Semisupervised Multilabel Learning With Joint Dimensionality Reduction

Tingzhao Yu and Wensheng Zhang

***Abstract*—Mutlilabel classification arises in various domains including computer vision and machine learning. Given a single instance, multilabel classification aims to learn a set of labels simultaneously. However, existing methods fail to address two key problems: 1) exploiting correlations among instances and 2) reducing computational complexity. In this letter, we propose a new semisupervised multilabel classification algorithm with joint dimensionality reduction. First, an elaborate matrix is designed for evaluating instance similarity; thus, it can take both labeled and unlabeled instances into consideration. Second, a linear dimensionality reduction matrix is added into the framework of multilabel classification. Besides, the dimensionality reduction matrix and the objective function can be optimized simultaneously. Finally, we design an efficient algorithm to solve the dual problem of the proposed model. Experiment results demonstrate that the proposed method is effective and promising.**

***Index Terms*—Alternating method, dimensionality reduction, dual problem, multilabel classification, semisupervised learning.**

I. Introduction

MULTILABEL classification is a generalization of multiclass classification. It allows to assign a set of labels to a single instance for accurate description. Multilabel classification has many potential applications in text classification [1], [2], scene classification [3], video segmentation [4], and music emotion classification [5], [6].

Existing multilabel classification methods can be divided into two categories [7], [8]: 1) methods based on problem transformation and 2) methods based on algorithm adaptation. Methods based on problem transformation transform the multilabel classification problems into existing well-established problems. These methods include: transforming into one or more independent single-label classification problem [3], [9], transforming into a chain of single-label classification problem [10], transforming into a label ranking problem [11], and transforming into an ensemble of multiclass classification problem [12], [13], [14]. Methods based on algorithm adaptation adapt or extend the existing state-of-the-art methods to solve multilabel problems explicitly. These methods include: adapting boosting [1], adapting decision tree [15], adapting support vector machine (SVM) [16], adapting maximum a posterior (MAP) principle [17], adapting maximum entropy principle [18], adapting back propagation (BP) neural networks [19], and adapting kNN [20].

Recently, researches on multilabel classification based on semisupervised learning arise. For example, semi-supervised multi-label sylvester equation (SMSE) [21] constructs two graphs on instance level and category level, respectively. The graph on instance level is defined based on both labeled and unlabeled instances, while the graph on category level is built on all categories. TRANSductive (TRANS) [22] learns a subspace representation of the labeled and unlabeled inputs, while simultaneously training a supervised large-margin multilabel classifier. Semi-supervised low-rank mapping (SLRM) [23] takes advantages of the nuclear norm regularization to capture the label correlations, while at the same time exploits manifold regularizer to capture the intrinsic structure among data.

However, these methods have difficulty in dealing with high-dimensional data. In this letter, we propose a new semisupervised multilabel classification algorithm with joint dimensionality reduction. Our work is mainly inspired by dimensionality reduction multi-label learning (DRMLL) [24]. DRMLL has studied a joint learning framework of dimensionality reduction and multilabel classification, but they take no use of relationships among instances. The contributions of this letter are summarized as follows.

1) Introduce an elaborate similarity matrix which can measure the similarity among all instances.
2) Introduce a linear dimensionality reduction matrix for joint learning; thus, it can deal with high-dimensional data.
3) Design an efficient algorithm to solve the dual problem, in which the objective function and dimensionality reduction matrix can be optimized simultaneously.

Note that semi-supervised dimension reduction-multi-label classification (SSDR-MC) [25] has already addressed semisupervised learning, dimensionality reduction, and multilabel learning simultaneously. It consists of two terms: 1) data reconstruction error $\|\mathbf{x}_i - \sum_j \mathbf{W}_{i,j}\mathbf{x}_i\|^2$ and 2) label reconstruction error $\|\mathbf{f}_i - \sum_j \mathbf{W}_{i,j}\mathbf{f}_i\|^2$, where $\mathbf{x}_i$ is the instance, $\mathbf{f}_i$ is the predicted label corresponding to instance $\mathbf{x}_i$, and $\mathbf{W}_{i,j}$ is the weight matrix for the dimensional reduction. In this letter, $\mathbf{x}$ represents instance, $\mathbf{W}$ represents the weight matrix, $\mathbf{w}_l$ is the lth column of $\mathbf{W}$, $\mathbf{Q}$ represents the dimension reduction matrix, and $\mathbf{y}$ represents the ground truth label. Variables with hat correspond to data with labels. We differ from SSDR-MC in four aspects.

1) SSDR-MC aims to minimize the data reconstruction error, while we deal with prediction error $\|\hat{\mathbf{y}}_l - \hat{\mathbf{X}}\mathbf{Q}\mathbf{w}_l\|^2$.

Manuscript received December 18, 2015; revised April 05, 2016; accepted April 13, 2016. Date of publication April 14, 2016; date of current version April 28, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaodong He.

The authors are with the Institute of Automation, Chinese Academy of Sciences, 100190 Beijing, China (e-mail: yutingzhao2013@ia.ac.cn; zhangwenshengia@hotmail.com).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2016.2554361

2) SSDR-MC explores the semisupervised information via data reconstruction error, while we introduce an additional part $\sum_{i,j=1}^{n} e_{i,j}\|\mathbf{Q}^T\mathbf{x}_i - \mathbf{Q}^T\mathbf{x}_j\|_2^2$ to measure the similarity among all instances, and $e_{i,j}$ is the weight.
3) SSDR-MC learns a matrix $\mathbf{W}$ for multilabel classification and dimensionality reduction, while we learn two matrices: 1) $\mathbf{W}$ for classification and 2) $\mathbf{Q}$ for dimensionality reduction, respectively.
4) SSDR-MC considers least square loss, while we consider hinge loss $\{1 - \hat{y}_{il} f(\hat{\mathbf{x}}_i)\}_+$.

II. Proposed Model

Our main idea is to design a semisupervised multilabel classification model with joint linear dimensionality reduction. We aim to design a transform matrix which can reduce the dimension of the original instance and at the same time preserve the inherent property. We define the linear dimensionality reduction matrix as $\mathbf{Q} \in \mathbb{R}^{k\times d}$, where k is the original dimension and d is the transformed dimension. Given an instance $\mathbf{x} \in \mathbb{R}^k$, we consider t linear classifiers $f_l : \mathbf{x} \to f_l(\mathbf{x}) = \boldsymbol{w}_l^T\mathbf{Q}^T\mathbf{x}, l = 1, \ldots, t$ for t labels, where $\boldsymbol{w}_l^T \in \mathbb{R}^d$ is the weight vector corresponding to the lth label. Our framework of semisupervised multilabel classification with joint dimensionality reduction is defined as

$$\min_{f,\mathbf{Q}} \sum_{l=1}^{t}\left(\sum_{i=1}^{nl}\mathcal{L}(\hat{y}_{il}, f_l(\hat{\mathbf{x}}_i), \mathbf{Q}) + \lambda\Omega(f_l)\right) + \gamma\Psi(\mathbf{Q}, \mathbf{X}) \quad (1)$$

where $\mathcal{L}$ is the loss function and it is related to the ground truth label $\hat{y}_{il}$, the predicted label $f_l(\hat{\mathbf{x}}_i)$, and the dimensionality reduction matrix $\mathbf{Q}$. Ω controls the complexity of f_l and Ψ is a measure of similarity between instances related to the transform matrix $\mathbf{Q}$. $\mathbf{X}$ is the data set with both labeled and unlabeled instances, and nl is the number of labeled instances. λ and γ are two balanced terms.

A. Choice of Loss Function

There are mainly three loss functions frequently used in real applications: 1) least square loss; 2) logistic loss; and 3) hinge loss. We consider hinge loss. Hinge loss is defined as

$$\mathcal{L}(\hat{y}_{il}, f(\hat{\mathbf{x}}_i)) = \{1 - \hat{y}_{il} f(\hat{\mathbf{x}}_i)\}_+ \quad (2)$$

where $\{\cdot\}_+ = \max\{0, \cdot\}$. The loss function of our model is

$$\sum_{l=1}^{t}\left(\sum_{i=1}^{nl}\mathcal{L}(\hat{y}_{il}, f(\hat{\mathbf{x}}_i), \mathbf{Q})\right) = \sum_{l=1}^{t}\sum_{i=1}^{nl}\{1 - \hat{y}_{il}\mathbf{w}_l^T\mathbf{Q}^T\hat{\mathbf{x}}_i\}_+. \quad (3)$$

B. Controller of Complexity

The complexity of function f_l can be measured by the linear weight vector w_l. For example, if we want the weight vector to be sparse, the 1-norm $\|\boldsymbol{w}_l\|_1 = \sum_{i=1}^{d}|w_{il}|$ is often used. If we want the weight vector to be smooth, the 2-norm $\|\boldsymbol{w}_l\|_2 = (\sum_{i=1}^{d} w_{il}^2)^{\frac{1}{2}}$ is usually the ideal choice. Other normally used regularizer is p-norm $\|\boldsymbol{w}_l\|_p = (\sum_{i=1}^{d} w_{il}^p)^{\frac{1}{p}}$. For simplicity, we choose squared 2-norm, thus

$$\Omega(f) = \sum_{l=1}^{t}\sum_{i=1}^{d} w_{il}^2 = \sum_{l=1}^{t}\|\boldsymbol{w}_l\|_2^2 = \|\mathbf{W}\|_F^2. \quad (4)$$

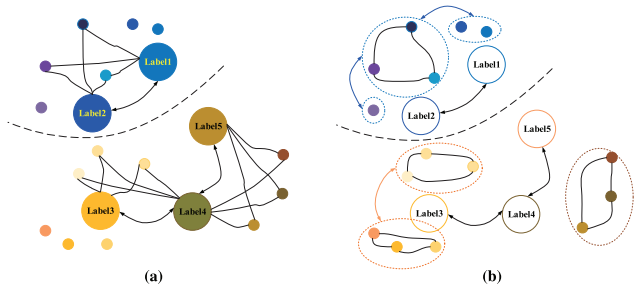

Fig. 1. (a) Traditional multilabel learning method, which exploits the correlation among all labels while ignores the unlabeled instances. Instances with labels are connected to their corresponding labels by solid lines, and the correlations among labels are represented by double-headed arrows. (b) New semisupervised multilabel learning method. It takes the advantage of all instances (both labeled and unlabeled). Similar instances are clustered into small sets and the small sets are connected by double-headed arrows.

C. Measurement of Similarity

Given two instances $\mathbf{x}_i$, $\mathbf{x}_j$ and the corresponding linear dimensionality reduction matrix $\mathbf{Q}$, if these two points are close in their original feature space, then the corresponding transformed data $\mathbf{Q}^T\mathbf{x}_i$, $\mathbf{Q}^T\mathbf{x}_j$ are required to be close to each other. Suppose all of the instances construct a graph. Each vertex is corresponding to one instance and the edge between vertexes is described by the similarity between instances. Typically, if two instances are connected, the edge weight is defined as $e_{i,j} = \exp(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma})$. Then, the measurement of similarity among instances in transformed space is defined by

$$\Psi(\mathbf{Q}) = \sum_{i,j=1}^{n} e_{i,j}\|\mathbf{Q}^T\mathbf{x}_i - \mathbf{Q}^T\mathbf{x}_j\|_2^2 \quad (5)$$

where n is the number of all instances. Furthermore, if we define an edge matrix $\mathbf{E}$, and a diagonal matrix $\mathbf{D}$, where

$$\mathbf{E} = \begin{cases} e_{i,j}, & \text{if } i,j \text{ is connected} \\ 0, & \text{otherwise} \end{cases},$$

$$\mathbf{D} = \begin{cases} \sum_{i=1}^{n} e_{i,j}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

then (5) is summarized as

$$\Psi(\mathbf{Q}) = \mathrm{tr}((\mathbf{XQ})^T\mathbf{S}(\mathbf{XQ})) \quad (6)$$

where $\mathbf{S} = \mathbf{D} - \mathbf{E}$ is the Laplacian matrix. A more detailed explanation can be found in Fig. 1.

D. New Model

Substituting the loss term (3), the complexity term (4), and the semisupervised term (5) in our framework (1), we get the final model, which is given by

$$\begin{aligned}\min_{f,\mathbf{Q}} &\sum_{l=1}^{t}\left(\sum_{i=1}^{nl}\mathcal{L}(\hat{y}_{il}, f(\hat{\mathbf{x}}_i), \mathbf{Q})\right) + \lambda\Omega(f) + \gamma\Psi(\mathbf{Q}) \\ &= \min_{f,\mathbf{Q}} \sum_{l=1}^{t}\sum_{i=1}^{nl}\{1 - \hat{y}_{il}\mathbf{w}_l^T\mathbf{Q}^T\hat{\mathbf{x}}_i\}_+ + \lambda\sum_{l=1}^{t}\|\mathbf{w}_l\|_2^2 \\ &\quad + \gamma\sum_{i,j=1}^{n} e_{i,j}\|\mathbf{Q}^T\mathbf{x}_i - \mathbf{Q}^T\mathbf{x}_j\|_2^2. \end{aligned} \quad (7)$$

TABLE I
PREDICTIVE PERFORMANCE AND COMPARISON ON NINE REAL DATA SETS OF FIVE ALGORITHMS (WITH *HALF* TRAINING INSTANCES LABELED)

Criterion	Algorithm	Image	Scene	Yeast	Slashdot	Enron	Language	Bibtex	CAL500	Corel5k	Simulate	Average
Hamming loss	SVM	0.3766	0.1072	0.2031	0.0186	0.0607	**0.1450**	**0.0127**	0.1379	0.1379	0.2010	0.1401
	DRMLL	0.3550	0.1070	0.2029	0.0187	**0.0601**	0.1462	0.0141	**0.1358**	0.0098	0.2120	0.1262
	ML-kNN	0.3486	0.1058	0.2020	0.9695	0.9757	0.9165	0.9984	0.9687	0.9999	0.9080	0.7393
	LIFT	0.3440	**0.0997**	0.2029	0.0186	0.0618	0.1559	0.0132	0.1399	0.0097	**0.1989**	0.1244
	Proposed	**0.3418**	0.1052	**0.1977**	**0.0183**	**0.0601**	0.1518	0.0139	0.1378	**0.0094**	0.2020	**0.1238**
One error	SVM	0.8350	0.2559	0.2399	0.3322	0.7621	**0.1826**	**0.3972**	0.1238	**0.1238**	0.3350	**0.3588**
	DRMLL	0.8430	0.2684	**0.2377**	**0.3283**	0.7821	0.1891	0.5849	0.1298	0.6740	0.3850	0.4422
	ML-kNN	0.8610	0.2759	0.2595	0.3294	0.7877	0.2065	0.6183	0.1188	0.7040	0.4350	0.4596
	LIFT	**0.8100**	**0.2166**	0.2443	0.3316	**0.7151**	0.2239	0.4163	0.1188	0.7020	0.3400	0.4119
	Proposed	**0.8100**	0.2584	0.2399	0.3339	0.7792	**0.1826**	0.5089	**0.1139**	0.6740	**0.3250**	0.4226
Coverage	SVM	2.8900	0.5510	6.5322	**0.8270**	**23.2721**	**47.4609**	**26.9857**	**132.7673**	132.7673	1.0550	37.5109
	DRMLL	2.9630	0.5602	6.5060	1.0584	24.5399	48.6043	47.1312	132.8903	133.6100	1.1100	39.8973
	ML-kNN	2.8430	0.6254	6.5322	1.0544	26.8262	47.6239	60.7698	133.0743	118.1600	1.2150	39.8724
	LIFT	**2.8170**	**0.4682**	6.5671	1.0589	25.8775	48.5913	27.2374	132.8069	123.1560	**0.9900**	36.9570
	Proposed	2.8550	0.5560	**6.4624**	1.1914	23.4630	48.5065	39.0545	134.7822	**110.5620**	1.0200	**36.8453**
Ranking loss	SVM	0.6453	0.0894	0.1777	**0.0341**	**0.2339**	**0.1810**	**0.0909**	0.1861	0.1861	0.1871	0.2012
	DRMLL	0.6617	0.0917	0.1775	0.0383	0.2535	0.1970	0.1842	**0.1860**	0.1544	0.2025	0.2147
	ML-kNN	0.6441	0.1036	0.1777	0.0395	0.2767	0.1920	0.2388	0.1937	0.1373	0.2333	0.2237
	LIFT	**0.6254**	**0.0728**	0.1783	0.0374	0.2694	0.1919	0.0966	0.1877	0.1374	0.1846	**0.1982**
	Proposed	0.6322	0.0801	**0.1734**	0.0414	0.2445	0.1949	0.1424	0.1931	**0.1269**	**0.1808**	0.2009
Average precision	SVM	0.3997	0.8462	0.7540	**0.6584**	0.2629	**0.6143**	**0.5356**	0.4693	0.2893	0.7790	0.5608
	DRMLL	0.3912	0.8403	0.7548	0.6536	0.2502	0.6016	0.3568	0.4893	0.2769	0.7552	0.5369
	ML-kNN	0.3901	0.8298	0.7493	0.6528	0.2440	0.5902	0.3115	0.4685	0.2537	0.7214	0.5211
	LIFT	0.4128	**0.8700**	0.7517	0.6533	**0.2686**	0.5889	0.5231	0.4853	0.2830	0.7864	0.5623
	Proposed	**0.4139**	0.8653	**0.7567**	0.6487	0.2400	0.6123	0.5257	**0.4895**	**0.2937**	**0.7877**	**0.5633**
CPU time	SVM	1.9518	2.4452	4.8364	7.5071	19.6619	77.3418	563.8428	3.0673	**3.1096**	0.9337	68.4697
	DRMLL	7.5070	11.9788	43.1901	17.5154	33.6068	80.5377	878.9733	18.8207	779.9969	3.6310	187.5758
	ML-kNN	**0.5951**	**1.2242**	**1.9534**	**1.6145**	**1.3583**	**1.9154**	**11.3093**	**1.8963**	5.9792	**0.2260**	**2.8071**
	LIFT	2.0040	64.3917	58.0759	54.8751	24.9635	37.5018	1905.8000	5.8685	662.9366	1.5872	281.8004
	Proposed	7.4611	23.9475	21.9127	36.3476	38.5803	130.6225	988.8365	11.4402	1465.1000	2.1703	272.6419
Iteration	Proposed	4	5	3	4	4	8	3	2	2	4	3.9000

III. PROPOSED ALGORITHM

In this section, we describe our algorithm for solving model (7). Model (7) is unconstrained. By transforming it into a constrained problem, we get

$$\min_{\substack{\{\mathbf{w}_l,\xi_i^l\}\\ \mathbf{Q}}} \sum_{l=1}^{t}\left(\frac{1}{2}\|\mathbf{w}_l\|^2 + C\sum_{i=1}^{nl}\xi_i^l\right)+\gamma\sum_{i,j=1}^{n} e_{i,j}\|\mathbf{Q}^T\mathbf{x}_i-\mathbf{Q}^T\mathbf{x}_j\|_2^2$$
$$\text{s.t.}\quad y_{il}(\mathbf{w}_l^T\mathbf{Q}^T\mathbf{x}_i+b_l)\geq 1-\xi_i^l,\xi_i^l\geq 0\quad \forall i,l,\quad \mathbf{Q}^T\mathbf{Q}=\mathbf{I} \tag{8}$$

where we introduce the slack variable ξ_i^l and intercept b_l for a clear corresponding with SVM [26]. Model (8) is the primal problem, but usually the primal problem is hard to handle. We consider the dual problem, which is given by

$$\min_{\mathbf{Q}}\max_{\boldsymbol{\alpha}^l}\sum_{l=1}^{t}\left(\sum_{i=1}^{nl}\alpha_i^l-\frac{1}{2}\left((\boldsymbol{\alpha}^l)^T\mathbf{Z}^l\hat{\mathbf{X}}\mathbf{Q}\mathbf{Q}^T\hat{\mathbf{X}}^T\mathbf{Z}^l\boldsymbol{\alpha}^l\right)\right)+\gamma\text{tr}\left((\mathbf{X}\mathbf{Q})^T\mathbf{S}(\mathbf{X}\mathbf{Q})\right)$$
$$\text{s.t.}\sum_{i=1}^{nl}y_i^l\alpha_i^l=0,\quad 0\leq\alpha_i^l\leq C\quad \forall l\ \mathbf{Q}^T\mathbf{Q}=\mathbf{I} \tag{9}$$

where $\mathbf{Z}^l$ is a diagonal matrix defined by $\mathbf{Z}^l=y_{il}$, if $i=j$, $\mathbf{Z}^l=0$ otherwise.

We assign an alternating algorithm to solve the proposed hinge loss involved dual problem. At each iteration, $\mathbf{Q}$ or $\boldsymbol{\alpha}^l$ is fixed. When $\mathbf{Q}$ is fixed, (9) can be simplified to

$$\max_{\boldsymbol{\alpha}^l}\sum_{l=1}^{t}\left(\sum_{i=1}^{nl}\alpha_i^l-\frac{1}{2}\left((\boldsymbol{\alpha}^l)^T\mathbf{Z}^l\hat{\mathbf{X}}\mathbf{Q}\mathbf{Q}^T\hat{\mathbf{X}}^T\mathbf{Z}^l\boldsymbol{\alpha}^l\right)\right)$$
$$\text{s.t.}\sum_{i=1}^{nl}y_i^l\alpha_i^l=0,\ 0\leq\alpha_i^l\leq C\quad\forall l \tag{10}$$

which can be regarded as t separated standard SVM problems. The kernel function is transformed from $\langle\mathbf{x}_i,\mathbf{x}_i\rangle$ into $\langle\mathbf{Q}^T\mathbf{x}_i,\mathbf{Q}^T\mathbf{x}_i\rangle$. The t standard SVMs can be solved simultaneously. When $\boldsymbol{\alpha}^l$ is fixed, $\mathbf{Q}$ can be solved via [27]

$$\max_{\mathbf{Q}:\mathbf{Q}^T\mathbf{Q}=\mathbf{I}} \operatorname{tr}\left(\mathbf{Q}^T\mathbf{L}\mathbf{Q}\right) \tag{11}$$

where $\mathbf{L} = \hat{\mathbf{X}}^T\hat{\mathbf{S}}\hat{\mathbf{X}} - \mathbf{X}^T\mathbf{S}\mathbf{X}$, $\hat{\mathbf{S}} = \sum_{l=1}^{t}\left(\mathbf{Z}^l\boldsymbol{\alpha}^l(\boldsymbol{\alpha}^l)^T\mathbf{Z}^l\right)$, and $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_{nl}]^T \in \mathbb{R}^{nl\times k}$ is the labeled instances. Note that $\hat{\mathbf{S}}$ is a measurement of similarity among labeled instances, while $\mathbf{S}$ is a measurement of similarity among all instances.

IV. Experiment Results

In this section, two groups of experiments on nine real data sets *image*,[1] *scene*,[2] *yeast*, *slashdot*,[3] *enron*,[2] *language*,[3] *bibtex*,[2] *CAL500*,[2] and *Corel5k*[2], and one simulated data set are designed to evaluate the performance of the proposed method. We construct a toy example with 600 instances of 294 dimensions and five labels. For comparison purpose, we use SVM [26], DRMLL [24], ML-kNN [20], and LIFT [28], and the five criterions [1], [8], [28] are *Hamming loss*, *One-error*, *Coverage*, *Ranking loss*, and *Average precision*. Besides, we choose the running *CPU time* as a criterion of time complexity.

A. Experiment Setup

In the first experiment, we demonstrate the performance of five algorithms on data sets referred before with half of the instances without their ground truth labels. We randomly split the training instances into two parts. One part preserves their labels while the other part not. In the second experiment, we demonstrate the tendency of the performance of five algorithms as the percentage of instances missing their labels increasing. In both section, the unlabeled instances are totally ignored during supervised learning, while in our semisupervised mode, both of the labeled and unlabeled instances together construct a matrix $\mathbf{X}$, which will be used to get $\mathbf{Q}$. The reduced dimension of $\mathbf{Q}$ is fixed by $d = 100$, and the balanced parameter γ is set to be 1.0, which means we pay equal attention to the performance term and the semisupervised term. All of the kernel functions involved in this letter is set to be "radial basis function (RBF)" kernel, and the parameters related to SVM is set to be default.

B. Results and Analysis

The experimental results are reported in Table I and Fig. 2.

It is obviously from Table I to see that the proposed method obtains the best overall performance among all the methods on *image*, *scene*, *yeast*, *CAL500*, and *Corel5k* data sets. We should note that the proposed method is indeed a generalization of SVM plus a similarity term and a dimensionality reduction term, and it is interesting to see that SVM conducts a overwhelming advantage than the proposed method on the left four data sets. The reason is that these four data sets, which SVM conducts the best performance, are all of high dimensionality. In our experiment, we fixed the reduced dimensionality to 100, which will lose some of the most discriminative information. While our method is also a generalization of DRMLL plus a semisupervised term, our method is superior than DRMLL when reducing to the same dimension of $d = 100$, due to the effect of unlabeled instances. The proposed method is suitable to handle data sets with more labels (e.g., *CAL500* and *Corel5k*), which means exploiting the similarity among instances do help to cluster instances into small subsets as illustrated in Fig. 1. Even though the proposed method is iterative, it will converge in an average of four iterations.

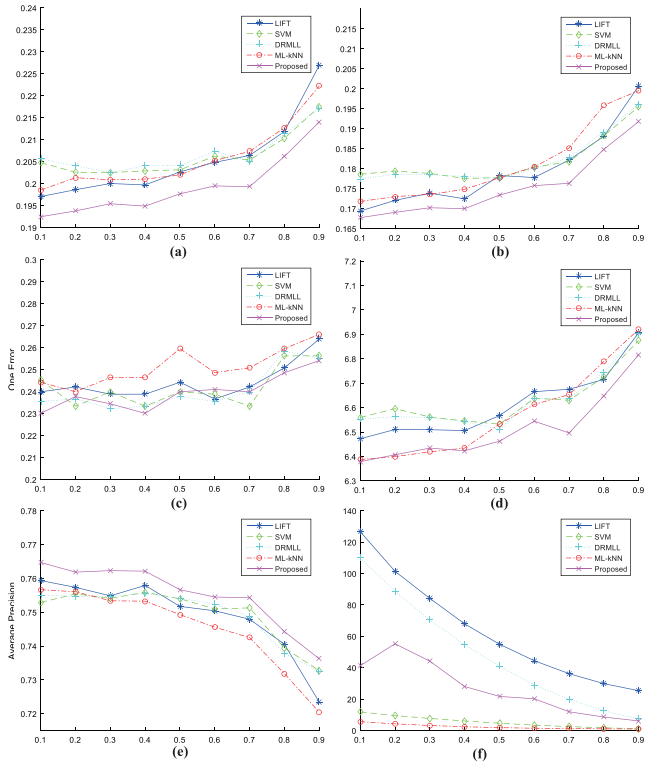

Fig. 2. Performance of five algorithms on *yeast* data set as the percentage of instances missing their label instances increasing. (a) Hamming loss. (b) Ranking loss. (c) One-error. (d) Coverage. (e) Average precision. (f) CPU time.

Fig. 2 gives the results on *yeast* data set as the percentage of instances without their ground truth labels increasing. All of the five algorithms decrease as the percentage of instances missing labels increasing. The proposed method (as be shown in pink color) has a much lower decreasing rate (with moderate time complexity) especially when the percentage are large, which evidently demonstrate that the proposed method is superior than the other methods within a semisupervised mode.

V. Conclusion

In this letter, we propose a new model for semisupervised multilabel learning with joint dimensionality reduction. Within this model, we introduce hinge loss into our framework and propose an efficient method based on the dual problem. Experiments on both simulated and real data sets illustrate the effectiveness of our new model and algorithm. In the future work, an adaptive dimensionality reduction matrix $\mathbf{Q}$ and a faster algorithm in solving the dimensionality reduction matrix $\mathbf{Q}$ will be considered. Besides, nonlinear dimensionality reduction techniques will be exploited.

[1][Online]. Available: http://cse.seu.edu.cn/people/zhangml
[2][Online]. Available: http://mulan.sourceforge.net/datasets-mlc.html
[3][Online]. Available: http://meka.sourceforge.net/#datasets

REFERENCES

[1] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Mach. Learn.*, vol. 39, no. 2, pp. 135–168, 2000.

[2] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Multilabel text classification for automated tag suggestion," in *Proc. Eur. Conf. Mach. Learn. Principles Pract. Knowl. Discov. Databases Discov. Challenge*, 2008, vol. 75, pp. 75–83.

[3] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, 2004.

[4] C. G. Snoek, M. Worring, J. C. Van Gemert, J. M. Geusebroek, and A. W. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proc. ACM Int Conf. Multimedia*, 2006, pp. 421–430.

[5] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions," in *Proc. 9th Int Conf. Music Inf. Retrieval*, 2008, vol. 8, pp. 325–330.

[6] C. Sanden and J. Z. Zhang, "Enhancing multi-label music genre classification through ensemble techniques," in *Proc. ACM SIGIR Int Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 705–714.

[7] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehous. Min.*, vol. 3, no. 3, pp. 1–13, 2007.

[8] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[9] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker, "Label ranking by learning pairwise preferences," *Artif. Intell.*, vol. 172, no. 16, pp. 1897–1916, 2008.

[10] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, pp. 333–359, 2011.

[11] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Mach. Learn.*, vol. 73, no. 2, pp. 133–153, 2008.

[12] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 1079–1089, Jul. 2011.

[13] J. Read, B. Pfahringer, and G. Holmes, "Multi-label classification using ensembles of pruned sets," in *Proc. 8th IEEE Int. Conf. Data Min.*, 2008, pp. 995–1000.

[14] C. Shi, X. Kong, S. Y. Philip, and B. Wang, "Multi-label ensemble learning," in *Machine Learning and Knowledge Discovery in Databases*. New York, NY, USA: Springer, 2011, pp. 223–239.

[15] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," in *Principles of Data Mining and Knowledge Discovery*. New York, NY, USA: Springer, 2001, pp. 42–53.

[16] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proc. Adv. Neural Inf. Process. Syst*, 2001, pp. 681–687.

[17] N. Ueda and K. Saito, "Single-shot detection of multiple categories of text using parametric mixture models," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2002, pp. 626–631.

[18] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, 2005, pp. 195–200.

[19] M. L. Zhang and Z. H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1338–1351, Oct. 2006.

[20] M. L. Zhang and Z. H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.

[21] G. Chen, Y. Song, F. Wang, and C. Zhang, "Semi-supervised multi-label learning by solving a sylvester equation," in *Proc. SIAM Conf. Data Min. (SDM)*, 2008, pp. 410–419.

[22] Y. Guo and D. Schuurmans, "Semi-supervised multi-label classification," in *Machine Learning and Knowledge Discovery in Databases*. New York, NY, USA: Springer, 2012, pp. 355–370.

[23] L. Jing, L. Yang, J. Yu, and M. K. Ng, "Semi-supervised low-rank mapping learning for multi-label classification," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1483–1491.

[24] S. Ji and J. Ye, "Linear dimensionality reduction for multi-label classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2009, vol. 9, pp. 1077–1082.

[25] B. Qian and I. Davidson, "Semi-supervised dimension reduction for multi-label classification," in *Proc. Nat. Conf. Artif. Intell*, 2010, vol. 10, pp. 569–574.

[26] V. N. Vapnik, *Statistical Learning Theory*. Hoboken, NJ, USA: Wiley, 1998, vol. 1.

[27] E. Kokiopoulou, J. Chen, and Y. Saad, "Trace optimization and eigenproblems in dimension reduction methods," *Numer. Linear Algebra Appl.*, vol. 18, no. 3, pp. 565–602, 2011.

[28] M. L. Zhang and L. Wu, "Lift: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.