Multi-label Classification by Semi-supervised Singular Value Decomposition

Michael Ng
Department of Mathematics
Hong Kong Baptist University

April 2018

Outline

- Background
- The Proposed Model
- Experimental Results
- Summary

Classification

- Training data
- Learning methods (classes or labels)
- Testing data for applications

Data

- Objects: attributes/variables/features/dimensions
- Objects: single instance, multi-instance
- Multiple classes, multi-label
- Universal Machine (deep learning) or Specific Learning Model

Multi-label Learning

- Label correlations
- Knowledge acquired from both features and label domain
- Lack of training data
- The performance of supervised learning algorithms may decay significantly
- Information from both multi-labeled data and unlabeled data (semi-supervised learning)

The Problem

- Given a set of labeled data with n_l instances $\{(\hat{\mathbf{x}}_i, \mathbf{y}_i)\}_{i=1}^{n_l}$, where $\hat{\mathbf{x}}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}^k$ are respectively the d-dimensional feature vector and k-dimensional label vector of the ith labeled data, the traditional multi-label learning aims to find a mapping function from $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1\ \hat{\mathbf{x}}_2 \cdots \hat{\mathbf{x}}_{n_l}]$ to $\mathbf{Y} = [\mathbf{y}_1\ \mathbf{y}_2 \cdots \mathbf{y}_{n_l}]$ using labeled data only.
- Each entry of the label vector indicates whether the current instance belongs to the corresponding class.
- In real applications, there are amounts of unlabeled data with n_u instances denoted as $\check{\mathbf{X}} = [\check{\mathbf{x}}_1\ \check{\mathbf{x}}_2 \cdots \check{\mathbf{x}}_{n_u}]$, where $\check{\mathbf{x}}_i \in \mathbb{R}^d$. The whole dataset is denoted as $\mathbf{X} = [\hat{\mathbf{X}}, \check{\mathbf{X}}]$ with n instances and $n = n_l + n_u$.
- Our goal is to effectively and efficiently find a good mapping from $\hat{\mathbf{X}}$ to $\mathbf{Y}$ by using the whole dataset $\mathbf{X}$.

The Model

- The proposed semi-supervised multi-label learning model:

$$\min_{f} \sum_{i=1}^{n_l} L(\mathbf{y}_i, f(\hat{\mathbf{x}}_i)) + \lambda\Phi(f) + \gamma\Psi(f).$$

  f indicates the desired mapping function for multi-label learning that we need to solve.
- The data fidelity term $L(\cdot)$ can be any loss function which measures the error between the given multi-labeled data and the prediction result generated by the mapping f. Minimization of $L(\mathbf{y}_i, f(\hat{\mathbf{x}}_i))$ keeps the mapping results fit the given label.
- $\Phi(f)$ and $\Psi(f)$ are the regularization terms based on some prior assumptions on desired f.

Data Fidelity Term

- $L(\cdot)$: least squares, the hinge, and the logistic loss functions
- The linear mapping $\mathbf{U} \in \mathbb{R}^{k \times d}$:

$$L_1(\mathbf{y}_i, f(\mathbf{U}, \hat{\mathbf{x}}_i)) = \|\mathbf{U}\hat{\mathbf{x}}_i - \mathbf{y}_i\|_2^2 = \sum_{j=1}^{k} ([\mathbf{y}_i]_j - [\mathbf{U}\hat{\mathbf{x}}_i]_j)^2$$

$$L_2(\mathbf{y}_i, f(\mathbf{U}, \hat{\mathbf{x}}_i)) = \sum_{j=1}^{k} \max\{0, 1 - [\mathbf{y}_i]_j \times [\mathbf{U}\hat{\mathbf{x}}_i]_j\}$$

$$L_3(\mathbf{y}_i, f(\mathbf{U}, \hat{\mathbf{x}}_i)) = \sum_{j=1}^{k} \log(1 + e^{-[\mathbf{y}_i]_j \times [\mathbf{U}\hat{\mathbf{x}}_i]_j})$$

- Convex functions
- Nonlinear setting can be considered.

Regularization of Complexity

- Make use of SVD for desired linear mapping function $\mathbf{U}$:

$$\sum_{j=1}^{r} \mathbf{p}_j(\mathbf{U})\sigma_j(\mathbf{U})\big(\mathbf{q}_j(\mathbf{U})\big)^T$$

  with $r = \min\{k, d\}$

- $\{\mathbf{p}_1(\mathbf{U}), \mathbf{p}_2(\mathbf{U}), \cdots, \mathbf{p}_r(\mathbf{U})$ are referred as label component vectors and $\{\mathbf{q}_1(\mathbf{U}), \mathbf{q}_2(\mathbf{U}), \cdots, \mathbf{q}_r(\mathbf{U})$ are called feature component vectors.
- The complexity of $\mathbf{U}$ is measured by summation of all the non-zero singular values of the matrix:

$$\Phi(\mathbf{U}) = \|\mathbf{U}\|_* = \sum_{j=1}^{r} \sigma_j(\mathbf{U}),$$

  $\| \star \|_*$ denotes the nuclear norm of a matrix.

Regularization of Complexity

- Suppose $r' < r$ singular values are kept for the mapping function throughout the minimization process on $\Phi(\mathbf{U})$, we transform each data point $\hat{\mathbf{x}}_i$ from the feature space to label space by:

$$\mathbf{U}\hat{\mathbf{x}}_i = \sum_{j=1}^{r'} \sigma_j(\mathbf{U})[\big(\mathbf{q}_j(\mathbf{U})\big)^T \hat{\mathbf{x}}_i]\mathbf{p}_j(\mathbf{U}),$$

  which exactly gives the intuitive idea of such regularization: to recognize and approximately represent each label vector by the linear combination of very small number of r' label component vectors based on the fact that label vectors of similar instances should be highly correlated.
- Such regularization can be helpful especially for case with very limited training data available. The low-rank regularization is also capable of correcting the missing labels in the training data.

Regularization of the Smoothness

- Force the optimal mapping function $\mathbf{U}$ to be smooth which can preserve the intrinsic geometry structure in feature space.
- For two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ that are close to each other in feature space, we expect that $\mathbf{y}_i$ (i.e., $\mathbf{U}\mathbf{x}_i$) and $\mathbf{y}_j$ (i.e., $\mathbf{U}\mathbf{x}_j$) should be also close to each other in label space.
- Express the intrinsic geometrical structure in feature space effectively, one useful approach is to construct a c-nearest neighbor graph via employ all the n instances available in feature space as vertices.

Regularization of the Smoothness

- The edge weight here is computed by adopting the heat kernel weight: for each instance $\mathbf{x}_i$, $a_{i,j} = a_{j,i} = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}\right)$ only if an edge is assigned between the instance $\mathbf{x}_j$ and $\mathbf{x}_i$. Otherwise, set $a_{i,j} = 0$ as $\mathbf{x}_i$ and $\mathbf{x}_j$ are not connected.
- Then $\mathbf{A} = [a_{i,j}]$ models the local invariance assumption by utilizing the so called manifold regularization technique.
- $$\Psi(\mathbf{U}) = \frac{1}{2}\sum_{i,j=1}^{n} a_{i,j} \, \|\mathbf{U}\mathbf{x}_i - \mathbf{U}\mathbf{x}_j\|_2^2 = tr((\mathbf{U}\mathbf{X})\mathbf{L}(\mathbf{U}\mathbf{X})^T)$$
- All the instances $\mathbf{X} = [\hat{\mathbf{X}}, \check{\mathbf{X}}]$ in the feature space can be included.

The Model

Semi-supervised **L**ow-**R**ank **M**apping:

$$\min_{\mathbf{U}} \sum_{i=1}^{n_l} \sum_{j=1}^{k} ([\mathbf{y}_i]_j - [\mathbf{U}\hat{\mathbf{x}}_i]_j)^2 + \lambda ||\mathbf{U}||_* + \gamma tr((\mathbf{UX})\mathbf{L}(\mathbf{UX})^T)$$

$$\min_{\mathbf{U}} \sum_{i=1}^{n_l} \sum_{j=1}^{k} \max\{0, 1 - [\mathbf{y}_i]_j \times [\mathbf{U}\hat{\mathbf{x}}_i]_j\} + \lambda ||\mathbf{U}||_* + \gamma tr((\mathbf{UX})\mathbf{L}(\mathbf{UX})^T)$$

$$\min_{\mathbf{U}} \sum_{i=1}^{n_l} \sum_{j=1}^{k} \log(1 + e^{-[\mathbf{y}_i]_j \times [\mathbf{U}\hat{\mathbf{x}}_i]_j}) + \lambda ||\mathbf{U}||_* + \gamma tr((\mathbf{UX})\mathbf{L}(\mathbf{UX})^T)$$

ADMM

The Error Bound

- We consider a distribution $\mathtt{D}$ for data points and labels.
- We receive n_l training points $\{(\hat{\mathbf{x}}_i, \mathbf{y}_i)\}_{i=1}^{n_l}$ sampled i.i.d. from $\mathtt{D}$.
- We assume that the ground truth label vectors $\mathbf{y}_i$ appear at s random locations $z_i^1, z_i^2, \cdots, z_i^s$ chosen from the set $[K] = \{1, 2, \cdots, k\}$ independent of $(\hat{\mathbf{x}}_i, \mathbf{y}_i)$.
- To minimize the empirical risk:

$$\hat{\mathcal{L}}(\mathbf{U}) \equiv \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{s} L(\mathbf{y}_i^{z_i^j}, f^{z_i^j}(\mathbf{U}, \mathbf{x}_i)),$$

- The population risk:

$$\mathcal{L}(\mathbf{U}) \equiv \mathop{\mathbb{E}}_{\mathbf{y},\mathbf{x},z} \left[\!\left[L(\mathbf{y}^z, f^z(\mathbf{U}, \mathbf{x})) \right]\!\right].$$

The Error Bound

A predictor $\mathbf{U}$ is determined by solving empirical risk minimization:

$$\hat{\mathcal{L}}(\mathbf{U}) \quad \text{subject to } \|\mathbf{U}\|_* + \gamma tr((\mathbf{UX})\mathbf{L}(\mathbf{UX})^T) \leq \tau$$

over a set of n training points. Then with probability at least $1-\delta$, we have

$$\mathcal{L}(\hat{\mathbf{U}}) \leq \inf_{\|\mathbf{U}\|_* \leq \tau} \mathcal{L}(\mathbf{U}) + \mathcal{O}\left(s\tau\sqrt{\frac{1}{n}}\right) + \mathcal{O}\left(s\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right),$$

with $\mathbb{E}\left[[\|\mathbf{x}\|_2^2]\right] \leq 1$. Therefore, we expect $\hat{\mathbf{U}}$ has good generalization properties in learning.

Related Work

- In order to identify the latent information in label space, the original label space as a hypercube and mined its principal components by

$$(PLST)\ \max_{\mathbf{P}} tr(\mathbf{P}^T\mathbf{Y}\mathbf{Y}^T\mathbf{P}) \text{ s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I},$$

  where $\mathbf{P} \in \mathbb{R}^{k \times b}$ consists of the normalized eigenvectors of $\mathbf{Y}\mathbf{Y}^T$ corresponding to its b largest eigenvalues.
- Extend it by integrating the labeled data information $\hat{\mathbf{X}}$ via

$$(CPLST)\ \max_{\mathbf{P}} tr(\mathbf{P}^T\mathbf{Y}\hat{\mathbf{X}}^{\dagger}\hat{\mathbf{X}}\mathbf{Y}^T\mathbf{P}) \text{ s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I},$$

  where $\hat{\mathbf{X}}^{\dagger}$ is the pseudo-inverse of $\hat{\mathbf{X}}$.

Related Work

- Maximize the recoverability of the label space and the predictability of the feature space via

$$(FAIE)\ \max_{\mathbf{C}} tr(\mathbf{C}^T(\mathbf{Y}^T\mathbf{Y} + \alpha\hat{\mathbf{X}}^T(\hat{\mathbf{X}}\hat{\mathbf{X}}^T)^{-1}\hat{\mathbf{X}})\mathbf{C}) \text{ s.t. } \mathbf{C}^T\mathbf{C} = \mathbf{I},$$

  where $\mathbf{C} \in \mathbb{R}^{n_l \times b}$ indicates the relationships between data instances and the latent space. We note that $\mathbf{C}$ cannot explicitly reflect the correlation between labels which is a main point in multi-label learning.

Experiments on Synthetic Data

The data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ (here n is the number of instances, m is the number of features) was generated via Gaussian distribution in $[0, 1]$ and then set the cell value to be 1 if it is larger than ζ, otherwise set it to be 0. This step makes the density of feature space. When ζ is large (small), the data is sparse (dense). For multi-label data, we can assume that labels are the combinations of different features. In this case, we take each feature as one label, and make the combination of any two features refer to one label. The resulting label information $\mathbf{Y} \in \mathbb{R}^{n \times k}$ with $k = d + \frac{d(d-1)}{2}$ can be built.

Criteria

- precision = true-pos / (true-pos + false-pos)
- recall = true-pos / (true-pos + false-neg)
- f1 score = 2 (precision x recall) (precision + recall)
- f1 score can be interpreted as a weighted average of the precision and recall
- macro is the unweighted average of the precision/recall taken separately for each class
- micro average on the contrary is an average over instances: therefore classes which have many instances are given more importance
- roc curve is a graph where the x-axis represents the number of true negatives and the y-axis the number of true positives (thresholding values for labels)

Experiments on Synthetic Data

#Features (d)		20	30	40	50	60	70	80	90	100
#Labels (k)		210	465	820	1275	1830	2485	3240	4095	5050
$\zeta = 0.2$	CPLST	0.132	0.259	0.586	1.280	2.088	4.255	10.389	25.352	49.441
	FAIE	0.602	0.712	0.933	0.806	0.986	1.110	1.305	1.230	1.602
	SLRM	0.134	0.301	0.910	1.550	6.730	15.553	24.753	51.442	98.165
$\zeta = 0.5$	CPLST	0.137	0.277	0.580	1.313	2.155	4.344	11.582	26.505	49.805
	FAIE	0.644	0.701	0.977	0.854	1.148	1.207	1.516	1.332	1.745
	SLRM	0.127	0.302	0.576	1.848	6.500	15.911	24.811	51.888	99.619
$\zeta = 0.7$	CPLST	0.128	0.274	0.581	1.360	2.167	4.382	11.435	27.710	49.957
	FAIE	0.603	0.732	0.941	0.809	1.090	1.120	1.280	1.349	1.656
	SLRM	0.180	0.304	1.578	1.625	6.653	16.059	24.017	54.561	97.146

Table: Running time (s) on Synthetic data by varying feature and label sizes but fixing number of samples.

Experiments on Synthetic Data

($\times 10^4$)	1	2	3	4	5	1	2	3	4	5
Trai.			10%					20%		
CPLST	1.155	2.138	2.519	3.505	3.912	1.725	2.503	3.437	4.880	7.999
FAIE	1.050	4.066	12.665	28.178	64.594	4.655	37.452	110.233	263.708	523.181
SLRM	1.590	2.056	2.874	3.098	4.221	1.525	1.957	2.662	2.930	3.892
CPLST	1.258	1.756	2.470	3.078	5.421	1.330	2.210	3.164	4.448	7.764
FAIE	1.065	3.284	10.606	27.166	81.030	3.115	26.778	101.703	257.433	520.166
SLRM	1.513	1.905	2.291	2.717	3.884	1.562	1.916	2.376	3.023	4.176
CPLST	1.169	2.006	2.692	2.894	5.235	1.392	2.520	2.973	4.273	6.448
FAIE	0.801	3.512	14.183	27.746	72.791	3.046	26.781	101.100	249.831	511.292
SLRM	1.664	1.942	2.108	3.049	4.697	1.839	2.190	2.367	3.220	5.088

Table: Running time (s) on Synthetic data by varying the number of samples and fixing the number of features and labels. $\zeta = 0.2, 0.5, 0.7$.

Experiments on Synthetic Data

Data Set	Synthetic		Emotion		Birds		MSRC	
	$\gamma > 0$	$\gamma = 0$	$\gamma > 0$	$\gamma = 0$	$\gamma > 0$	$\gamma = 0$	$\gamma > 0$	$\gamma = 0$
AUC	1.0000	0.9865	0.8155	0.8061	0.7138	0.6876	0.8253	0.5467
Macro-F1	0.9639	0.9522	0.6733	0.6332	0.3284	0.3025	0.4481	0.2232
Micro-F1	0.9645	0.9529	0.6988	0.6338	0.4574	0.4251	0.5890	0.4424
Accuracy	0.9285	0.9153	0.5853	0.5406	0.3208	0.3059	0.3866	0.2699
Data Set	Mediamill		CAL500		Corel5k		SUN	
	$\gamma > 0$	$\gamma = 0$	$\gamma > 0$	$\gamma = 0$	$\gamma > 0$	$\gamma = 0$	$\gamma > 0$	$\gamma = 0$
AUC	0.7969	0.7456	0.5585	0.5621	0.5762	0.5014	0.7126	0.6085
Macro-F1	0.1413	0.1355	0.1655	0.1609	0.0497	0.0359	0.2603	0.2265
Micro-F1	0.6476	0.6388	0.4818	0.4115	0.2700	0.2174	0.5043	0.4508
Accuracy	0.4691	0.4532	0.3087	0.2604	0.1566	0.1410	0.3388	0.2946

Table: Comparison of classification performance of SLRM on one synthetic dataset and seven real world multimedia datasets with $\gamma > 0$ and $\gamma = 0$.

Experiments on Real Data

Dataset	Domain	n	d	k	*cardinality*
Emotion	music	593	72	6	1.869
Birds	audio	645	258	19	1.104
MSRC	image	591	512	23	2.508
CAL500	music	502	68	174	26.044
Corel5K	image	5000	499	374	3.522
SUN	image	14240	512	102	15.526
Mediamill	video	43907	210	101	4.376

Table: Multi-label dataset summary.

Dataset	Evaluation	CPLST	FAIE	MLLOC	MC	MIML	SLRM
Emotion	AUC	0.7513	0.7427	0.8021	0.7866	0.8082	**0.8155**
	Macro-F1	0.5986	0.5880	0.6567	0.5872	0.6619	**0.6733**
	Micro-F1	0.6009	0.5918	0.6892	0.6054	0.6734	**0.6988**
	Accuracy	0.5015	0.4904	0.5790	0.4949	0.5773	**0.5853**
	Running time (s)	0.006	0.008	3.44	6.97	11.23	0.011
Birds	AUC	0.6735	0.6600	0.6738	0.6715	0.7115	**0.7236**
	Macro-F1	0.2297	0.2347	0.2309	0.2875	0.3013	**0.3284**
	Micro-F1	0.4059	0.4040	0.4000	0.3822	0.4138	**0.4574**
	Accuracy	0.3073	0.2919	0.2962	0.2797	0.2873	**0.3208**
	Running time (s)	0.013	0.026	3.65	7.36	31.35	0.169
MSRC	AUC	0.7887	0.7780	0.5400	0.7857	0.8133	**0.8253**
	Macro-F1	0.3317	0.3467	0.1048	0.2541	0.4083	**0.4481**
	Micro-F1	0.5109	0.5357	0.3692	0.4196	0.5538	**0.5890**
	Accuracy	0.3281	0.3344	0.2070	0.2353	0.2801	**0.3866**
	Running time (s)	0.059	0.141	33.49	35.55	687.78	0.731
Mediamill	AUC	0.7938	0.7793	0.7918	0.7563	0.7705	**0.7969**
	Macro-F1	0.0982	0.1266	0.1399	0.1269	0.1298	**0.1413**
	Micro-F1	0.5785	0.6422	0.6381	0.6273	0.6412	**0.6476**
	Accuracy	0.4264	0.4265	0.4326	0.4509	0.4465	**0.4691**
	Running time (s)	0.278	10.09	4928.37	2534.60	8953.65	0.790
CAL500	AUC	0.5471	0.5468	0.5155	0.5211	0.5454	**0.5585**
	Macro-F1	0.1547	0.1541	0.1309	0.1366	0.1399	**0.1655**
	Micro-F1	0.4401	0.4410	0.4626	0.4703	0.4704	**0.4818**
	Accuracy	0.3022	0.3024	0.3027	0.3074	0.3016	**0.3087**
	Running time (s)	0.013	0.018	438.04	19.32	1343.89	0.318
Corel5k	AUC	0.5534	0.5547	**0.5786**	0.5317	0.5573	0.5762
	Macro-F1	0.0383	0.0411	0.0273	0.0419	0.0422	**0.0497**
	Micro-F1	0.2241	0.2220	0.2230	0.2305	0.2322	**0.2700**
	Accuracy	0.1256	0.1162	0.1332	0.1447	0.1306	**0.1566**
	Running time (s)	1.53	2.67	17021.46	1441.99	3957.35	15.36
SUN	AUC	0.7020	0.6950	0.6753	0.6760	0.6661	**0.7126**
	Macro-F1	0.2196	0.2630	0.1923	0.2507	**0.2852**	0.2687
	Micro-F1	0.4605	0.4936	0.4441	0.4670	0.4521	**0.5043**
	Accuracy	0.3009	0.3287	0.2877	0.3054	0.2954	**0.3388**
	Running time (s)	0.2050	1.1104	571.69	1927.21	4016.15	0.7182

Experiments on Real Data

	Related		Related		Related		Related
aerop.	road(0.157) sky(0.133)	build.	body(0.231) car(0.217)	face	body(0.425) build.(0.231)	sheep	grass(0.072) tree(0.066)
bicycle	tree (0.057) build.(0.050)	car	build.(0.152) road(0.143)	flower	face(0.119) grass(0.081)	sign	road(0.089) build.(0.067)
bird	build.(0.094) grass(0.074)	cat	road(0.037) grass(0.020)	grass	cow(0.168) sky(0.123)	sky	tree(0.256) road(0.242)
boat	water(0.166) tree(0.142)	chair	grass(0.042) build.(0.039)	horse	grass(0.016) tree(0.015)	tree	road(0.271) sky(0.256)
body	face(0.425) build.(0.217)	cow	grass(0.218) tree(0.106)	mount.	water(0.084) boat(0.056)	water	boat(0.168) tree(0.142)
book	face(0.133) body(0.133)	dog	road(0.069) body(0.058)	road	tree(0.271) sky(0.242)		

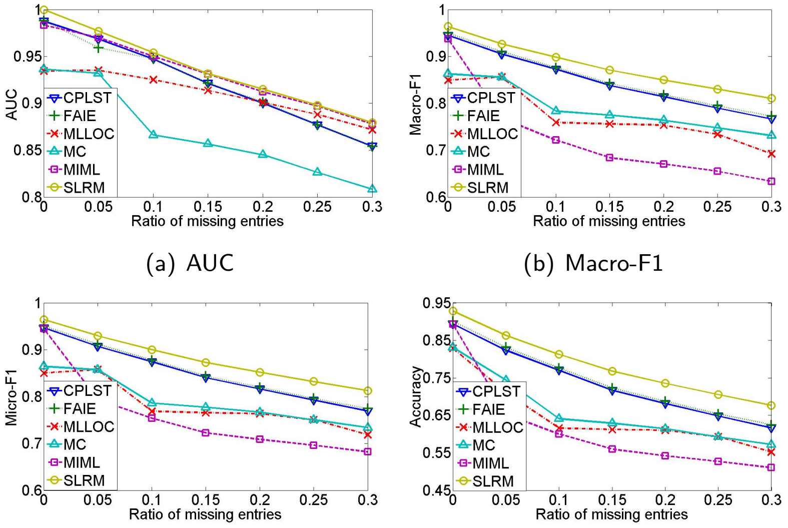
Demonstration of label correlation identified by SLRM on MRSC data.

Experiments on Real Data

Dataset	Emotion			CAL500			Corel5k		
	LS	LL	HL	LS	LL	HL	LS	LL	HL
AUC	0.8155	0.8144	0.8099	0.5585	0.5535	0.5471	0.5762	0.6358	0.6090
Macro-F1	0.6733	0.6832	0.6626	0.1655	0.1580	0.1675	0.0497	0.0461	0.0456
Micro-F1	0.6988	0.6961	0.6814	0.4818	0.4674	0.4626	0.2700	0.2579	0.2406
Accuracy	0.5853	0.6070	0.5966	0.3087	0.3053	0.3009	0.1566	0.1686	0.1604
Time (s)	0.011	8.776	4.005	0.318	333.875	45.123	15.360	55746.251	882.683

Table: Effect of loss function on semi-supervised multi-label classification.

Experiments on Real Data



(a) AUC

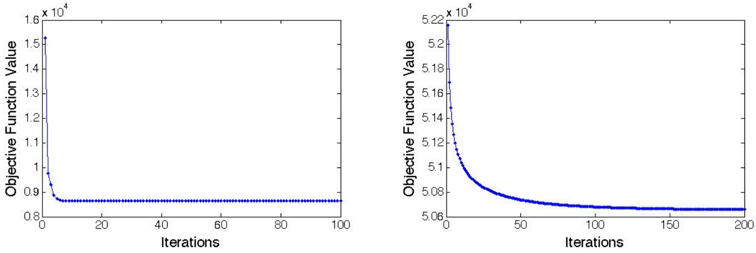
(b) Macro-F1

(c) Micro-F1

(d) Accuracy

Figure: Comparison results under varying the ratio of missing entries in label matrix (Y) of Synthetic data set with 1000 samples, 50 features, 1275 labels and 10% data as training set.

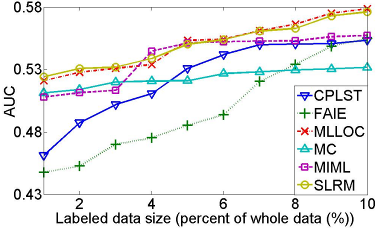
Experiments on Real Data



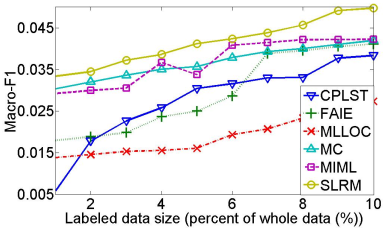
(a) Corel5K (b) Mediamill

Figure: Convergence of SLRM on Corel5K and Mediamill.

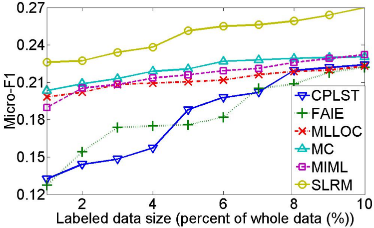
Experiments on Real Data



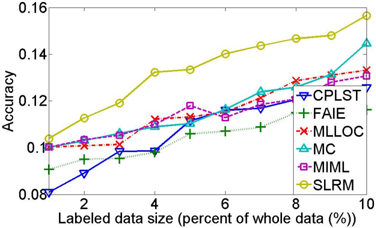
(a) AUC



(b) Macro-F1



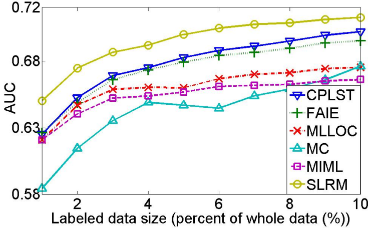
(c) Micro-F1



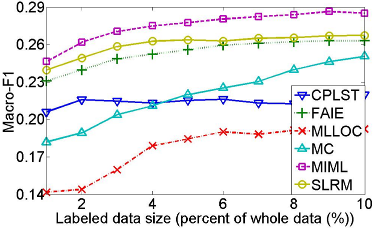
(d) Accuracy

Figure: Comparison of seven methods under varying the labeled data sizes on *Corel5K*.

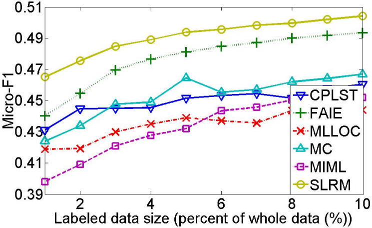
Experiments on Real Data



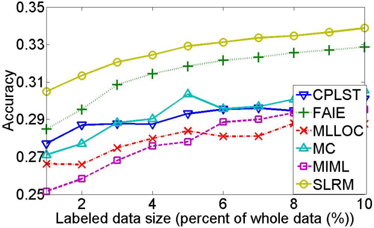
(a) AUC



(b) Macro-F1



(c) Micro-F1



(d) Accuracy

Figure: Comparison of seven methods under varying the labeled data sizes on *SUN*.

Experiments on Real Data

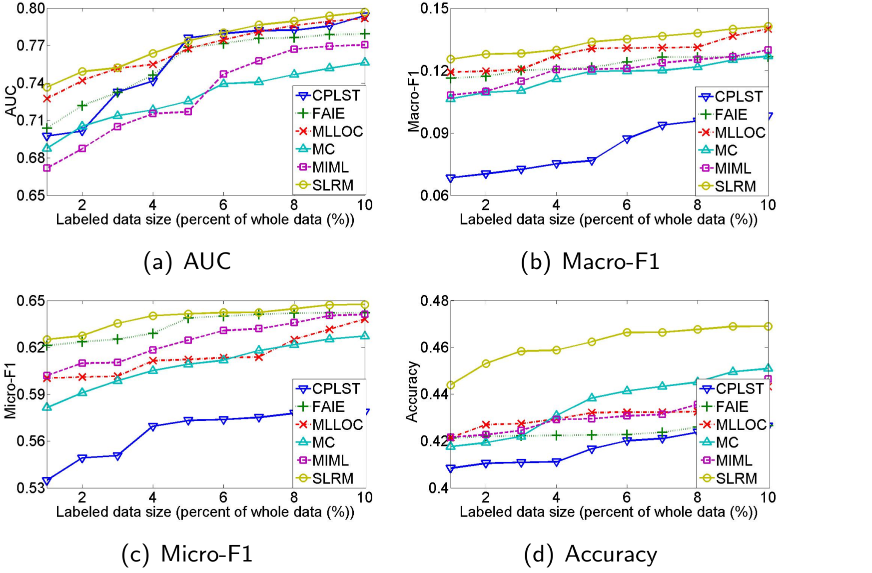
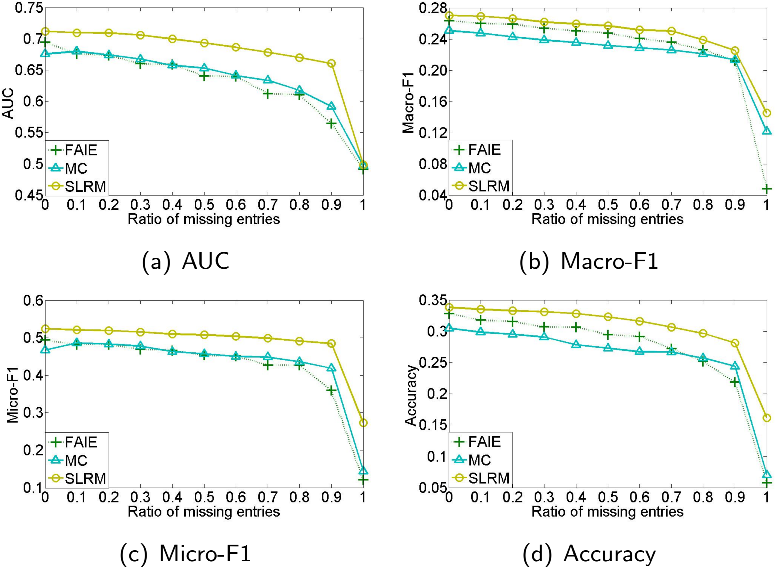


Figure: Comparison of seven methods under varying the labeled data sizes on *Mediamill*.

Experiments on Real Data

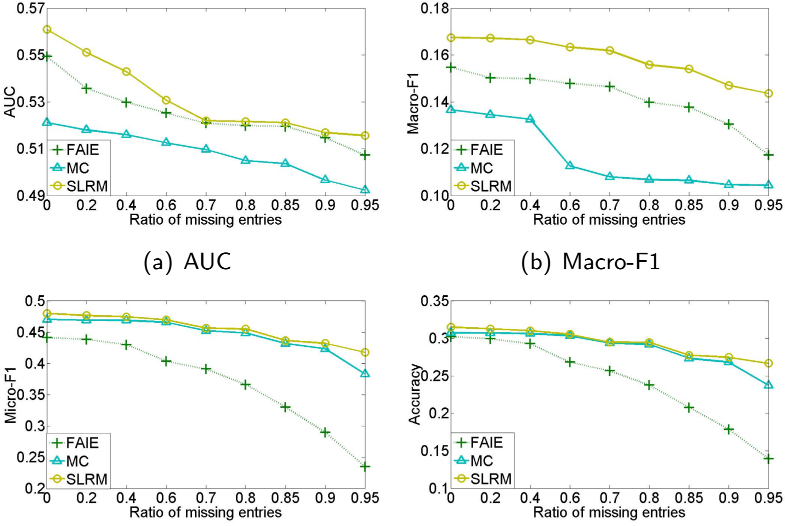


(a) AUC (b) Macro-F1

(c) Micro-F1 (d) Accuracy

Figure: Comparison results under varying the ratio of missing entries in label matrix (Y) *SUN*.

Experiments on Real Data



(a) AUC

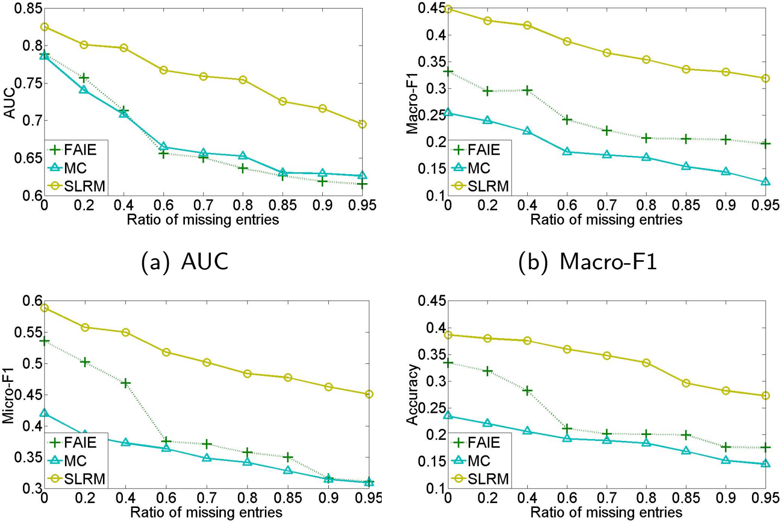
(b) Macro-F1

(c) Micro-F1

(d) Accuracy

Figure: Comparison results under varying the ratio of missing entries in label matrix (Y) *CAL500*.

Experiments on Real Data



(a) AUC

(b) Macro-F1

(c) Micro-F1

(d) Accuracy

Figure: Comparison results under varying the ratio of missing entries in label matrix (Y) *MSRC*.

Experiments on Real Data

TRUE:	flower	TRUE:	aeroplane, grass, sky	TRUE:	dog, grass, tree, body	TRUE:	building, sky, road, tree
SLRM:	flower	SLRM:	aeroplane, grass, sky	SLRM:	dog, grass, face, body	SLRM:	building, car, road, tree
CPLST:	grass	CPLST:	building, sky, tree	CPLST:	cow, grass, tree, sky	CPLST:	building, grass, tree, sky
FAIE:	grass	FAIE:	building, road, sky	FAIE:	grass, body, face, road	FAIE:	building, sky, road, tree
MLLOC:	grass	MLLOC:	grass, road, sky	MLLOC:	grass, tree, road, sky	MLLOC:	grass, sky, road, tree
MC:	grass	MC:	building, grass, sky	MC:	building, grass, road, sky	MC:	building, sky, road, grass
MIML:	grass	MIML:	building, road, sky	MIML:	cow, grass, tree, sky	MIML:	building, sky, grass, tree

Figure: Image label prediction examples from MRSC data.

Further Comparison

- Li et al. proposed a Conditional Restricted Boltzmann Machines model to characterize the label correlations by introducing a hidden level on the label level, and model the conditional marginal distribution of the label according to the observed input feature information.
- We test the effect of labeled data size for *Corel5K*. We also evaluate the effect of the ratio of missing labels on SLRM and CRBM for *CAL500* (the large cardinality)
- These results confirm that the semi-supervised strategy is helpful to mine the intrinsic structure from both labeled and unlabeled data and improve the final prediction performance.
- The results demonstrate that the low-rank term is more proper to determine the label correlations than the strategy adopted in CRBM.

Further Comparison

Method	Labeled data size (percentage of whole data (%))									
	1	2	3	4	5	6	7	8	9	10
CRBM	0.4623	0.4709	0.4747	0.4783	0.4909	0.4984	0.5161	0.5199	0.5272	0.5326
SLRM	0.5241	0.5309	0.5319	0.5388	0.5503	0.5539	0.5608	0.5629	0.5730	0.5762
CRBM	0.0157	0.0206	0.0243	0.0262	0.0294	0.0312	0.0317	0.0319	0.0331	0.0336
SLRM	0.0334	0.0345	0.0372	0.0386	0.0412	0.0424	0.0439	0.0457	0.0491	0.0497
CRBM	0.1649	0.1751	0.1790	0.1845	0.1864	0.1913	0.1967	0.2097	0.2151	0.2171
SLRM	0.2261	0.2273	0.2341	0.2383	0.2513	0.2550	0.2562	0.2592	0.2640	0.2700
CRBM	0.0891	0.0933	0.0961	0.1021	0.1087	0.1102	0.1164	0.1183	0.1210	0.1212
SLRM	0.1039	0.1127	0.1192	0.1323	0.1334	0.1402	0.1437	0.1467	0.1480	0.1566

Comparison of CRBM and SLRM under varying the labeled data sizes on *Corel5K*. 1. AUC; 2. Macro-F1; 3. Micro-F1; 4. Accuracy.

Further Comparison

Method	Ratio of missing labels (%)								
	0	20	40	60	70	80	85	90	95
CRBM	0.5417	0.5355	0.5253	0.5186	0.5152	0.5127	0.5110	0.5067	0.5012
SLRM	0.5585	0.5511	0.5429	0.5307	0.5219	0.5216	0.5211	0.5169	0.5156
CRBM	0.1410	0.1358	0.1302	0.1285	0.1221	0.1185	0.1122	0.1101	0.1092
SLRM	0.1655	0.1653	0.1646	0.1614	0.1600	0.1539	0.1521	0.1451	0.1417
CRBM	0.4463	0.4437	0.4373	0.4259	0.4039	0.3939	0.3818	0.3694	0.3453
SLRM	0.4818	0.4786	0.4764	0.4717	0.4585	0.4570	0.4385	0.4340	0.4196
CRBM	0.3039	0.2965	0.2859	0.2533	0.2422	0.2215	0.2109	0.2037	0.1859
SLRM	0.3087	0.3083	0.3078	0.3063	0.3003	0.2988	0.2871	0.2784	0.2643

Comparison of CRBM and SLRM under varying the ratio of missing entries in label matrix (Y) on *CAL500*. 1. AUC; 2. Macro-F1; 3. Micro-F1; 4. Accuracy.

Summary

- In order to tackle the multi-label classification problems, in this paper, we have proposed and developed a new model SLRM to identify an effective mapping function from feature space to label space.
- The proposed SLRM model can capture the label correlations by enforcing nuclear norm regularization on mapping function.
- SLRM also makes use of amounts of unlabeled data to smooth the mapping function by considering the intrinsic geometric structure among.
- Th extension of the current linear mapping to a non-linear one, i.e. looking for a function $U(\star)$, such that $Y = U(X)$. To make the problem tractable, one potential approach is to consider the finite-order-polynomial based approximation $U(X) = p_k(X)U$ where $p_n(X) = [X^0, X^1, \cdots, X^n]$ indicates the basis of polynomial respect to X with order up to k.
- It is also interesting to design model to automatically predict the number of positive labels for the new data.