Classifier Chains for Multi-label Classification

Jesse Read*, Bernhard Pfahringer, Geoffrey Holmes, Eibe Frank

* currently at École Polytechnique.

2009–2019

18 Sep. 2019, ECML-PKDD, Würzburg

Outline

Introduction: Multi-Label Classification

We want a model to assign labels to input instances, e.g.,

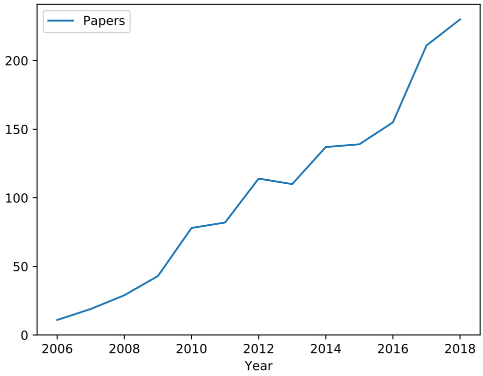
$\mathbf{x} =$ 

Given a set of labels, e.g.,

$$\mathcal{Y} = \{\texttt{beach}, \texttt{people}, \texttt{foliage}, \texttt{sunset}, \texttt{urban}\}$$

we want to predict a subset, e.g., $\{\texttt{beach}, \texttt{foliage}\} \subseteq \mathcal{Y}$ for $\mathbf{x}$.

New papers with "multi-label classification" in the title (Google Scholar), per year.

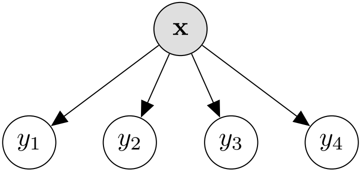


A random selection of papers citing MEKA (a multi-label learning framework):

- *[. . .] Multi-label Sentiment Classification of Health Forums*
- *Using Multi-Label Classification for Improved Question Answering*
- *Predictive Skill Based Call Routing [. . .]*
- *[. . .] Methods for Prediagnosis of Cervical Cancer*
- *[. . .] Expert Systems for Reasoning in Clinical Depressive Disorders*
- *Multi-label classification for intelligent health risk prediction*
- *Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos*
- *Spectral features for audio based vehicle and engine classification*
- *Ensemble-Based Location Tracking Using Passive RFID*
- *[. . .] big data streams analysis: The case of object trajectory prediction*
- *Multi-task network embedding*
- *Multi-Target Classification and Regression in Wineinformatics*

Binary Relevance: The Baseline

$\boldsymbol{X}$	Y_1	Y_2	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	1	0	0
$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	0	0	1
$\tilde{\mathbf{x}}$	?	?	?	?



The binary relevance method = *one binary classifier trained for each label*, i.e., independent models.

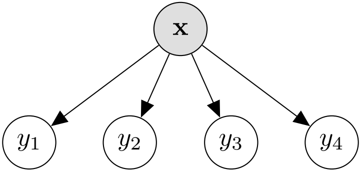
Binary Relevance: The Baseline

$\boldsymbol{X}$	Y_1
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0
$\tilde{\mathbf{x}}$	?

$\boldsymbol{X}$	Y_2
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	1
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0
$\tilde{\mathbf{x}}$	?

$\boldsymbol{X}$	Y_3
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0
$\tilde{\mathbf{x}}$	?

$\boldsymbol{X}$	Y_4
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	1
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0
$\tilde{\mathbf{x}}$	?

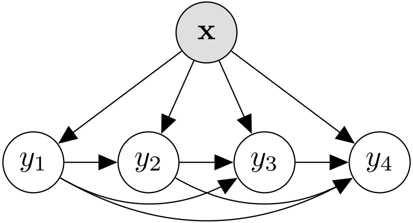


The binary relevance method = *one binary classifier trained for each label*, i.e., independent models.

Outline

Classifier Chains

A "chain" of classifiers[1]:



The output of each classifier (classification, $\in \{0, 1\}$) becomes an additional feature for all following classifiers.

- Takes into account label dependence
- Works well "off-the-shelf"
- A transformation method (*base classifier* as a hyperparameter)
- Similar running time as independent classifiers (in practice)

[1] Read et al., ECML-PKDD 2009

As a transformation (L standard binary classification problems):

$\boldsymbol{X}$	Y_1
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0
$\tilde{\mathbf{x}}$	$\widehat{y}_1$

$\boldsymbol{X}$	Y_1	Y_2
$\mathbf{x}^{(1)}$	0	1
$\mathbf{x}^{(2)}$	1	0
$\mathbf{x}^{(3)}$	0	1
$\mathbf{x}^{(4)}$	1	0
$\mathbf{x}^{(5)}$	0	0
$\tilde{\mathbf{x}}$	$\widehat{y}_1$	$\widehat{y}_2$

$\boldsymbol{X}$	Y_1	Y_2	Y_3
$\mathbf{x}^{(1)}$	0	1	1
$\mathbf{x}^{(2)}$	1	0	0
$\mathbf{x}^{(3)}$	0	1	0
$\mathbf{x}^{(4)}$	1	0	0
$\mathbf{x}^{(5)}$	0	0	0
$\tilde{\mathbf{x}}$	$\widehat{y}_1$	$\widehat{y}_2$	$\widehat{y}_3$

$\boldsymbol{X}$	Y_1	Y_3	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	1	0	0
$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	0	0	1
$\tilde{\mathbf{x}}$	$\widehat{y}_1$	$\widehat{y}_2$	$\widehat{y}_3$	$\widehat{y}_4$

where $\mathbf{x}^{(i)}$ is the i-th training example, $\tilde{\mathbf{x}}$ is a test example, $\widehat{y}_j$ the prediction of the j-th classifier.

What about the order of the labels? – A poor order could lead to *error propagation*.

An *Ensemble* of Classifier Chains: Build many chains, each with a random order, and combine the predictions.

- Works well (robust against error propagation)
- Still was tractable (on the datasets at the time)

What about the order of the labels? – A poor order could lead to *error propagation*.

An *Ensemble* of Classifier Chains: Build many chains, each with a random order, and combine the predictions.

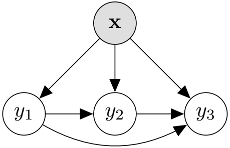
- Works well (robust against error propagation)
- Still was tractable (on the datasets at the time)

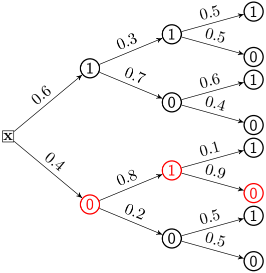
> But *how* does it work? What is it optimising?
> Can we get a better chain? ...

Outline

View 1: Probabilistic Classifier Chains

A probabilistic interpretation[2]:





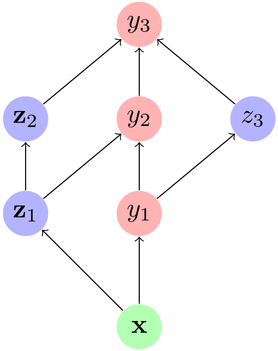
$$\hat{\mathbf{y}} = \underset{\mathbf{y}\in\{0,1\}^L}{\operatorname{argmax}}\, P(y_1|\mathbf{x}) \prod_{j=2}^{L} P(y_j|\mathbf{x}, y_1, \ldots, y_{j-1})$$

- It's a MAP estimate, optimising subset 0/1 loss
- Inference becomes a search
  - standard classifier chain = greedy search.
  - exhaustive search: try all 2^L combinations/paths

[2]Dembczyński, Cheng, and Hüllermeier, ICML 2010; and followup work

View 2: Classifier Chains as a Deep Network

Classifier chains as a neural network[3] (with delay nodes z):

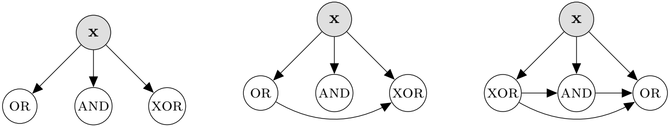


- It's deep in the label space!
- "Hidden nodes" come for free
- labels = a higher-level feature representation

[3] Read and Hollmén, IDA 2014; Cisse, Al-Shedivat, and Bengio, ICML 2016; and others

How to Order/Structure the Chain

1. Random (ensembles)? Effective but large/uninteresting
2. Existing hierarchy? May not be as useful as you think
   - Nice to look at, but no guarantee it suits given method/metric
   - We[4] won Kaggle LSHTC14 (large scale *hierarchical* text classification), *ignoring the hierarchy*
3. Based on label dependence?[5] It depends, consider:



   Only one works with 'default parameters' (linear SVM, greedy inference).
4. Search the label-structure space[6]: Slow!, but
   - Many local maxima that are easy to reach – i.e., it can work!
   - Don't need to discard suboptimal models – dynamic order

[4] Puurula, Read, and Bifet 2014

[5] Zaragoza et al., IJCAI 2011; and others

[6] Kumar et al., ECML-PKDD 2012; Read, Martino, and Luengo, Pat. Rec. 2014; Gasse, U. Lyon 2017; etc.

Connection to RAkEL, etc.

It a sense, classifier chains is similar to RAkEL[7] :

$$\underset{\mathbf{y}\in\{0,1\}^L}{\operatorname{argmax}} P(y_1|\mathbf{x}) \prod_{j=2}^{L} P(y_j|\mathbf{x}, y_1, \ldots, y_{j-1}) \approx \underset{\mathbf{y}\in\mathcal{S}}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x})$$

- Classifier chains provides a mechanism to search the space
- RAkEL reduces the space itself, $\mathcal{S} \subset \{0,1\}^k$
- Both rely on ensembles

Other connections:

- Probabilistic graphical models (HMMs, CRFs, . . .)
- Neural networks (ResNets, . . .)

[7] Tsoumakas and Vlahavas, ECML-PKDD 2007 (Test of Time Award 2017)

Recent Developments

A small selection of analyses and improvements of classifier chains:

- Study of methods for search inference[8] and label ordering[9]
- A closer look at error propagation[10]
- With feature selection integrated[11]
- Class imbalance, Dynamic chains. . .
- Conditional entropy based chains[12]

---

[8] Dembczyński, Waegeman, and Hüllermeier, ECAI 2012; Mena et al., IJCAI 2015

[9] Goncalves, Plastino, and Freitas, ICTAI 2013

[10] Senge, Coz, and Hüllermeier, DAMLKD 2014

[11] Teisseyre, Neurocomp. 2017

[12] Jun et al., Neurocomp. 2019

What about Regressor Chains?

Regressor chains: direct/off-the-shelf application; but

- results are not great (vs independent classifiers)[13]
- 4 papers with "regressor chains"/regression chains in the title[14] vs 77 for "classifier chains"/chain classifiers (Google Scholar, Sep. 2019)

What happens?

[13] As surveyed by, e.g., Borchani et al., Wiley. 2015

[14] incl. Melki et al., Inf. Sci. 2017 using SVR; Read and Martino, ArXiv 2019

What about Regressor Chains?

Regressor chains: direct/off-the-shelf application; but

- results are not great (vs independent classifiers)[13]
- 4 papers with "regressor chains"/regression chains in the title[14] vs 77 for "classifier chains"/chain classifiers (Google Scholar, Sep. 2019)

What happens? A

- 'default' choice of base regression model (linear regression)
- 'default' choice of error metric (squared error)
- error propagation in $\mathbb{R}^L$

means that (compared to individual classifiers) no improvement in the best case, and potentially catastrophic otherwise.

[13] As surveyed by, e.g., Borchani et al., Wiley. 2015
[14] incl. Melki et al., Inf. Sci. 2017 using SVR; Read and Martino, ArXiv 2019

Perspectives

For classifier chains, and multi-label learning in general:

- For 'small' datasets ($L \leq 1000$?, $L \leq 10000$?):
  - Predictive performance has plateaued;
  - emphasis on interpretability of label relationships; applications
  - Open problems on evaluation (what metrics, etc.), dealing with weak labels, sparsity, online learning, ...
- For large datasets ('extreme' multi-label classification):
  - Different metrics, focus, etc; typically text
  - Intersection with structured-output and deep-learning architectures
  - Recipe of many recent papers: title
    "Deep X for Extreme Multi-label [Text] Classification"
    where $X \subset$ {Seq2Seq, Neural Networks, Convolution, Attention, LSTM, Adversarial, Sparse, Autoencoder, Latent, ...}.

ECMLPKDD Würzburg 2019 | Home | Programme | Submissions | Attending | Organisation | Sponsorship & Exhibition | Su

Research Track | Applied Data Science Track | Journal Track | **All**

Show 10 entries   Search (Title, Author, Session): multi-label

Tue, 14:20 - 14:40 @ 1.011 (Poster@Wed)   Ranking
Learning to Calibrate and Rerank Multi-label Predictions (391)
Cheng Li (Northeastern University), Virgil Pavlu (Northeastern University), Javed Aslam (Northeastern University), Bingyu Wang (Northeastern University), Kechen Qin (Northeastern University)
Reproducible Research

Wed, 11:40 - 12:00 @ 0.004 (AOK-HS) (Poster@Wed)   Single
Assessing the multi-labelness of multi-label data (562)
Laurence A. F. Park (Western Sydney University), Yi Guo (Western Sydney University), Jesse Read (École Polytechnique)

Thu, 16:20 - 16:40 @ 0.002 (Poster@Thu)   Multi-Label Learning
Data scarcity, robustness and extreme multi-label classification (J29)
Rohit Babbar, Bernhard Schölkopf

Thu, 16:40 - 17:00 @ 0.002 (Poster@Thu)   Multi-Label Learning
Neural Message Passing for Multi-Label Classification (438)
Jack Lanchantin (University of Virginia), Arshdeep Sekhon (University of Virginia), Yanjun Qi (University of Virginia)
Reproducible Research

Thu, 17:00 - 12:20 @ 0.002 (Poster@Thu)   Multi-Label Learning
Synthetic Oversampling of Multi-Label Data based on Local Label Distribution (624)
Bin Liu (Aristotle University of Thessaloniki), Grigorios Tsoumakas (Aristotle University of Thessaloniki)
Reproducible Research

Thu, 17:20 - 17:40 @ 0.002 (Poster@Thu)   Multi-Label Learning
PP-PLL: Probability Propagation for Partial Label Learning (296)
Kaiwei Sun (Chongqing University of Posts), Zijian Min (Telecommunications)
Reproducible Research

Thu, 17:40 - 18:00 @ 0.002 (Poster@Thu)   Multi-Label Learning
Dynamic Principal Projection for Cost-Sensitive Online Multi-Label Classification (J30)
Hong-Min Chu, Kuan-Hao Huang, Hsuan-Tien Lin

Showing 1 to 7 of 7 entries (filtered from 162 total entries)   Previous | 1 | Next

Classifier Chains for Multi-label Classification

Jesse Read*, Bernhard Pfahringer, Geoffrey Holmes, Eibe Frank

* currently at École Polytechnique.

Thank You!!!