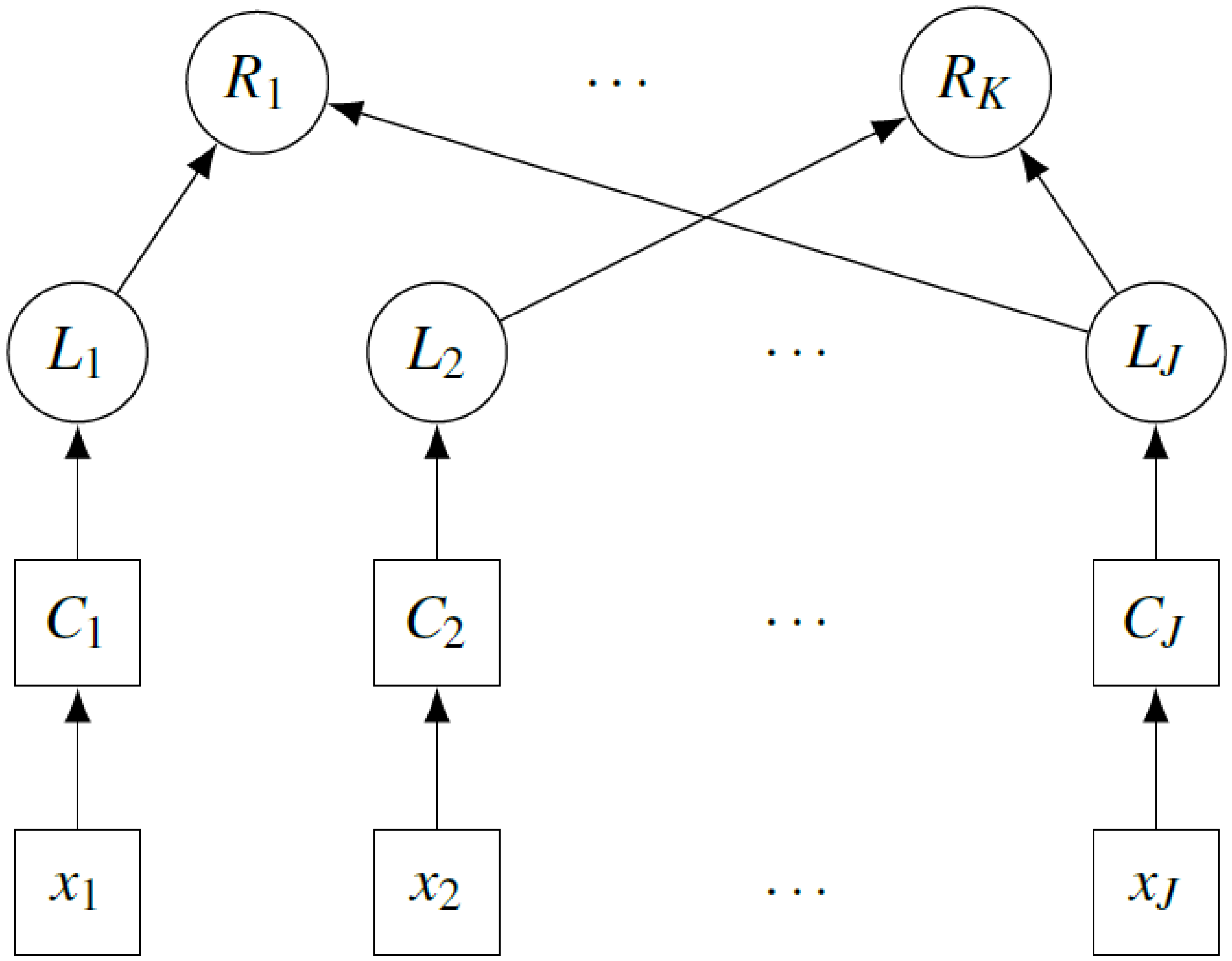
technische universität dortmund

pRSL: Interpretable Multi-label Stacking by Learning Probabilistic Rules

Michael Kirchhof, Lena Schmid, Christopher Reining, Michael ten Hompel, Markus Pauly

TU Dortmund University, Department of Statistics



Motivation

P(overhead work) = 0.5

P(standing) = 0.95

P(hands high) = 0.7

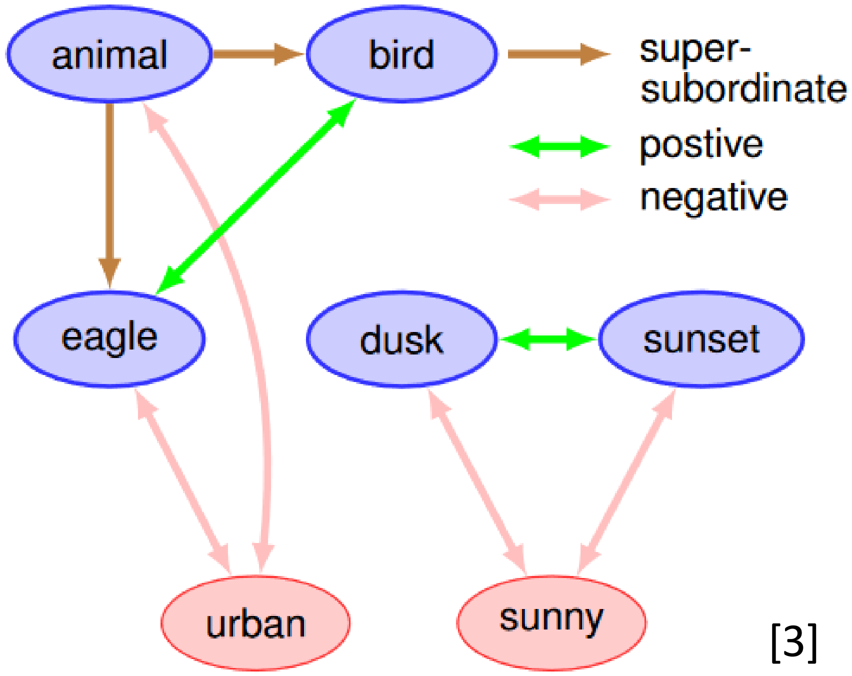
P(overhead work | other beliefs) = ?

Related Work

- Goal: Model multi-label distribution

$$P(L_1 = \ell_1, \ldots, L_J = \ell_J | \boldsymbol{x}) = P(\boldsymbol{L} = \boldsymbol{\ell} | \boldsymbol{x})$$

- Attribute-class approaches
- Knowledge graphs
- Bayesian networks
- Probabilistic rules

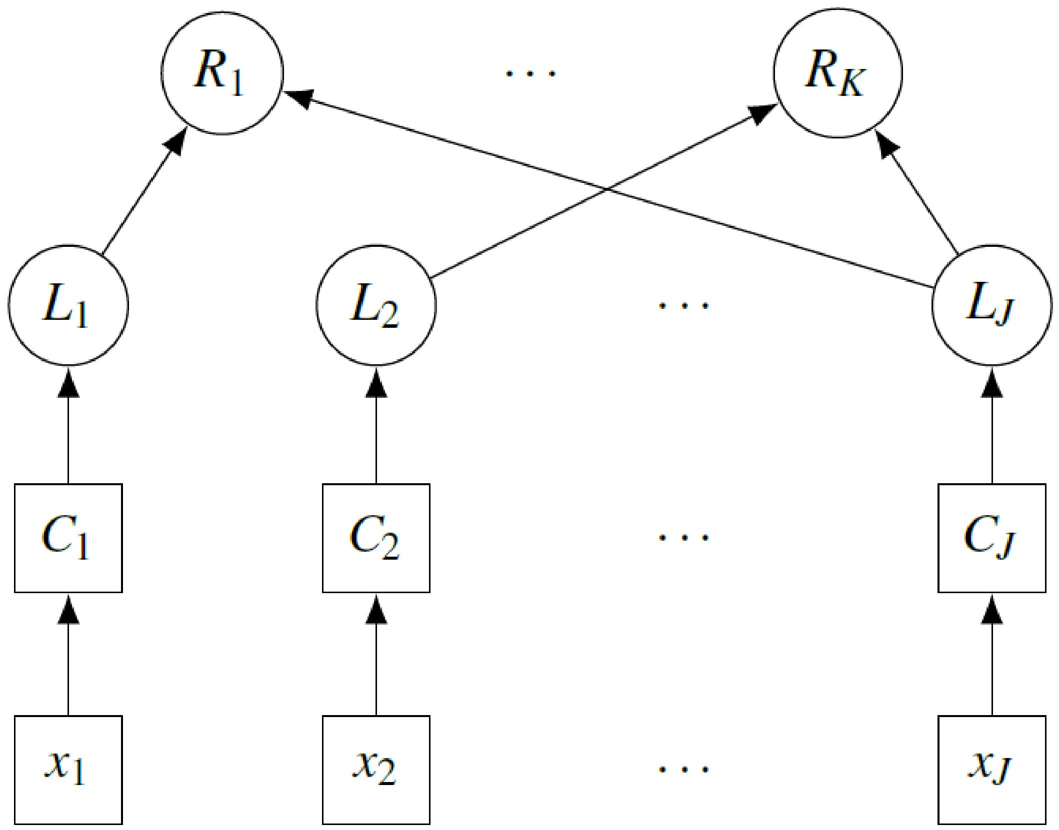


[3]

pRSL

- Classifiers outputs are prior beliefs on the labels
- Labels are connected via probabilistic logical rules
- Conditioning on rules gives a-posteriori beliefs:

$$
\begin{aligned}
&P(\boldsymbol{L}=\boldsymbol{\ell}|\boldsymbol{R}=1,\boldsymbol{x}) \\
\propto\, &P(\boldsymbol{L}=\boldsymbol{\ell},\boldsymbol{R}=1|\boldsymbol{x}) \\
=\, &P(\boldsymbol{L}=\boldsymbol{\ell}|\boldsymbol{x})P(\boldsymbol{R}=1|\boldsymbol{L}=\boldsymbol{\ell})
\end{aligned}
$$



Rules

Propositional Logic

- Soften up truth tables
- Arbitrary inter-label relations

ℓ_1	ℓ_2	$P(R_k = 1 \mid L = \ell)$
a_1	b_1	0.8
a_1	b_2	0.8
a_2	b_1	0.8
a_2	b_2	0.2

R_k: $a_2 \rightarrow b_1$ (p = 0.8)

Multicategorical Noisy-or

- Parametrized with inhibition probabilities $q^k_{j\ell_j}$
- Extended to multicategorical case

$$P(R_k = 0 | \boldsymbol{L} = \boldsymbol{\ell}) = \prod_{j=1}^{J} q^k_{j\ell_j}$$

Example

- Detect overhead work
- Sensors:
  - L_1: Camera = {movement, normal work, overhead work}
  - L_2: Shoes = {gait, stand}
  - L_3: Height = {high, center, low}
- Rules:
  - R_1: $s \wedge h \rightarrow o$ (p = 0.8)
  - R_2: $n \rightarrow c \vee l$ (p = 0.9)
  - R_3: $g \leftrightarrow m$ (p = 1)

Example Calculations

- Given $P(L_1|x_1) = (0.1, 0.4, 0.5)$, $P(L_2|x_2) = (0.05, 0.95)$, $P(L_3|x_3) = (0.5, 0.3, 0.2)$, what is $P(L_1 = o|R = 1, x)$?

ℓ_1	ℓ_2	ℓ_3	$P(\boldsymbol{L}=\boldsymbol{\ell}\|x)$	$P(R_1=1\|\boldsymbol{L}=\boldsymbol{\ell})$	$P(R_2=1\|\boldsymbol{L}=\boldsymbol{\ell})$	$P(R_3=1\|\boldsymbol{L}=\boldsymbol{\ell})$	$P(\boldsymbol{L}=\boldsymbol{\ell}\|R=1,x)$
w	s	h	$0.1 \cdot 0.95 \cdot 0.5$	0.2	0.9	0	0
n	s	h	$0.4 \cdot 0.95 \cdot 0.5$	0.2	0.1	1	0.0078
o	s	h	$0.5 \cdot 0.95 \cdot 0.5$	0.8	0.9	1	0.3517
w	g	l	$0.1 \cdot 0.05 \cdot 0.2$	0.8	0.9	1	0.0015
n	s	c	$0.4 \cdot 0.95 \cdot 0.3$	0.8	0.9	1	0.1688
o	s	c	$0.5 \cdot 0.95 \cdot 0.3$	0.8	0.9	1	0.2110
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

- $P(L_1|R = 1, x) = (0.01, 0.05, 0.94)$, $P(L_2|R = 1, x) = (0.01, 0.99)$, $P(L_3|R = 1, x) = (0.48, 0.31, 0.21)$
- (o, s, h) is the most likely combination

Implementation Details

- Rules parametrized as multi-categorical noisy-or
- Loopy belief propagation allows approximate inference in $O(JK)$
- Learn rules in $O(JK^2)$ by inverse likelihood trick

$$\frac{\partial}{\partial q}\log(P(\boldsymbol{L}=\boldsymbol{\ell}|\boldsymbol{R}=\boldsymbol{1},\boldsymbol{x})) = \frac{1}{P(\boldsymbol{L}=\boldsymbol{\ell}|\boldsymbol{R}=\boldsymbol{1},\boldsymbol{x})}\frac{\partial}{\partial q}P(\boldsymbol{L}=\boldsymbol{\ell}|\boldsymbol{R}=\boldsymbol{1},\boldsymbol{x})$$

- Beta Regularization

$$\log(P(\boldsymbol{L}=\boldsymbol{\ell}^*,\boldsymbol{Q}=\boldsymbol{q}|\boldsymbol{x})) = \log(P(\boldsymbol{L}=\boldsymbol{\ell}^*|\boldsymbol{Q}=\boldsymbol{q},\boldsymbol{x})) + \log(P(\boldsymbol{Q}=\boldsymbol{q})) \mathrel{\widehat{=}} \text{Loss} + \text{Regularizer}$$

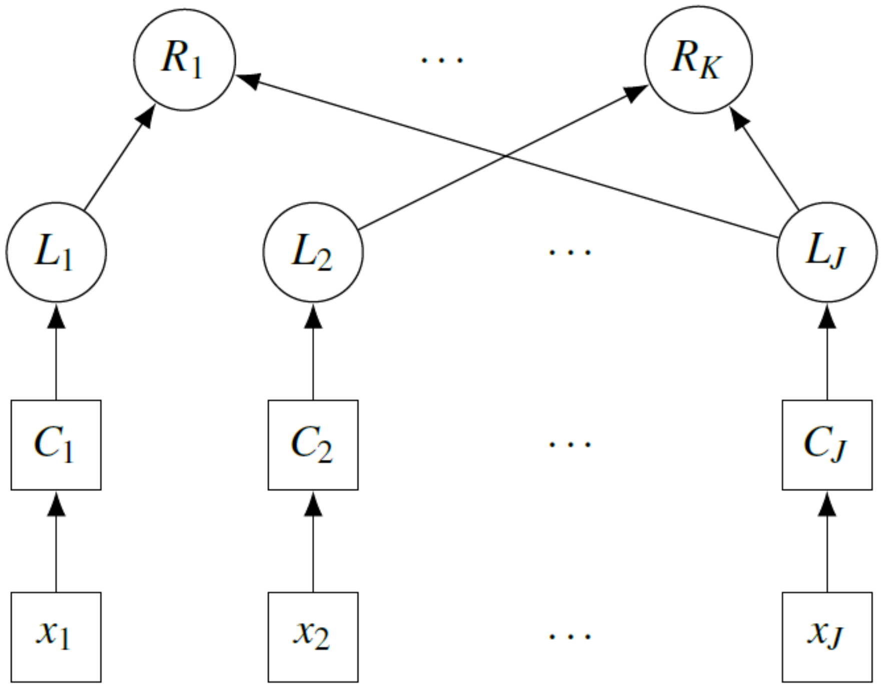
Benchmark Results

- Datasets: Emotions, Yeast, Birds, Medical, Enron, Mediamill
- Comparison Methods: 2-layer NN, MLWSE (2020), BOOMER (2020)
- Measures: Joint-label Accuracy, Hamming Loss, log-likelihood

- Each method has strengths on different datasets and metrics
- pRSL particularly good on datasets with high density
- Approximate inference methods are scalable

Summary

- pRSL builds up multi-label distribution via probabilistic rules
- All types of logical inter-label relations
- Mathematically derived learning strategies outperformed heuristics
- Applied to zero-shot human activity recognition task



$$\begin{aligned} & P(\boldsymbol{L} = \boldsymbol{\ell} | \boldsymbol{R} = 1, \boldsymbol{x}) \\ \propto & P(\boldsymbol{L} = \boldsymbol{\ell}, \boldsymbol{R} = 1 | \boldsymbol{x}) \\ = & P(\boldsymbol{L} = \boldsymbol{\ell} | \boldsymbol{x}) P(\boldsymbol{R} = 1 | \boldsymbol{L} = \boldsymbol{\ell}) \end{aligned}$$

References in Order of Appearance

1. Friedrich Niemann, Christopher Reining, Fernando Moya Rueda, Nilah Ravi Nair, Janine Anika Steffens, Gernot A Fink, and Michael ten Hompel. Lara: Creating a dataset for human activity recognition in logistics using semantic attributes. Sensors, 20(15):4083, 2020.
2. Yuval Atzmon and Gal Chechik. Probabilistic AND-OR Attribute Grouping for Zero-Shot Learning. In Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, 2018.
3. Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi–label Zero–Shot Learning with Structured Knowledge Graphs. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, pages 1576–1585, 2018.
4. Rongbo Shen, Fuhao Zou, Jingkuan Song, Kezhou Yan, and Ke Zhou. EFUI: An Ensemble Framework using Uncertain Inference for Pornographic Image Recognition. Neurocomputing, 322:166–176, 2018.
5. Eyal Krupka, Kfir Karmon, Noam Bloom, Daniel Freedman, Ilya Gurvich, Aviv Hurvitz, Ido Leichter, Yoni Smolin, Yuval Tzairi, Alon Vinnikov, et al. Toward Realistic Hands Gesture Interface: Keeping it Simple for Developers and Machines. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 1887–1898, 2017.
6. Judea Pearl. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, 1988.
7. Yuelong Xia, Ke Chen, and Yun Yang. Multi–label Classification with Weighted Classifier Selection and Stacked Ensemble. Information Sciences, 2020.
8. Michael Rapp, Eneldo Loza Mencía, Johannes Fürnkranz, Vu-Linh Nguyen, and Eyke Hüllermeier. Learning Gradient Boosted Multi–label Classification Rules. In Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2020, 2020.
9. Dembczynski, Krzysztof, Weiwei Cheng, and Eyke Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. *ICML*. 2010.
10. Kevin P Murphy. Machine Learning: A Probabilistic Perspective. MIT press, 2012.

Regularization

Hyperparameters are interpretable
and can be set automatically

Prior: $(1-S) \sim \text{Exp}(\eta)$

Reg.: $$R(\boldsymbol{q}) = \log\left(\prod_{k=1}^{K}\prod_{j=1}^{J} \eta e^{-\eta(1-s)}\right)$$

$$\propto \eta \sum_{k=1}^{K}\sum_{j=1}^{J} 1 - s(k,j)$$

Prior: $S \sim \text{Beta}(\beta_1, \beta_2)$

Reg.: $$R(\boldsymbol{q}) = \log\left(\prod_{k=1}^{K}\prod_{j=1}^{J} \frac{1}{B(\beta_1,\beta_2)} s^{\beta_1-1}(1-s)^{\beta_2-1}\right)$$

$$\propto \frac{1}{\eta}((\beta_1-1)\sum_{k=1}^{K}\sum_{j=1}^{J}\log(s(k,j)+\varepsilon)+$$

$$(\beta_2-1)\sum_{k=1}^{K}\sum_{j=1}^{J}\log(1-s(k,j)+\varepsilon)),$$

Normalization constant gives "natural"
(and optimal) regularizer strength

Inference and Learning

- Loopy belief propagation allows approximate inference in $O(JK)$
- Learn rules in $O(JK^2)$ by inverse likelihood trick