k = 1$.

Vertical Reduction



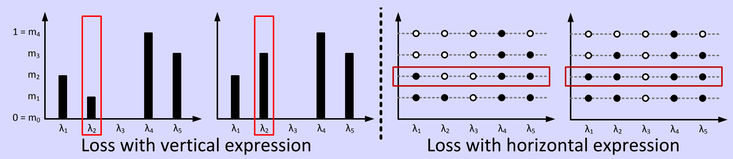
– One classifier $h_i : \mathcal{X} \longrightarrow M$ is induced for each label λ_i;
– h_i is solving an ordinal classification problem;
– To exploit potential label dependencies, the problems should be solved simultaneously, not independently.

Horizontal Reduction



– A fuzzy subset of labels $L_{\mathbf{x}}$ can be represented "horizontally" by its level-cuts $[L_{\mathbf{x}}]_\alpha$ (e.g., $[L_{\mathbf{x}}]_{m_2} = \{\lambda_1, \lambda_4, \lambda_5\}$);
– $L_{\mathbf{x}}$ can be recovered by $L_{\mathbf{x}}(\lambda) = \max\{m_i \in M \mid \lambda \in [L_{\mathbf{x}}]_{m_i}\}$;
– For each level $\alpha \in \{m_1, m_2, \ldots, m_k\}$, a mapping $h^{(\alpha)} : \mathcal{X} \longrightarrow 2^{\mathcal{L}}, \mathbf{x} \mapsto [\mathcal{L}]_\alpha$ is learned;
– Overall, we are solving k conventional multilabel classification problems;
– Consistency condition: $h^{(m_i)}(\mathbf{x}) = 1 \;\Rightarrow\; h^{(m_{i-1})}(\mathbf{x}) = 1$.

Loss Functions



1. How to extend existing loss functions to the graded setting? 2. Given a graded loss function, what loss should be minimized for the reduced sub-problems?

– For example, Hamming loss $E_H(h(\mathbf{x}), L_{\mathbf{x}}) = \frac{1}{|\mathcal{L}|}|h(\mathbf{x}) \Delta L_{\mathbf{x}}|$ can be extended to a horizontal representation $E^*_H(h(\mathbf{x}), L_{\mathbf{x}}) = \frac{1}{k|\mathcal{L}|}\sum_{i=1}^{k}|[h(\mathbf{x})]_{m_i} \Delta [L_{\mathbf{x}}]_{m_i}|$ and an equivalent vertical one $E^*_H(h(\mathbf{x}), L_{\mathbf{x}}) = \frac{1}{k|\mathcal{L}|}\sum_{i=1}^{|\mathcal{L}|} \mathrm{AE}(h(\mathbf{x})(\lambda_i), L_{\mathbf{x}}(\lambda_i))$;
– However, there are other losses that only exhibit a horizontal or a vertical representation, but not both.

Experiments and Conclusions

– A dataset from social psychology;
– Two settings
  - *Binary learning*: the whole data is binarized;
  - *Graded learning*: predictions and test data are binarized.
• Both, vertical and horizontal, decompositions work well;
• Training a learner on graded data can be useful even if only a *binary prediction* is requested.

ACKNOWLEDGEMENT: Weiwei Cheng was financially supported by the ICML 2010 Student Travel Scholarship Program.
Poster presented at ICML 2010, 27th International Conference on Machine Learning, Haifa, Israel, June 2010