Feature-Induced Labeling Information Enrichment for Multi-Label Learning

Qian-Wen Zhang , Yun Zhong, Min-Ling Zhang

AAAI-2018

Contents

Background

Multi-class Learning

Classifier

$$\begin{array}{cl} \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \\ -1 \\ \vdots \end{bmatrix} & \begin{array}{l} \text{horse} \\ \text{grassland} \\ \text{sky} \\ \text{sea} \\ \text{cow} \\ \dots \end{array} \end{array}$$

Multi-label Learning

Classifier

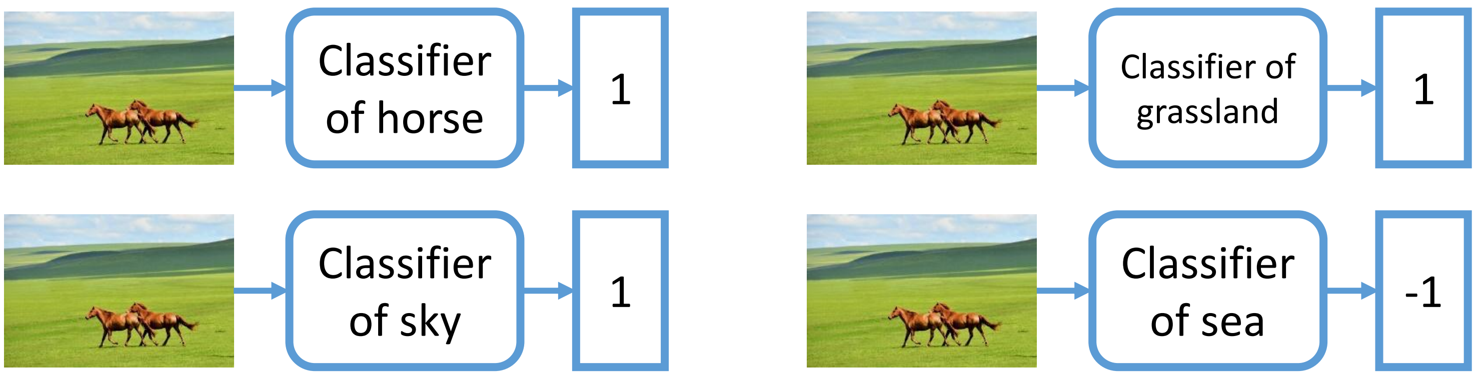
$$\begin{array}{cl} \begin{bmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ \vdots \end{bmatrix} & \begin{array}{l} \text{horse} \\ \text{grassland} \\ \text{sky} \\ \text{sea} \\ \text{cow} \\ \dots \end{array} \end{array}$$

Background

Formalization

- Feature space: d-dimensional $\mathcal{X} = \mathbb{R}^d$
- Label space: q class labels $\mathcal{Y} = \{y_1, y_2, \ldots, y_q\}$
- Training set: p instances $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq p\}$
- Task: learn a predictive model $h : \mathcal{X} \to 2^{\mathcal{Y}}$

Cross Training



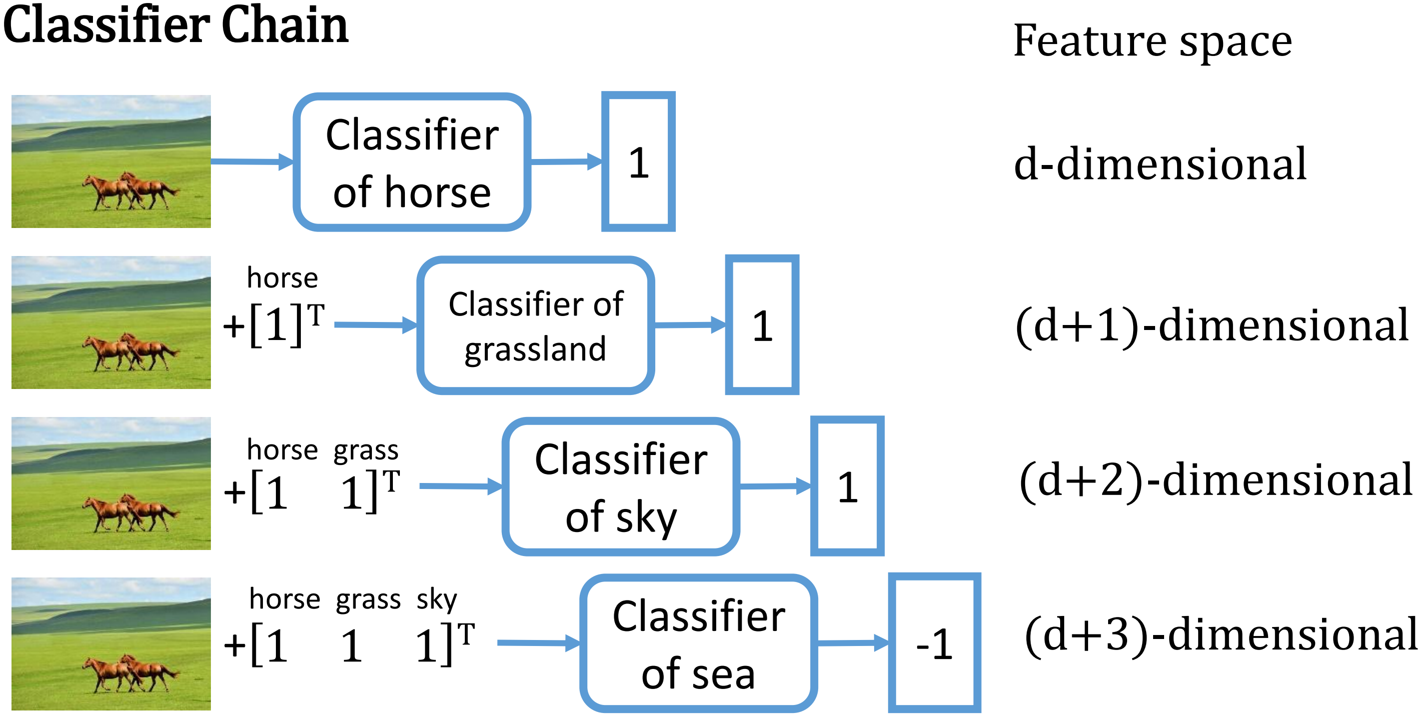
…… $[1 \quad 1 \quad 1 \quad -1 \quad \cdots]^{\mathrm{T}}$ Ignoring the co-existence of other labels.

Background

Label Correlation

- If we know the image already has label horse, the probability of it having label grassland would be high.

Classifier Chain



Contents

- Background
- Introduction
- Approach
- Experiments
- Conclusion

Introduction

Motivation

- Categorical labeling information is actually a simplification of the rich semantics encoded by multi-label training examples.

20% sky, 20% horses and 60% grassland

特朗普就职典礼人数少过奥巴马 摄影师被要求P图

2018-09-08 19:08:51 来源: 观察者网(上海) 举报

103

易信

微信

QQ空间

微博

更多

（原标题：英媒：特朗普干预就职典礼照片 摄影师被要求修图展现人多）

【文/观察者网】特朗普去年上任伊始，就和媒体在就职典礼人数上吵了几个回合。近两年过去，这个问题如今又被提了出来。

英国《卫报》6日独家消息称，特朗普当时因不满参与自己就职典礼的人数看起来比奥巴马少，和时任白宫发言人肖恩·斯派塞（Sean Spicer）亲自致电照片提供方进行干预，结果摄影师只能重新修图，裁去空白部分，让人群看起来多一些。

《卫报》根据《信息自由法案》从美国内政部获得的记录文件显示，特朗普就职后的第一天，即2017年1月21日清晨，他就与发布就职典礼照片的美国国家公园管理局（NPS）的代理主管迈克尔·雷诺兹（Michael Reynolds）通了电话。

80% Trump, 20% Obama

- MLFE: Enrich the labeling information to induce multi-label predictive model with strong generalization performance.

Introduction

Basic Strategy

Leveraging the structural information in the feature space.

- The underlying structure of feature space is characterized by the sparse reconstruction relationships among training examples.
- The reconstruction information is utilized to guide the enrichment process of turning categorical labeling information into numerical labeling information.
- The desired multi-label predictive model is learned from training examples with enriched labeling information based on tailored multivariate regression techniques.

Contents

- Background
- Introduction
- Approach
  - Structural Information Discovery
  - Labeling Information Enrichment
  - Predictive Model Induction
- Experiments
- Conclusion

Approach

Intuition

- To characterize the underlying structure of feature space, MLFE works by modeling the relationship between one example and all the other examples via sparse reconstruction.

$w_1=0.2$ $w_2=0.6$ $w_3=0.2$

Approach

- Feature space: d-dimensional $\mathcal{X} = \mathbb{R}^d$
- Label space: q class labels $\mathcal{Y} = \{y_1, y_2, \ldots, y_q\}$
- Training set: p instances $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le p\}$

Structural Information Discovery

- To characterize the underlying structure of feature space, MLFE works by modeling the relationship between one example and all the other examples via sparse reconstruction.

$$\mathbf{A}_i = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_p]$$

d×(p-1)matrix
all training instances other than $\boldsymbol{x}_i$

$$\boldsymbol{v}_i = [w_{1i}, \ldots, w_{i-1,i}, w_{i+1,i}, \ldots, w_{pi}]^\top$$

(p-1)-dimensional
reconstruction coefficients

- The coefficient vector $\boldsymbol{v}_i$ is learned by solving the following optimization problem:

$$\min_{\boldsymbol{v}_i} \underbrace{\|\mathbf{A}_i \boldsymbol{v}_i - \boldsymbol{x}_i\|_2^2}_{\text{linear reconstruction error}} + \underbrace{\lambda \|\boldsymbol{v}_i\|_1}_{\text{sparsity of the co-efficients}}$$

Solved by ADMM

Contents

Approach

Labeling Information Enrichment

- binary labeling vector $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \ldots, t_{iq})^\top \quad \boldsymbol{t}_i \in \{1, -1\}^q$
- numerical labeling vector $\boldsymbol{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iq})^\top \in \mathbb{R}^q$

$$\min_{\boldsymbol{v}_i} \underbrace{\|\mathbf{A}_i \boldsymbol{v}_i - \boldsymbol{x}_i\|_2^2}_{\text{linear reconstruction error}} + \lambda \underbrace{\|\boldsymbol{v}_i\|_1}_{\text{sparsity of the co-efficients}}$$

- reconstruction error over the training set $E(\mathbf{W}) = \sum_{i=1}^{p} \|\boldsymbol{x}_i - \sum_{j=1}^{p} w_{ji} \boldsymbol{x}_j\|_2^2$

Suppose that the structural relationship specified in the feature space also holds in the output space.

- reconstruction error in the label space $\Phi(\mathbf{U}) = \sum_{i=1}^{p} \|\boldsymbol{u}_i - \sum_{j=1}^{p} w_{ji} \boldsymbol{u}_j\|_2^2$
  $\mathbf{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p]$

Approach

Labeling Information Enrichment

- ***Suppose that*** ***the structural relationship specified in the feature space also holds in the output space.***

$$\Phi(\mathbf{U}) = \sum_{i=1}^{p} \|\boldsymbol{u}_i - \sum_{j=1}^{p} w_{ji}\boldsymbol{u}_j\|_2^2 \quad \mathbf{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p]$$

- The enriched labeling information is generated by solving the following optimization problem:

$$\min_{\mathbf{U}} \sum_{i=1}^{p} \|\boldsymbol{u}_i - \sum_{j=1}^{p} w_{ji}\boldsymbol{u}_j\|_2^2$$

A standard QP problem

$$\text{s.t.} \quad c_1 \leq t_{ij}u_{ij} \leq c_2 \ \ (1 \leq i \leq p, 1 \leq j \leq q)$$

- The constraint ensures that the numerical label possesses the same sign with the binary label and takes value with reasonable magnitude.

Contents

- Background
- Introduction
- Approach
  - Structural Information Discovery
  - Labeling Information Enrichment
  - Predictive Model Induction
- Experiments
- Conclusion

Approach

Predictive Model Induction

- The original multi-label training set can be transformed into its enriched version $\tilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \boldsymbol{u}_i) \mid 1 \leq i \leq p\}$
- Induce the predictive model by employing multi-output regression techniques.

$$\Omega(\boldsymbol{\Theta}, \boldsymbol{b}) = \frac{1}{2}\sum_{j=1}^{q}\|\boldsymbol{\theta}_j\|_2^2 + \beta_1 \sum_{i=1}^{p} \Omega_1(u_i) + \beta_2 \sum_{i=1}^{p}\sum_{j=1}^{q} \Omega_2(o_{ij}) + \beta_3 \sum_{i=1}^{p}\sum_{j=1}^{q} \Omega_3(r_{ij})$$

To minimize the objective function Ω(Θ, b), MLFE employs the quasi-Newton iterative method named Iterative Re-Weighted Least Square (IRWLS)

$$\Omega_1(u_i) = \begin{cases} 0, & u_i < \epsilon \\ (u_i - \epsilon)^2, & u_i \geq \epsilon \end{cases}$$

ϵ-insensitive cost

$$\Omega_2(o_{ij}) = \begin{cases} 0, & o_{ij} > 0 \\ -o_{ij}, & o_{ij} \leq 0 \end{cases}$$

$$o_{ij} = t_{ij}\left(\boldsymbol{\theta}_j^\top \varphi(\boldsymbol{x}_i) + b_j\right)$$

penalize the case where the sign of predictive output is different to that of original binary label

$$\Omega_3(r_{ij}) = \begin{cases} 1, & r_{ij} > 0 \\ 0, & r_{ij} \leq 0 \end{cases}$$

$$r_{ij} = \boldsymbol{\theta}_j^\top \varphi(\boldsymbol{x}_i) + b_j.$$

penalize the case where the predictive model yields large number of relevant labels for the training example

Contents

- Background
- Introduction
- Approach
- Experiments
- Conclusion

Experiments

Table 2: Predictive performance of each comparing algorithm (mean ± std. deviation) on the regular-scale data sets.

Comparing algorithms	One-error ↓							
	cal500	emotions	medical	llog	msra	image	scene	yeast
MLFE	0.129±0.047	0.260±0.030	**0.113±0.041**	**0.669±0.044**	**0.040±0.008**	**0.258±0.020**	**0.149±0.023**	0.231±0.015
LIFT	**0.116±0.040**	**0.255±0.037**	0.152±0.039	0.705±0.049	0.057±0.014	0.277±0.023	0.162±0.023	**0.229±0.013**
RELIAB	0.159±0.062	0.277±0.036	0.168±0.039	0.746±0.028	0.066±0.017	0.350±0.023	0.270±0.023	0.247±0.017
ML^2	0.166±0.078	0.268±0.032	0.129±0.028	0.699±0.038	**0.040±0.013**	0.267±0.022	0.156±0.029	0.254±0.027
CLR	0.269±0.061	0.322±0.032	0.360±0.170	0.830±0.058	0.144±0.027	0.437±0.019	0.344±0.027	0.241±0.012
RAKEL	0.611±0.084	0.315±0.074	0.246±0.038	0.879±0.026	0.239±0.031	0.412±0.029	0.339±0.027	0.280±0.016
Comparing algorithms	**Coverage ↓**							
	cal500	emotions	medical	llog	msra	image	scene	yeast
MLFE	0.764±0.013	**0.278±0.022**	**0.031±0.012**	0.227±0.027	**0.518±0.010**	**0.163±0.014**	**0.016±0.009**	**0.452±0.010**
LIFT	0.759±0.020	0.280±0.025	0.048±0.022	0.173±0.020	0.539±0.011	0.172±0.014	0.020±0.009	0.459±0.010
RELIAB	**0.738±0.013**	0.299±0.029	0.047±0.016	**0.161±0.020**	0.541±0.011	0.199±0.012	0.107±0.011	0.457±0.004
ML^2	0.805±0.013	0.279±0.022	**0.031±0.011**	0.181±0.019	0.522±0.011	0.165±0.012	0.018±0.009	0.455±0.011
CLR	0.795±0.008	0.334±0.020	0.080±0.068	0.186±0.044	0.618±0.013	0.247±0.016	0.137±0.017	0.480±0.008
RAKEL	0.964±0.006	0.348±0.021	0.089±0.019	0.340±0.023	0.670±0.010	0.253±0.009	0.174±0.015	0.564±0.008
Comparing algorithms	**Ranking loss ↓**							
	cal500	emotions	medical	llog	msra	image	scene	yeast
MLFE	0.185±0.008	**0.142±0.020**	0.020±0.010	0.233±0.027	**0.118±0.006**	**0.135±0.014**	**0.052±0.010**	**0.167±0.007**
LIFT	0.182±0.004	**0.142±0.025**	0.033±0.017	0.157±0.021	0.125±0.005	0.147±0.015	0.056±0.009	0.170±0.006
RELIAB	**0.177±0.005**	0.161±0.031	0.031±0.012	0.128±0.018	0.131±0.005	0.181±0.013	0.090±0.010	0.180±0.008
ML^2	0.210±0.009	0.144±0.023	**0.019±0.008**	0.170±0.020	0.119±0.006	0.138±0.011	0.055±0.011	0.172±0.009
CLR	0.224±0.008	0.199±0.024	0.065±0.059	**0.152±0.039**	0.190±0.009	0.243±0.018	0.119±0.016	0.187±0.005
RAKEL	0.364±0.014	0.217±0.026	0.067±0.015	0.292±0.028	0.232±0.011	0.250±0.012	0.154±0.014	0.250±0.005

Experiments

Table 2: Predictive performance of each comparing algorithm (mean ± std. deviation) on the regular-scale data sets.

Comparing algorithms	Average precision ↑							
	cal500	emotions	medical	llog	msra	image	scene	yeast
MLFE	0.490±0.025	0.815±0.020	**0.914±0.024**	0.393±0.036	0.827±0.008	**0.833±0.014**	**0.917±0.014**	**0.767±0.010**
LIFT	0.503±0.015	**0.817±0.027**	0.874±0.029	**0.402±0.039**	0.830±0.008	0.819±0.015	0.909±0.013	0.762±0.008
RELIAB	**0.507±0.019**	0.797±0.028	0.869±0.028	0.393±0.034	0.818±0.009	0.776±0.013	0.840±0.014	0.744±0.011
ML^2	0.456±0.027	0.811±0.022	0.903±0.021	0.396±0.036	**0.836±0.007**	0.827±0.014	0.913±0.016	0.757±0.014
CLR	0.436±0.019	0.762±0.024	0.687±0.192	0.295±0.075	0.741±0.013	0.718±0.014	0.795±0.018	0.745±0.008
RAKEL	0.332±0.019	0.766±0.031	0.802±0.027	0.233±0.026	0.694±0.014	0.725±0.013	0.780±0.018	0.710±0.009
Comparing algorithms	Macro-averaging F1 ↑							
	cal500	emotions	medical	llog	msra	image	scene	yeast
MLFE	0.237±0.022	**0.670±0.052**	**0.720±0.073**	**0.461±0.062**	**0.553±0.015**	**0.658±0.024**	**0.819±0.026**	0.425±0.023
LIFT	0.176±0.021	0.630±0.042	0.690±0.079	0.399±0.057	0.516±0.017	0.621±0.035	0.797±0.015	0.388±0.022
RELIAB	**0.301±0.022**	0.650±0.039	0.712±0.053	0.392±0.058	0.546±0.014	0.556±0.035	0.665±0.025	0.405±0.024
ML^2	0.236±0.021	0.650±0.047	0.674±0.061	0.370±0.060	0.548±0.012	0.646±0.029	0.799±0.029	**0.443±0.025**
CLR	0.211±0.025	0.601±0.038	0.600±0.129	0.395±0.062	0.499±0.017	0.525±0.022	0.620±0.025	0.400±0.018
RAKEL	0.187±0.020	0.618±0.036	0.672±0.058	0.366±0.051	0.492±0.020	0.540±0.012	0.644±0.019	0.430±0.014
Comparing algorithms	Micro-averaging F1 ↑							
	cal500	emotions	medical	llog	msra	image	scene	yeast
MLFE	0.374±0.024	**0.684±0.043**	**0.816±0.032**	**0.211±0.042**	**0.725±0.009**	**0.657±0.021**	**0.810±0.026**	**0.649±0.014**
LIFT	0.323±0.016	0.659±0.024	0.775±0.036	0.177±0.037	0.716±0.015	0.622±0.031	0.787±0.015	0.645±0.011
RELIAB	**0.483±0.012**	0.647±0.036	0.725±0.029	0.181±0.037	0.714±0.007	0.560±0.030	0.654±0.025	0.637±0.011
ML^2	0.358±0.027	0.661±0.039	0.782±0.026	0.057±0.021	0.722±0.008	0.645±0.028	0.788±0.029	0.641±0.014
CLR	0.326±0.019	0.614±0.037	0.598±0.157	0.176±0.049	0.624±0.010	0.525±0.019	0.612±0.026	0.628±0.012
RAKEL	0.355±0.018	0.634±0.031	0.685±0.031	0.148±0.027	0.613±0.015	0.540±0.011	0.636±0.023	0.632±0.009

Experiments

Table 3: Predictive performance of each comparing algorithm (mean ± std. deviation) on the large-scale data sets.

Comparing algorithms	One-error ↓						
	slashdot	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5
MLFE	0.372±0.020	0.646±0.021	0.413±0.013	0.463±0.014	0.472±0.023	0.453±0.020	0.445±0.015
LIFT	0.397±0.026	0.669±0.014	0.415±0.019	0.455±0.012	0.473±0.020	0.457±0.022	0.445±0.017
RELIAB	0.516±0.017	0.718±0.015	0.467±0.020	0.476±0.011	0.477±0.024	0.462±0.015	0.467±0.017
ML^2	**0.363±0.018**	**0.627±0.023**	**0.396±0.021**	**0.448±0.014**	**0.461±0.020**	**0.438±0.017**	**0.437±0.017**
CLR	0.979±0.005	0.741±0.018	0.501±0.027	0.507±0.019	0.533±0.037	0.499±0.017	0.503±0.018
RAKEL	0.615±0.020	0.872±0.014	0.623±0.023	0.592±0.017	0.598±0.018	0.592±0.013	0.595±0.021
Comparing algorithms	Coverage ↓						
	slashdot	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5
MLFE	0.117±0.008	**0.263±0.013**	**0.100±0.007**	**0.094±0.005**	**0.096±0.004**	**0.081±0.006**	**0.094±0.006**
LIFT	0.107±0.009	0.286±0.013	0.132±0.009	0.139±0.006	0.142±0.006	0.120±0.009	0.134±0.004
RELIAB	0.134±0.005	0.300±0.011	0.137±0.009	0.121±0.005	0.125±0.006	0.113±0.011	0.119±0.006
ML^2	**0.101±0.007**	0.288±0.012	0.110±0.008	0.105±0.007	0.109±0.005	0.088±0.007	0.108±0.007
CLR	0.258±0.009	0.287±0.015	0.112±0.008	0.105±0.006	0.114±0.024	0.095±0.007	0.107±0.006
RAKEL	0.218±0.012	0.874±0.012	0.417±0.012	0.359±0.022	0.369±0.014	0.358±0.020	0.363±0.015
Comparing algorithms	Ranking loss ↓						
	slashdot	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5
MLFE	0.098±0.008	**0.109±0.006**	**0.039±0.003**	**0.039±0.002**	**0.040±0.002**	**0.034±0.003**	**0.038±0.002**
LIFT	0.092±0.008	0.122±0.005	0.053±0.003	0.059±0.002	0.062±0.002	0.051±0.004	0.055±0.002
RELIAB	0.118±0.005	0.130±0.005	0.058±0.003	0.045±0.002	0.052±0.002	0.047±0.004	0.048±0.002
ML^2	**0.084±0.006**	0.163±0.008	0.042±0.003	0.043±0.003	0.046±0.003	0.037±0.003	0.043±0.003
CLR	0.245±0.010	0.131±0.008	0.048±0.002	0.046±0.002	0.054±0.020	0.044±0.002	0.046±0.003
RAKEL	0.198±0.013	0.586±0.011	0.233±0.007	0.209±0.012	0.222±0.009	0.224±0.013	0.209±0.012

Experiments

Table 3: Predictive performance of each comparing algorithm (mean ± std. deviation) on the large-scale data sets.

Comparing algorithms	Average precision ↑						
	slashdot	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5
MLFE	0.715±0.015	**0.316±0.012**	0.615±0.010	0.622±0.006	0.614±0.013	0.640±0.013	0.626±0.007
LIFT	0.695±0.019	0.291±0.010	0.582±0.013	0.579±0.008	0.569±0.010	0.596±0.010	0.586±0.009
RELIAB	0.610±0.012	0.269±0.009	0.563±0.010	0.588±0.009	0.586±0.013	0.611±0.010	0.586±0.009
ML^2	**0.728±0.015**	0.315±0.013	**0.629±0.013**	**0.630±0.008**	**0.622±0.011**	**0.647±0.014**	**0.631±0.006**
CLR	0.261±0.006	0.247±0.009	0.564±0.012	0.579±0.011	0.554±0.050	0.589±0.013	0.576±0.012
RAKEL	0.516±0.012	0.120±0.007	0.391±0.009	0.429±0.010	0.423±0.010	0.431±0.011	0.421±0.012
Comparing algorithms	Macro-averaging F1 ↑						
	slashdot	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5
MLFE	**0.455±0.048**	**0.338±0.014**	0.282±0.024	0.263±0.023	0.243±0.021	0.290±0.036	0.265±0.024
LIFT	0.427±0.036	0.324±0.014	0.219±0.038	0.163±0.020	0.151±0.020	0.203±0.034	0.165±0.022
RELIAB	0.433±0.047	0.303±0.019	**0.332±0.026**	**0.332±0.023**	**0.333±0.022**	**0.335±0.039**	**0.332±0.012**
ML^2	0.424±0.050	0.331±0.015	0.228±0.025	0.225±0.020	0.216±0.019	0.263±0.037	0.232±0.020
CLR	0.165±0.035	0.276±0.015	0.278±0.028	0.269±0.016	0.255±0.035	0.297±0.023	0.286±0.013
RAKEL	0.363±0.033	0.257±0.013	0.266±0.029	0.237±0.024	0.243±0.023	0.256±0.020	0.255±0.016
Comparing algorithms	Micro-averaging F1 ↑						
	slashdot	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5
MLFE	**0.550±0.016**	0.177±0.015	0.411±0.011	0.411±0.011	0.397±0.016	0.426±0.013	0.414±0.010
LIFT	0.509±0.020	0.077±0.011	0.311±0.020	0.297±0.013	0.289±0.014	0.327±0.019	0.314±0.012
RELIAB	0.449±0.014	**0.226±0.010**	**0.417±0.007**	**0.468±0.011**	**0.431±0.013**	**0.485±0.009**	**0.466±0.009**
ML^2	0.531±0.015	0.126±0.016	0.378±0.016	0.383±0.014	0.377±0.017	0.411±0.012	0.394±0.015
CLR	0.008±0.003	0.123±0.019	0.361±0.008	0.356±0.015	0.338±0.029	0.365±0.017	0.368±0.014
RAKEL	0.362±0.014	0.135±0.009	0.341±0.008	0.337±0.008	0.335±0.012	0.349±0.010	0.350±0.010

Contents

- Background
- Introduction
- Approach
- Experiments
- Conclusion

Conclusion

- A novel approach is proposed to learning from multi-label data by leveraging the structural information in feature space.
- The key strategy is to convey the structural information modeled by sparse reconstruction in feature space to facilitate generating enriched labeling information in output space.
- The effectiveness of feature-induced labeling information enrichment is clearly validated with extensive experiments on benchmark multi-label data sets.