Efficient Monte Carlo Optimization for Multi-Label Classifier Chains

Jesse Read, Luca Martino, David Luengo
{jesse,luca}@tsc.uc3m.es, david.luengo@upm.es
Universidad Carlos III & Universidad Politécnica de Madrid (Spain)

Introduction: Multi-label Classification

Multi-label Classification is the supervised learning problem where an instance is associated with multiple classes, rather than with a single class, as in traditional binary or multi-class problems.

Task: learn, from training data $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$, a function:

$$\mathbf{h} : \mathcal{X} \to \mathcal{Y}$$

Mapping input in $\mathcal{X} = \mathbb{R}^D$ to some output in $\mathcal{Y} = \{0,1\}^L$; where
$\mathbf{x}^{(i)} = [x_1, \ldots, x_D] \in \mathcal{X}$ is some *data instance*, and
$\mathbf{y}^{(i)} = [y_1, \ldots, y_L] \in \mathcal{Y}$ is some label vector, where $y_j = 1$ if the j-th label is *relevant* to this i-th example (else $y_j = 0$); e.g.,

$\mathbf{x} =$ 

beach sunset foliage field mountain urban

$\mathbf{y} = [\,1\ 0\ 1\ 0\ 0\ 0\,]$

For a new test instance $\tilde{\mathbf{x}}$, we obtain $\hat{\mathbf{y}} = \mathbf{h}(\tilde{\mathbf{x}})$.

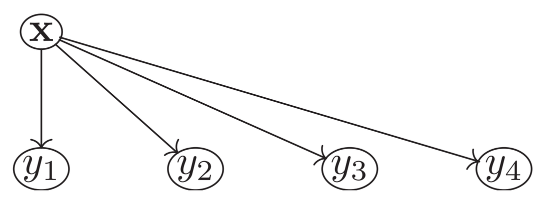
Table: Other Applications / Datasets

	$\mathcal{X}$	$\mathcal{Y}$	L	N	D	$\sum_{j=1}^{L} y_j$
Music	audio data	emotions	6	593	72	1.87
Scene	image data	scene labels	6	2407	294	1.07
Yeast	genes	biological fns	14	2417	103	4.24
Genbase	genes	biological fns	27	661	1185	1.25
Medical	medical text	diagnoses	45	978	1449	1.25
Enron	e-mails	labels, tags	53	1702	1001	3.38
Reuters	news articles	categories	103	6000	500	1.46

Binary Relevance (BR)

- The **Binary Relevance** method (BR): builds one binary classifier for each label

$$\hat{y}_j = h_j(\tilde{\mathbf{x}})$$



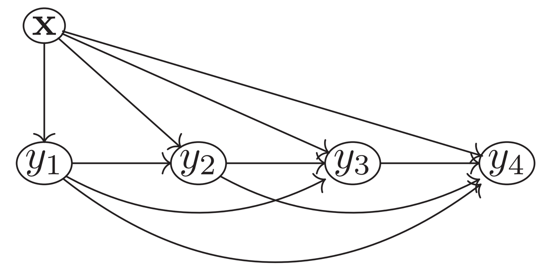
A natural approach to multi-label classification, use any *off-the-shelf* binary classifier. However, does not model label dependencies;

$$p(\mathbf{y}|\mathbf{x}) \neq \prod_{j=1}^{L} p(y_j|\mathbf{x})$$

Chain Classifiers

Chain Classifiers model label dependencies with:

$$p(\mathbf{y}|\mathbf{x}) = p(y_1|\mathbf{x}) \prod_{j=2}^{L} p(y_j|\mathbf{x}, y_1, \ldots, y_{j-1}) \qquad (1)$$



- The **Classifier Chain** (CC) [3] is a greedy approximation:

$$\hat{y}_j = h_j(\tilde{\mathbf{x}}, \hat{y}_1 \ldots, \hat{y}_{j-1}) \equiv \underset{y_j \in \{0,1\}}{\operatorname{argmax}}\, p(y_j|\tilde{\mathbf{x}}, \hat{y}_1, \ldots, \hat{y}_{j-1})$$

for each $j = 1, \ldots, L$. May propagate errors along the chain.

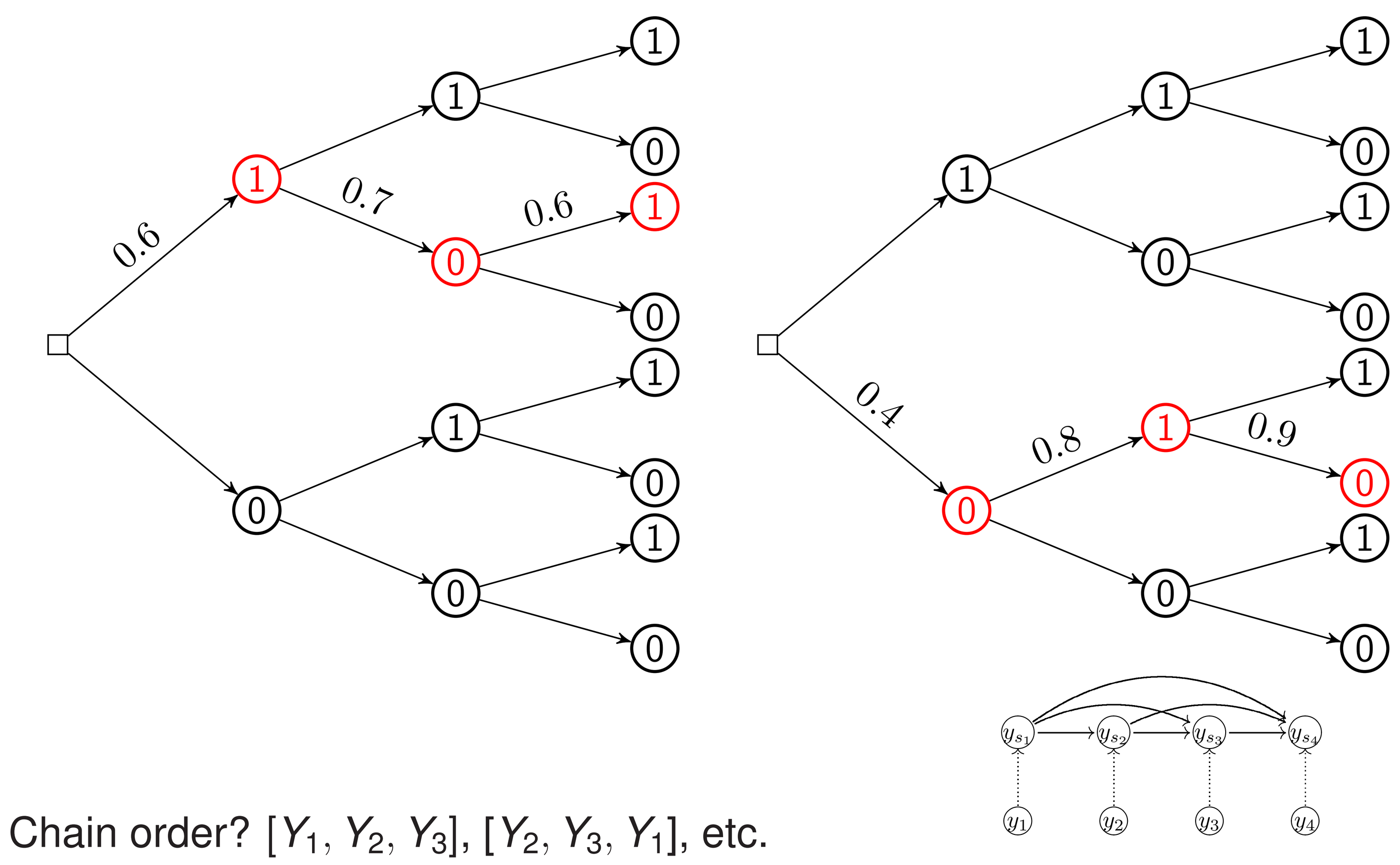
- **Probabilistic Classifier Chains** (PCC) [1] tests all 2^L possible $\mathbf{y}$ on (1):

$$\hat{\mathbf{y}} = \underset{\mathbf{y} \in \mathcal{Y}}{\operatorname{argmax}}\, p(\mathbf{y}|\tilde{\mathbf{x}})$$

This is intractable (for $L > 15$), and also ignores chain order.

- **Ensembles of CC** (ECC) [3] averages results of 10 CCs each with random chain orders.

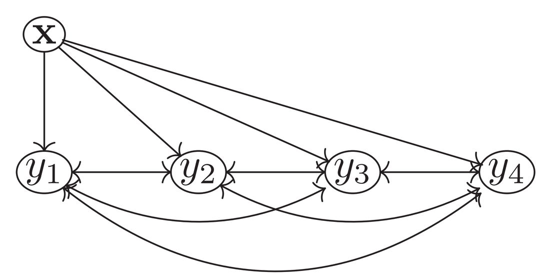
Example: Classifier Chains Prediction



Chain order? $[Y_1, Y_2, Y_3]$, $[Y_2, Y_3, Y_1]$, etc.

Alternative Approach

- **Conditional Dependency Networks** (CDN) [2] fully connected network (among $Y_1, \ldots, Y_L$) instead of a chain.



Inference with Gibbs sampling (over T iterations):
$y_j \sim p(y_j|\tilde{\mathbf{x}}, y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_L)$

Monte Carlo Optimization for Classifier Chains (MCC)

We present:
$\boxed{\text{MCC}}$: Monte Carlo search to find good $\hat{\mathbf{y}}|\tilde{\mathbf{x}}$ (inference time)
$\boxed{\text{M2CC}}$: MCC + find a good chain order $\hat{\mathbf{s}}$ at training time
where $\mathbf{s}$ parameterizes some order of the labels $1, \ldots, L$ w.r.t. $\mathbf{y}$

Training: Find good $\hat{\mathbf{s}}$, build $p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{s}})$

INPUT
$\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{D}$: training data
$\pi(\mathbf{s}|\mathbf{s}_{t-1})$: proposal function
T': number of iterations
ALGORITHM
build model $p(\mathbf{y}|\mathbf{x}, \mathbf{s}_0)$, with random $\mathbf{s}_0$
For $t = 1, \ldots, T'$:
  draw $\mathbf{s}' \sim \pi(\mathbf{s}|\mathbf{s}_{t-1})$
  build model $p(\mathbf{y}|\mathbf{x}, \mathbf{s}')$.
  if $J(\mathbf{s}') > J(\mathbf{s}_{t-1})$
    $\mathbf{s}_t \leftarrow \mathbf{s}'$ accept.
  else
    $\mathbf{s}_t \leftarrow \mathbf{s}_{t-1}$ reject.
OUTPUT
$\hat{\mathbf{s}} = \mathbf{s}_{T'}$ estimated label sequence.
$p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{s}})$ trained model

Inference: Find a good $\hat{\mathbf{y}} \mid \tilde{\mathbf{x}}$, $p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{s}})$

INPUT
$\tilde{\mathbf{x}}$: test instance.
$p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{s}})$: probabilistic model (from training stage).
T: number of iterations
ALGORITHM
obtain an initial path, $\mathbf{y}_0$, using CC.
For $t = 1, \ldots, T$:
  draw $\mathbf{y}' \sim p(\mathbf{y}|\tilde{\mathbf{x}}, \hat{\mathbf{s}})$
  if $p(\mathbf{y}'|\tilde{\mathbf{x}}, \hat{\mathbf{s}}) > p(\mathbf{y}_t|\tilde{\mathbf{x}}, \hat{\mathbf{s}})$
    $\mathbf{y}_t \leftarrow \mathbf{y}'$ accept.
  else
    $\mathbf{y}_t \leftarrow \mathbf{y}_{t-1}$ reject.
OUTPUT
$\hat{\mathbf{y}} = \mathbf{y}_T$: predicted label assignment.

where $\pi(\mathbf{s}|\mathbf{s}_{t-1})$ swaps 2 elements in $\mathbf{s}$; J is a payoff function: $J(\mathbf{s}) : \sum_{i=1}^{N} p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \mathbf{s})$. $T' = 0$ reduces to MCC.

Results

Table: 5×CV: average *exact match* $\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}^{(i)} = \mathbf{h}(\tilde{\mathbf{x}}^{(i)})\right)$: *avg. val* • *rank*

Dataset	BR	CC	PCC	ECC	CDN	MCC	M2CC
Music	0.30 (5)	0.29 (7)	0.35 (2)	0.31 (4)	0.30 (5)	0.35 (2)	0.36 (1)
Scene	0.54 (6)	0.55 (5)	0.64 (2)	0.61 (4)	0.53 (7)	0.64 (2)	0.66 (1)
Yeast	0.14 (5)	0.15 (4)	DNF	0.19 (3)	0.07 (6)	0.21 (1)	0.21 (1)
Genbase	0.94 (4)	0.96 (2)	DNF	0.94 (4)	0.94 (4)	0.96 (2)	0.97 (1)
Medical	0.58 (6)	0.62 (4)	DNF	0.64 (1)	0.60 (5)	0.63 (2)	0.63 (2)
Enron	0.07 (5)	0.10 (2)	DNF	0.11 (1)	0.07 (5)	0.10 (2)	0.10 (2)
Reuters	0.29 (5)	0.35 (4)	DNF	0.36 (2)	0.27 (6)	0.37 (1)	0.36 (2)
avg. rank	5.14	4.00	2.00	2.71	5.43	1.71	1.43

Table: Build Times (seconds, averaged and rounded).

	L	BR	CC	ECC	PCC	CDN	MCC	M$_2$CC
				$M = 10$		$T = 1000$	$T = 100$	$T' = 50$
Music	6	0	0	2	1	6	5	45
Scene	6	12	11	44	15	92	90	1347
Enron	53	102	92	349	DNF	3091	3884	10821
Reuters	101	106	120	1259	DNF	14735	1837	5740

- MCC similar accuracy to PCC, but *tractable to larger datasets*
- M2CC improves on MCC: *chain order can make a difference*
- M(2)CC improves on CDN: *finding a good chain can lead to better inference than in a fully connected network* (and faster!)
- M(2)CC improves on ECC: better than using 10× CC

Key References

[1] Dembczyński et al. *Bayes-Optimal Multi-label Classification via Probabilistic Classifier Chains.* ICML 2010.
[2] Guo and Gu. *Multi-label Classification using Conditional Dependency Networks.* IJCAI 2010.
[3] Read et al. *Classifier Chains for Multi-label Classification.* Mach. Learn. 2011.

Source code available in MEKA framework: `http://meka.sourceforge.net`