ICML
International Conference
On Machine Learning

Dual Perspective of Label-Specific Feature Learning for Multi-Label Classification

Jun-Yi Hang, Min-Ling Zhang

Southeast University, China

ICML'22, BALTIMORE, MD

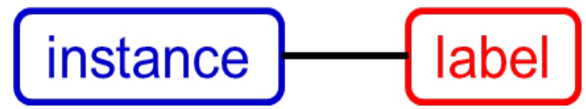
Outline

- **Introduction**
- The DELA Approach
- Experiments
- Conclusion

Multi-Label Classification (MLC)

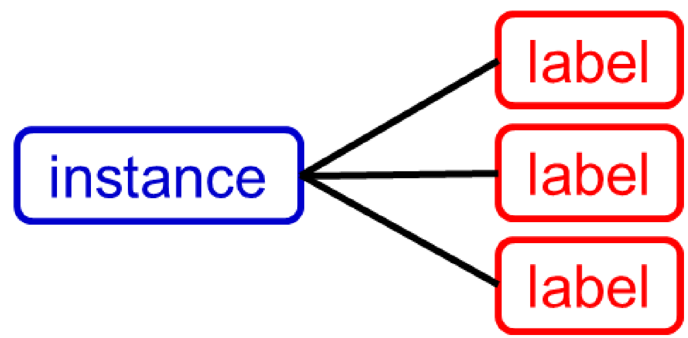
Traditional supervised classification

- Each instance only has one label



Multi-label classification (MLC)

- Each instance can have multiple labels simultaneously



Label-Specific Features (LSF)

Suboptimal

Common strategy

- Binary decomposition
- Classification with the *identical representation*

Fail to consider each label's own discriminative properties!

For example

- Recognizing *plane* category prefers *shape*-based features
- Recognizing *sky* category prefers *color*-based features
- …

Label-Specific Features (LSF)

Suboptimal

Common strategy

- Binary decomposition
- Classification with the *identical representation*

Fail to consider each label's own discriminative properties!

Improved strategy (LSF)

Facilitate the discrimination of each class label by *tailoring its own features*

✓ LSF - *The most pertinent and discriminative features* for each class label

For example

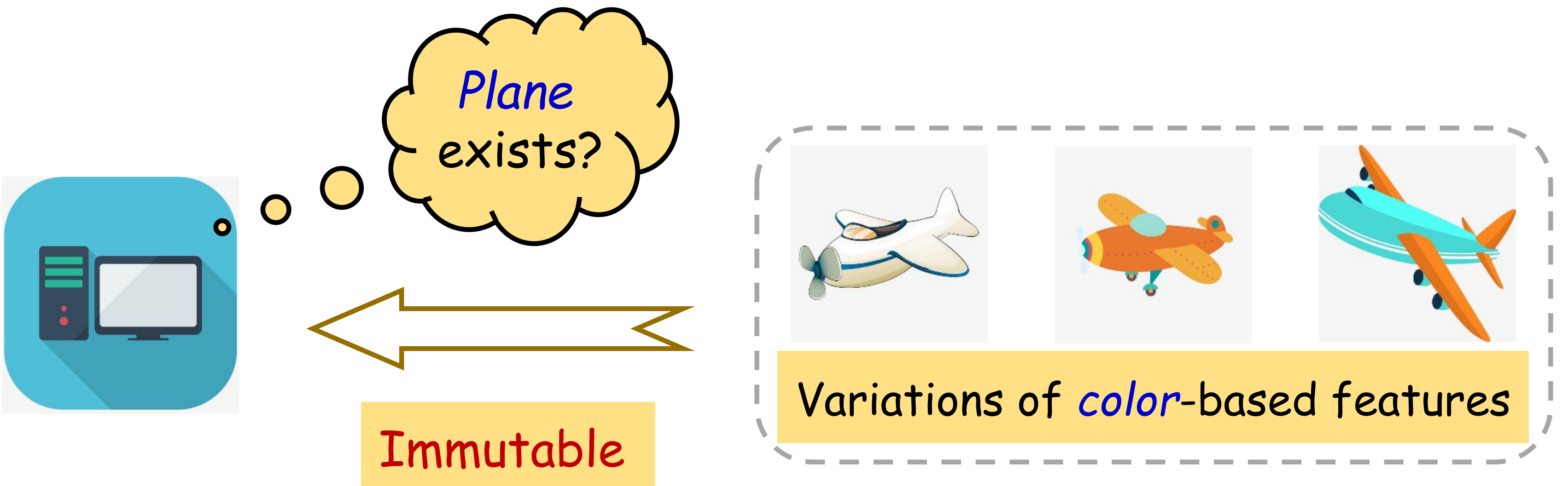
- Recognizing *plane* category prefers *shape*-based features
- Recognizing *sky* category prefers *color*-based features
- …

Dual Perspective of LSF

Our proposal

Basic idea

- Identify *non-informative features* for each class label
- Endow classifiers with *immutability* on these identified features



Outline

- Introduction
- **The DELA Approach**
- Experiments
- Conclusion

Overview

To achieve

Goal 1: identify non-informative features

Goal 2: endow classifiers with immutability

With

Expected risk minimizing (ERM) problem

$$\min_{\phi,S,\vartheta,\Theta} \mathbb{E}_{p(\mathbf{x},\mathbf{y})p_{\vartheta}(\epsilon)}\left[\sum_{k=1}^{t} \mathcal{L}(f_k(e_{\phi}(\mathbf{x})+\mathbf{i}_{S_k}\odot\epsilon;\boldsymbol{\theta}_k), y_k)\right]$$

Perturb non-informative features with random noise

- S_k : a subset of identified non-informative features for the k[th] label

Probabilistically Relaxed ER

Original ERM problem

$$\min_{\phi,S,\vartheta,\Theta} \mathbb{E}_{p(\mathbf{x},\mathbf{y})p_{\vartheta}(\boldsymbol{\epsilon})}[\sum_{k=1}^{t} \mathcal{L}(f_k(e_{\phi}(\mathbf{x})+\mathbf{i}_{S_k}\odot\boldsymbol{\epsilon};\boldsymbol{\theta}_k),y_k)]$$

An intractable subset selection problem is involved!

Relaxed ERM problem

$$\min_{\phi,P,\vartheta,\Theta} \mathbb{E}_{p(\mathbf{x},\mathbf{y})p_{\vartheta}(\epsilon)}[\sum_{k=1}^{t} \mathbb{E}_{p(\mathbf{b}_k)}[\mathcal{L}(f_k(e_{\phi}(\mathbf{x})+\mathbf{b}_k\odot\boldsymbol{\epsilon};\boldsymbol{\theta}_k),y_k)]]$$

A discrete stochastic node is involved!

Further relaxed ERM problem

$$\min_{\phi,P,\vartheta,\Theta} \mathbb{E}_{p(\mathbf{x},\mathbf{y})p_{\vartheta}(\epsilon)}[\sum_{k=1}^{t} \mathbb{E}_{p(\mathbf{c}_k)}[\mathcal{L}(f_k(e_{\phi}(\mathbf{x})+r(\mathbf{c}_k)\odot\boldsymbol{\epsilon};\boldsymbol{\theta}_k),y_k)]]$$

Indicator vector of subset S_k

$$\mathbf{i}_{S_k} \in \{0,1\}^{d_z}$$

Bernoulli gates

$$\mathbf{b}_k \in \{0,1\}^{d_z}$$
$$p(\mathbf{b}_k) = \mathcal{B}(\mathbf{p}_k)$$

Concrete gates

$$\mathbf{c}_k \in [0,1]^{d_z}$$
$$p(\mathbf{c}_k) = \mathcal{C}(\mathbf{p}_k,\tau)$$

Constraint on Noise Level

Expected discrepancy to target noise level

$$\mathbb{E}_{p(\mathbf{c}_k)}[KL(p_{\boldsymbol{\phi},\boldsymbol{\vartheta}}(\mathbf{z}_k|\mathbf{x},\mathbf{c}_k)||q(\mathbf{z}_k))]$$

- $p_{\boldsymbol{\phi},\boldsymbol{\vartheta}}(\mathbf{z}_k|\mathbf{x},\mathbf{c}_k)$: distribution of perturbed stochastic features for the kth label
- $q(\mathbf{z}_k)$: an instance-agnostic prior distribution

Sufficient perturbation endows classifiers with immutability

Overall objective function

$$\min_{\boldsymbol{\phi},P,\boldsymbol{\vartheta},\Theta} \mathbb{E}_{p(\mathbf{x},\mathbf{y})} \sum_{k=1}^{t} \mathbb{E}_{p(\mathbf{c}_k)} \Big[\mathbb{E}_{p(\mathbf{z}_k|\mathbf{x},\mathbf{c}_k)} [\mathcal{L}(f_k(\mathbf{z}_k;\boldsymbol{\theta}_k), y_k)] + \beta \cdot KL(p(\mathbf{z}_k|\mathbf{x},\mathbf{c}_k)||q(\mathbf{z}_k)) \Big]$$

Information Theory Explanation

Connection to the information bottleneck

Corollary. *The overall objective function of* DELA *is an upper bound of the label-wise* ***information bottlenecks****, when the loss function $\mathcal{L}(\cdot,\cdot)$ is instantiated by cross entropy loss*

$$\sum_{k=1}^{t} -I(\mathbf{z}_k; y_k) + \beta \cdot I(\mathbf{z}_k; \mathbf{x})$$

$$\leq \mathbb{E}_{p(\mathbf{x},\mathbf{y})} \sum_{k=1}^{t} \mathbb{E}_{p(\mathbf{c}_k)} \Big[\mathbb{E}_{p(\mathbf{z}_k|\mathbf{x},\mathbf{c}_k)} [-\log q(y_k|\mathbf{z}_k)] + \beta \cdot KL(p(\mathbf{z}_k|\mathbf{x},\mathbf{c}_k)||q(\mathbf{z}_k)) \Big]$$

Discrimination process in DELA conforms to the ***optimal information transportation*** process from x to y!

Outline

- Introduction
- The DELA Approach
- **Experiments**
- Conclusion

Experimental Setup

Comparing Approaches

LIFT, LLSF, C2AE, MPVAE, CLIF, PACA

Evaluation Metrics

Average precision, Macro-averaging AUC,

Hamming loss, One-error, Coverage, Ranking loss

Evaluation Protocol

Ten-fold cross-validation + Wilcoxon signed-ranks test

Experimental Setup – Con't

Data Sets

Ten benchmark multi-label data sets

Dataset	$\|\mathcal{D}\|$	$dim(\mathcal{D})$	$L(\mathcal{D})$	$F(\mathcal{D})$	$LCard(\mathcal{D})$	Domain
corel5k	5000	499	374	Nominal	3.522	Images[1]
rcv1-s1	6000	944	101	Numeric	2.880	Text[1]
Corel16k-s1	13766	500	153	Nominal	2.859	Images[1]
delicious	16105	500	983	Nominal	19.020	Text[1]
iaprtc12	19627	1000	291	Numeric	5.719	Images[2]
espgame	20770	1000	268	Numeric	4.686	Images[2]
mirflickr	25000	1000	38	Numeric	4.716	Images[2]
tmc2007	28596	981	22	Nominal	2.158	Text[1]
mediamill	43907	120	101	Numeric	4.376	Video[1]
bookmarks	87856	2150	208	Nominal	2.028	Text[1]

[1] http://mulan.sourceforge.net/datasets.html
[2] http://lear.inrialpes.fr/people/guillaumin/data.php

$|\mathcal{D}|$: #Examples

$dim(\mathcal{D})$: #Features

$L(\mathcal{D})$: #Labels

$F(\mathcal{D})$: Feature type

$LCard(\mathcal{D})$: Average #labels per instance

Comparative Studies

Summary of the Wilcoxon signed-ranks test for DELA against other comparing approaches at 0.05 significance level
(p-values are shown in the brackets)

DELA against	LIFT	LLSF	C2AE	MPVAE	CLIF	PACA
Average precision	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]
Macro-averaging AUC	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**tie** [0.0059]
Hamming loss	**win** [0.0352]	**win** [0.0313]	**win** [0.0020]	**win** [0.0078]	**win** [0.0117]	**win** [0.0020]
One-error	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0039]
Coverage	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]
Ranking loss	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]

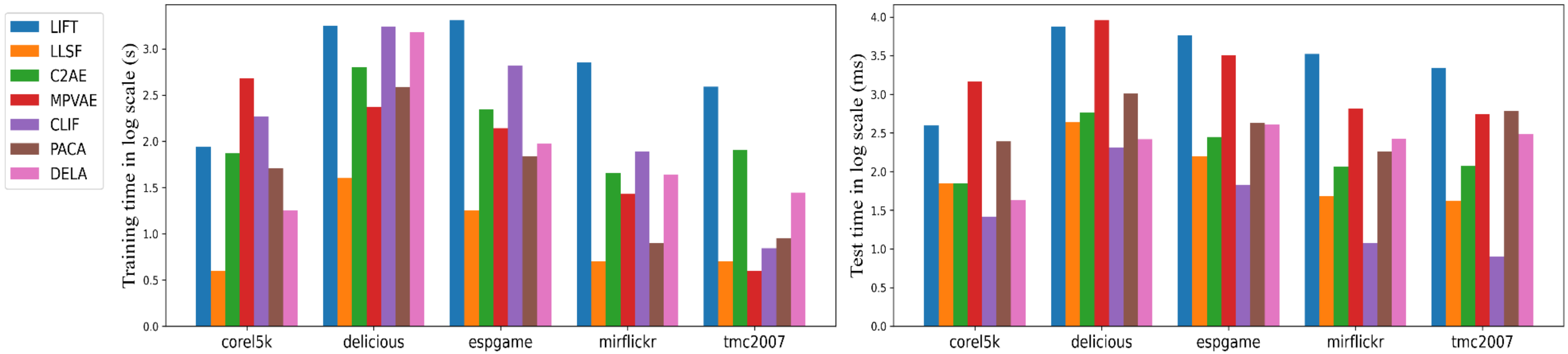
DELA vs. Others

ranks 1st in 92% cases

achieves statistically superior performance

Empirical Running Time

Running time (training/test) of each comparing approach on six benchmark data sets



(a) Training time of comparing approaches

(b) Test time of comparing approaches

DELA vs. Others comparable in time overhead

Outline

- Introduction
- The DELA Approach
- Experiments
- **Conclusion**

Conclusion

Main Contributions

- Propose *a dual perspective* for label-specific feature learning by *endowing classifiers with immutability on identified label-specific non-informative features*
- Provide justification with information theory

Future Work

Explore alternative implementations towards the promising dual perspective

Thanks !