Recurrent Bayesian Classifier Chains for Exact Multi-Label Classification

Walter Gerych, Thomas Hartvigsen, Luke Buquicchio, Emmanuel Agu, Elke Rundensteiner

Multi-Label Classification

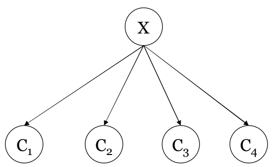
Data of the form $x, c_1, c_2, \ldots, c_L \sim (X, C_1, C_2, \ldots, C_L)$
such that ci = 1 if class i applies to x, and $c_i = 0$ otherwise.

Applicable in a variety of domains:

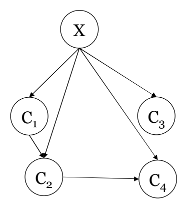
	Computer Vision	Predictive Medicine	Text Mining
X =			
$\mathbb{C}$ =	{Person, Laptop, Soda can}	{Hypertension, Arrhythmia}	{Editorial, Science, Health}

Multi-label classification method benefit by modeling label dependencies, instead of treating each class as an independent binary task:

Binary Approach



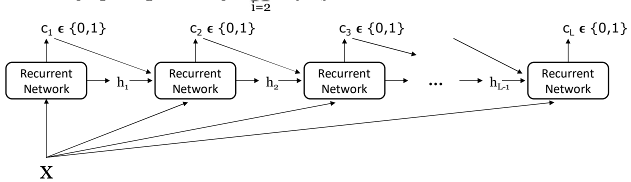
Modeling Label Dependencies



State-Of-The-Art: Recurrent Classifier Chains

Recurrent Classifier Chains (RCCs) sequentially predict classes, conditioning each prediction on the preceding classes

$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2} P(C_i|C_{<i}, X)$$



Limitations:

- Noisy conditioning: predictions conditioned on independent classes
- Error propagation: mistake made on 1 class effects all subsequent predictions
- Long term memory: early classes have diminishing effect on late classes

Our Approach: Recurrent Bayesian Classifier Chains

Recurrent Bayesian Classifier Chain (RBCC) key components:

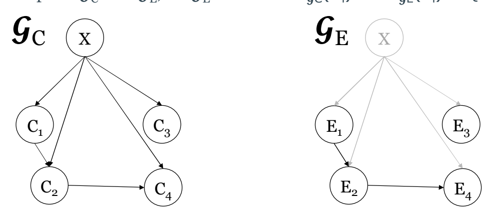
1. Infer Bayesian network of label dependencies
2. Modify RCC architecture to only use parent classes for inference

Tackles challenges by:

- Eliminating noisy conditioning
- Minimizing error propagation
- Removing need for long-term memory

Learning Label Dependency Graph

- We model class interdependencies by assuming Bayesian network structure $\mathcal{G}_C$
- Decompose joint probability as $P(C_1, C_2, \ldots, C_L|X) = \Pi\, P(C_i|\, Pa_{\mathcal{G}C}(C_i))$
- Challenge: X is typically continuous; difficult for network learning algorithms
- Solution: Replace $\mathcal{G}_C$ with $\mathcal{G}_E$, for $\mathcal{G}_E$ such that $Pa_{\mathcal{G}C}(C_i) = Pa_{\mathcal{G}E}(E_i) \cup \{X\}$



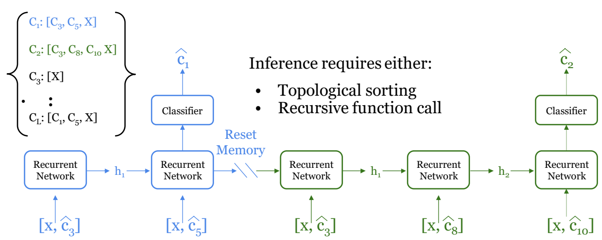
Our choice for $\mathcal{G}_E$: graph of class *errors*

$C_i = k_i(X) + E_i \Rightarrow E_i = C_i - f(x)$; k_i is found by maximizing data likelihood

- New model for joint probability: $P(C_1, C_2, \ldots, C_L|X) = \Pi\, P(C_i|\, Pa_{\mathcal{G}E}(E_i), X)$

RBCC Model

- We construct a dictionary of class dependencies from $\mathcal{G}_E$, where each key is a class and each value is the set of corresponding parents
- For each instance, the prediction of each class c_i is made using a recurrent network that steps through the value of each parent.
- The internal state of the hidden network is reset after each class prediction to avoid noisy conditioning



Inference requires either:

- Topological sorting
- Recursive function call

Experimental Evaluation

Compared Methods

- Recurrent Classier Chains (RCC)
- Topological-Sort RCC (TS-RCC)
- Order-Free RCC (OF-RCC)
- Bayesian Classifier Chains (BCC)
- Binary Decomposition (BD)

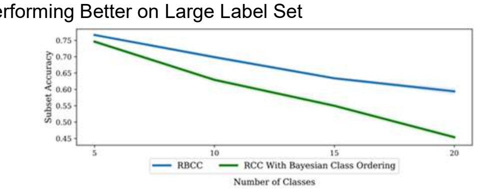
Classification Performance

Evaluation Metrics	Methods					
	RBCC (Ours)	RCC	TS-RCC	OF-RCC	BCC	BD
Subset Accuracy ↑	**0.240** ± 0.008	0.212 ± 0.002	0.192 ± 0.010	0.169 ± 0.009	0.210 ± 0.000	0.202 ± 0.002
Hamming Loss ↓	**0.186** ± 0.003	0.204 ± 0.001	0.209 ± 0.004	0.218 ± 0.004	0.199 ± 0.001	0.189 ± 0.000
Macro-F1 ↑	0.556 ± 0.008	0.526 ± 0.004	0.506 ± 0.004	**0.569** ± 0.004	0.551 ± 0.005	0.517 ± 0.008
Micro-F1 ↑	**0.670** ± 0.006	0.639 ± 0.002	0.628 ± 0.004	0.662 ± 0.004	0.653 ± 0.003	0.638 ± 0.003

Table 2: Classification results for the Yelp dataset. Bolded is best performer, underlined is second best.

- Our method outperforms all others on strictest multi-label metric, nearly always outperforms state-of-the-art on other metrics
- Additional results on 5 other datasets is available in our paper

Performing Better on Large Label Set



- The difference in performance between our method and a comparable RCC increases as the number of possible classes increases

Conclusion

In this work we:

- Identified flaws with state-of-the-art multi-label approach (RCC)
- Proposed new multi-label approach that leverages label dependence and independence to improve RCC training and inference
- Experimentally showed the practical improvement of our approach

Acknowledgements

- WPI WASH Research group
- WPI DAISY Lab
- DARPA WASH Grant #FA8750-18-2-0077

References

Nam, Jinseok, et al. "Maximizing subset accuracy with recurrent neural networks in multi-label classification." NeurIPS 2017.
Zhang, Min-Ling, et al. "Multi-label learning by exploiting label dependency." KDD 2010.
Shang-Fu Chen, et al. "Order-free RNN with visual attention for multi-label classification." AAAI 2018.
Tsoumakas , Grigorios Tsoumakas aet al. "Multi label classification: An overview." IJDWM 2007.