Conditional Bernoulli Mixtures for Multi-label Classification

Northeastern University

Cheng Li, Bingyu Wang, Virgil Pavlu, and Javed Aslam {chengli, rainicy, vip, jaa}@ ccs.neu.edu

ICML 2016

Problem

Multi-label classification: assigns a subset of candidate labels to an object (image, document, video, etc.)

`{clouds, lake, sky, sunset, water, reflection}`$\subseteq$`{airport, animal, clouds, book, lake, sky, sunset, cars, water, reflection...}`

Existing Approaches

Binary Relevance: predict each binary label independently
✗ ignore label dependencies

Power-Set: treat each subset as a class + multi-class
✗ poor scalability; cannot predict unseen subsets

CRF: specify label dependencies with graphical models
✗ only model limited (e.g. pair-wise) dependencies

PCC: predict next label based on previous labels
✗ intractable exact inference

CDN: full conditional + Gibbs sampling
✗ cannot handle perfect correlations/anti-correlations

Proposed Model

Approximate the conditional joint by a Conditional Bernoulli Mixture (CBM) with fully factorized mixture components

$$p(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{K} \pi(z=k|\mathbf{x};\alpha) \prod_{\ell=1}^{L} b(y_\ell|\mathbf{x};\beta_\ell^k)$$

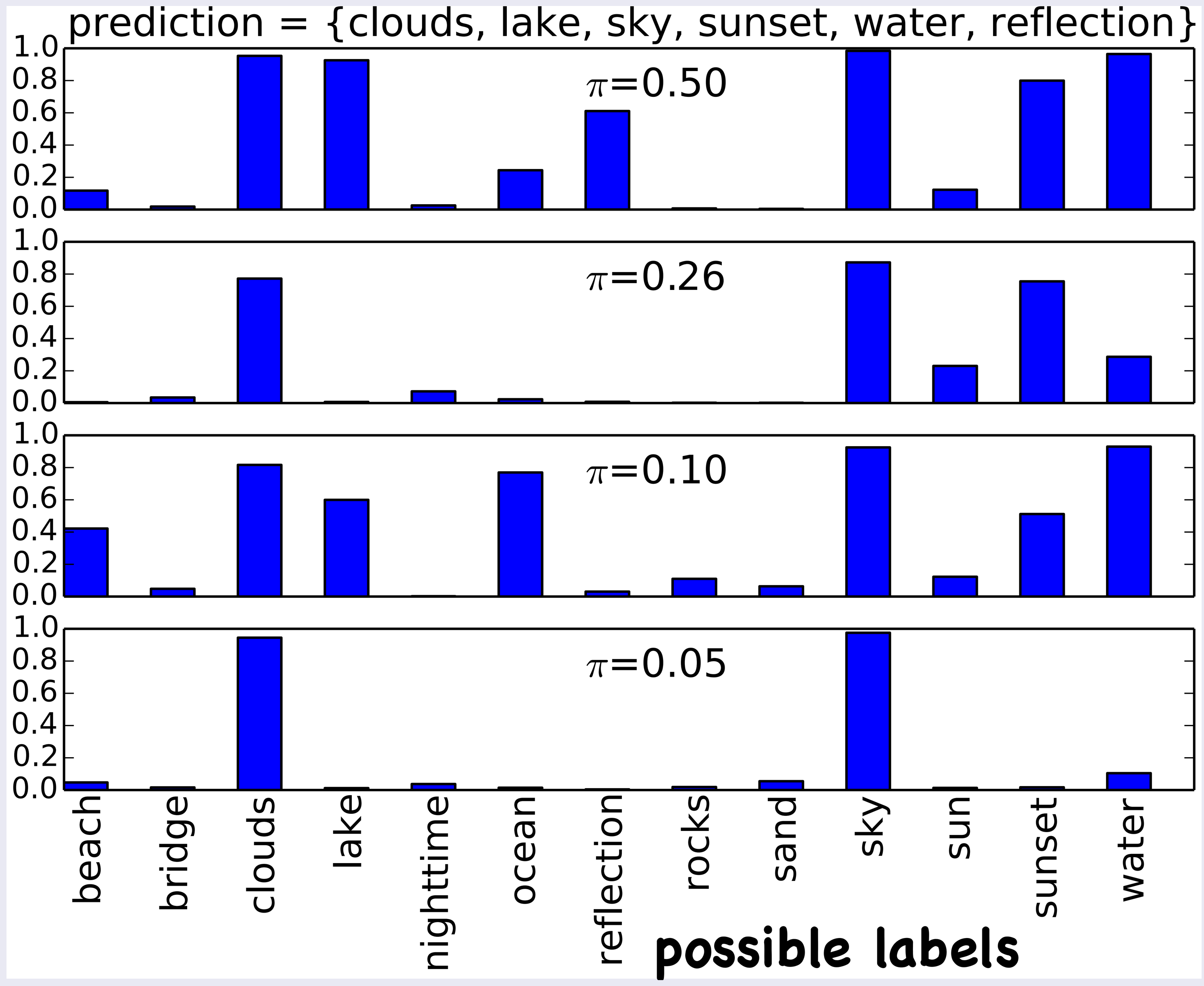
$\pi(z=k|\mathbf{x};\alpha)$: probability of belonging to component k; instantiated with a multi-class classifier

$b(y_\ell|\mathbf{x};\beta_\ell^k)$: probability of getting label y_ℓ in component k; instantiated with a binary classifier

- ✓ capture label dependencies: $p(\mathbf{y}|\mathbf{x}) \neq \prod_{\ell=1}^{L} p(y_\ell|\mathbf{x})$
- ✓ require no prior knowledge on the form of label dependencies
- ✓ subsume Binary Relevance and Power-Set as special cases
- ✓ can predict unseen subsets
- ✓ simple EM training
- ✓ efficient inference for both marginal modes and joint mode

Capturing Label Dependencies: an Example

Top 4 most influential CBM components for the example image



- `water`, `lake`, `sunset` have high marginal probabilities; `reflection` has a low marginal probability – independent predictions miss `reflection`
- `reflection` is positively correlated with `lake`, `water`, and `sunset`
  $\rho_{\texttt{reflection,lake}} = 0.5$, $\rho_{\texttt{reflection,water}} = 0.4$, $\rho_{\texttt{reflection,sunset}} = 0.17$
- predicting the most probable subset includes `reflection`

Training with EM

Given training dataset $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$, use EM to minimize an upper bound of negative log likelihood:

$$\sum_{n=1}^{N} \mathbb{KL}(\Gamma(z_n)||\pi(z_n|\mathbf{x}_n;\alpha)) + \sum_{k=1}^{K}\sum_{\ell=1}^{L}\sum_{n=1}^{N} \gamma_n^k \mathbb{KL}(\mathrm{Ber}(Y_{n\ell};y_{n\ell})||b(Y_{n\ell}|\mathbf{x}_n;\beta_\ell^k))$$

$\Gamma(z_n) = (\gamma_n^1, \gamma_n^2, ..., \gamma_n^K)$ is the posterior membership distribution $p(z_n|\mathbf{x}_n, \mathbf{y}_n)$.
$\mathrm{Ber}(Y_{n\ell};y_{n\ell})$ is the Bernoulli distribution with head probability $y_{n\ell}$.

E step:

$$\gamma_n^k = \frac{\pi(z_n=k|\mathbf{x}_n;\alpha)\prod_{\ell=1}^{L} b(y_{n\ell}|\mathbf{x}_n;\beta_\ell^k)}{\sum_{k=1}^{K}\pi(z_n=k|\mathbf{x}_n;\alpha)\prod_{\ell=1}^{L} b(y_{n\ell}|\mathbf{x}_n;\beta_\ell^k)}$$

M step: nice decomposition into a series of separate optimization problems

$$\alpha_{new} = \operatorname*{argmin}_{\alpha} \sum_{n=1}^{N} \mathbb{KL}(\Gamma(z_n)||\pi(z_n|\mathbf{x}_n;\alpha)) \quad \rightarrow \text{multi-class problem}$$

$$\beta_{\ell\,new}^k = \operatorname*{argmin}_{\beta_\ell^k} \sum_{n=1}^{N} \gamma_n^k \mathbb{KL}(\mathrm{Ber}(Y_{n\ell};y_{n\ell})||b(Y_{n\ell}|\mathbf{x}_n;\beta_\ell^k)) \quad \rightarrow \text{binary problem}$$

Two concrete instantiations:

- with logistic regressions learners: EM + LBFGS
- with gradient boosted trees learners: EM + functional gradient descent

Prediction by Dynamic Programming

Main target measure: subset accuracy

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{1}[\mathbf{y}_n = \hat{\mathbf{y}}_n]$$

Optimal prediction strategy for subset accuracy: joint mode

$$h^*(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$$

Find joint mode by dynamic programming:

- To get a high overall probability, at least one component probability must be high
- In each component, list label subsets in a decreasing probability order with DP
- Iterate round-robin across components and prune remaining suboptimal subsets

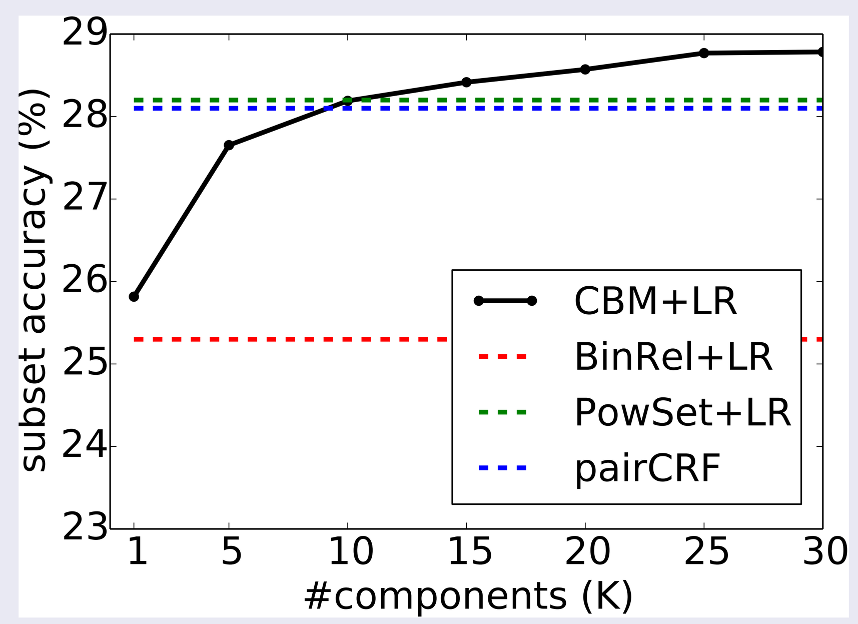
Results

Test subset accuracy of different methods on five datasets. All numbers are in percentages.

	dataset	SCENE	RCV1	TMC2007	MEDIAMILL	NUS-WIDE
	domain	image	text	text	video	image
	#labels / #label subsets	6 / 15	103 / 799	22 / 1341	101 / 6555	81 / 18K
	#features / #datapoints	294 / 2407	47K / 6000	49K / 29K	120 / 44K	128 / 270K
Method	Learner					
BinRel	LR	51.5	40.4	25.3	9.6	24.7
PowSet	LR	68.1	**50.2**	28.2	9.0	26.6
CC	LR	62.9	48.2	26.2	10.9	26.0
PCC	LR	64.8	48.3	26.8	10.9	26.3
ECC-label	LR	60.6	46.5	26.0	11.3	26.0
ECC-subset	LR	63.1	49.2	25.9	11.5	26.0
CDN	LR	59.9	12.6	16.8	5.4	17.1
pairCRF	linear	68.8	46.4	28.1	10.3	26.4
CBM	LR	69.7	49.9	**28.7**	13.5	**27.3**
BinRel	GB	59.3	30.1	25.4	11.2	24.4
PowSet	GB	**70.5**	38.2	23.1	10.1	23.6
CBM	GB	**70.5**	43.0	27.5	**14.1**	26.5

Analysis

Test subset accuracy on TMC dataset with varying number of components K for CBM+LR



- $K = 1$, CBM only estimates marginals and performs similarly to Binary Relevance
- $K > 1$, CBM becomes a better joint estimator and achieves better subset accuracy
- $K = 30$, performance asymptotes