Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Learning with Feature Network and Label Network Simultaneously

Yingming Li,[†] Ming Yang,[†] Zenglin Xu,[‡] Zhongfei (Mark) Zhang[†]
[†] College of Information Science & Electronic Engineering, Zhejiang University, China
[‡]School of Computer Science and Engineering, Big Data Research Center
University of Electronic Science and Technology of China
yingming@zju.edu.cn, cauchym@zju.edu.cn,
zenglin@gmail.com, zhongfei@zju.edu.cn

Abstract

For many supervised learning problems, limited training samples and incomplete labels are two difficult challenges, which usually lead to degenerated performance on label prediction. To improve the generalization performance, in this paper, we propose Doubly Regularized Multi-Label learning (DRML) by exploiting feature network and label network regularization simultaneously. In more details, the proposed algorithm first constructs a feature network and a label network with marginalized linear denoising autoencoder in data feature set and label set, respectively, and then learns a robust predictor with the feature network and the label network regularization simultaneously. While DRML is a general method for multi-label learning, in the evaluations we focus on the specific application of multi-label text tagging. Extensive evaluations on three benchmark data sets demonstrate that DRML outstands with a superior performance in comparison with some existing multi-label learning methods.

Introduction

With the research on tagging learning for decades (Nigam et al. 1998; Elisseeff and Weston 2001; Yu, Yu, and Tresp 2005; Hsu et al. 2009; Liu and Tsang 2015), recent years have witnessed the increasing applications of tagging learning in many fields ranging from social media searching to classification of medical reports due to its capability of improving data organization and management. Consequently, many tagging methods (Liu, Jin, and Yang 2006; Zhang and Zhou 2007; 2014; Li, Yang, and Zhang 2016) have been developed based on different requirements from different areas. However, most existing tagging methods assume that the amount of given training data is sufficient and the given training labels are complete. In contrast, for many supervised learning problems, they often face two challenges: limited training samples and incomplete training labels, which usually lead to degenerated performance on label prediction.

Given a limited amount of labeled training data and a very high-dimensional feature space, a common solution is to regularize a model by penalizing a specific norm of its parameters. The most commonly used norms in supervised learning are L_1 and L_2, which assume that model parameters are independent. However, dependencies between parameters usually exist in the real-world applications. For example, in biomedical domain, gene features have structured input since genes are organized as pathways; the learned model parameters (feature weights for a linear classifier) should be more effective by keeping the structural relationship between features. Further, dependencies can also be inferred from data, e.g., manifold-based feature graph can be used to regularize the model parameters and show its effectivity (Li and Li 2008). However, the feature network based on feature manifold only considers the positive correlation between features and ignores negative correlations between features. It is inappropriate since negative correlations also help to reduce the search space of the model parameters.

On the other hand, recent work, for example (Chen, Zheng, and Weinberger 2013), considers regularized learning with label network to mitigate the influence of incomplete training label set. It assumes that the given label set is incomplete and proposes a label network based on marginalized linear denoising autoencoder to exploit the relationship among tags. Consequently, a label network regularized learning method is presented to cope with the incomplete tagging problem. The proposed method significantly improves over the prior state-of-the-art. However, it still suffers from learning with limited training samples, which influence the generalization performance.

To improve the generalization performance of tagging, it is necessary to consider both feature network and label network. To achieve this goal, we propose to train robust predictors with feature network and label network simultaneously. In particular, we first learn a feature network and a label network with marginalized linear denoising autoencoders on feature set and label set, respectively. Take the learning of feature network for example, we learn the feature network by a marginalized linear denoising autoencoder, which is a one-layer linear denoising neural network, and train a network weight matrix $\mathbf{B}^x$ to make $\mathbf{B}^x\tilde{\mathbf{x}}$ approximate $\mathbf{x}$, where $\tilde{\mathbf{x}}$ is a corrupted version of sample $\mathbf{x} \in \mathbb{R}^d$ by random dropout corruption on each feature dimension. The learned network weight $\mathbf{B}^x_{ij}$ indicates the relationship between feature i and feature j.

Further, we present the Doubly Regularized Multi-Label learning (referred as DRML) model, which learns a robust predictor with the feature network and the label network regularization simultaneously. On the one hand, DRML extends the general multi-label learning with a feature network regularization; on the other hand, it exploits label network to enrich the incomplete training label set through a marginalized linear denoising autoencoder on the label set. In addition, it also leads to an optimization problem which is jointly convex and can be solved through alternative optimization with simple closed-form updates. Finally, we demonstrate the effectiveness and promise of DRML through extensive evaluations in three real data sets in comparison with the peer methods in the literature.

Related Work

Given a small training data set, network regularized learning is common to improve the generalization performance. In the cases of semi-supervised and unsupervised learning, graph Laplacian (Belkin and Niyogi 2001; 2003) is usually introduced to exploit the geometry of the marginal distribution. (Belkin, Niyogi, and Sindhwani 2006) proposes a graph Laplacian regularized geometric framework for data-dependent regularization. (Cai et al. 2008) proposes a Laplacian probabilistic semantic indexing method for topic modeling. In addition, graph Laplacian is also used for label propagation (Zhou et al. 2003; Wang and Zhang 2006).

Further, the previous work (Li and Li 2008; Sandler et al. 2008; Gu and Zhou 2009; Fei, Quanz, and Huan 2010) has demonstrated that learning with feature network can lead to improvement in generalization. (Li and Li 2008) introduces a network-constrained regularization for linear regression in order to incorporate the feature graph information into numerical data analysis. Based on the prior knowledge of features, (Sandler et al. 2008) presents a framework for regularized learning with feature network. (Gu and Zhou 2009) constructs a feature graph to explore the geometric structure of feature manifold.

Recently dropout training methods (Hinton et al. 2012; Srivastava et al. 2014) are proposed to combat overfitting by artificially corrupting the training data. By randomly dropping subsets of features at each iteration of a training process, (Hinton et al. 2012) reduces the influence of overfitting with dropout training. (Srivastava et al. 2014) introduces dropout training into the supervised neural network learning to improve the performance. In addition, (Vincent et al. 2008) trains robust denoising autoencoders with dropout noise. (Chen et al. 2012) proposes a marginalized denoising autoencoder to learn stacked features for supervised learning.

On the other hand, the given training data usually contain incomplete label, which belong to a special case of label noise problem. A number of denoising methods have been proposed for the label noise problem. Filtered preprocessing of the data and robust design of the algorithms are two common ways of tackling with label noise. The former focuses on removing the noise from the training set as much as possible (Van Hulse and Khoshgoftaar 2006) while the latter attempts to reduce the impact of the noise in the classification by designing robust algorithms (Lin and de Wang 2004). In addition, tag refinement is another effective way for noisy tagging in the literature (Wang et al. 2007). By investigating the robustness of SVMs against adversarial label noise, (Biggio, Nelson, and Laskov 2011) exploits a kernel matrix correction to improve the robustness of SVM. (Chen, Zheng, and Weinberger 2013) proposes to enrich the user tags with a label network learned from marginalized linear denoising autoencoder on training label set.

Learning with Feature Network and Label Network Simultaneously

In this section, we first introduce how to construct a novel feature network based on marginalized linear denoising autoencoder and to regularize the model parameters with this new feature network. Then a framework of label network regularization is also presented to enrich the incomplete label set. Further, we incorporate the label network regularization into the framework of learning with feature network. Consequently, a novel multi-label learning method, the Doubly Regularized Multi-Label (DRML), is developed to solve the problem of learning with feature network and label network simultaneously.

Learning with Feature Network

Regularization using feature network is especially appropriate for learning with high-dimensional feature space such as that encountered in medical text tagging where training data are very limited due to restricted corpus collection and expensive tagging process. Given a small set of training data, the local distances required by traditional manifold based methods may be difficult to be estimated accurately. Thus, manifold-based feature graphs are not so effective for learning with limited data. In this section, we first construct a new feature network with a marginalized linear denoising autoencoder and then propose a framework of feature network regularized learning.

Feature Network Construction In particular, we reduce the problem of constructing a feature network $\mathbb{N}_F$ into obtaining a feature structure relationship matrix, in which rows and columns correspond to features, and matrix elements indicate the relationship between features. Further, we describe how to estimate the feature relationship matrix from the data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ with a marginalized linear denoising autoencoder. Our intention is to explore the feature relationship network by a reconstruction mechanism. Let $x_i \in \mathbb{R}^d$ be the i-th sample, which is under consideration now. Imagine that we first corrupt this sample by random feature deletion with probability $p_x \geq 0$ and then reconstruct the original $\mathbf{x}_i$ from the "corrupted version" $\tilde{\mathbf{x}}_i$ with a feature relationship mapping matrix $\mathbf{B}^x : \mathbb{R}^d \to \mathbb{R}^d$. Here, for each sample $\mathbf{x}$ and dimension t, $p(\tilde{x}_t = 0) = p_x$ and $p(\tilde{x}_t = x_t) = 1 - p_x$. Consequently, we train this feature relationship mapping by minimizing the squared reconstruction loss,

$$\mathcal{L}(\mathbf{B}^x) = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{B}^x \tilde{\mathbf{x}}_i||^2 \qquad (1)$$

where $\mathbf{B}^x \in \mathbb{R}^{d\times d}$ can be considered as a feature relationship matrix that predicts the presence of features given the existing features in $\tilde{\mathbf{x}}$.

To reduce variance in $\mathbf{B}^x$, we take repeated samples of $\tilde{\mathbf{x}}$. In particular, we select each sample of the training set $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ and corrupt it m times by following the component-wise dropout distribution $p(\tilde{\mathbf{x}}|\mathbf{x})$. We can create corresponding corrupted examples $\tilde{\mathbf{x}}_{ij}$ (with $j = 1, \ldots, m$) for each $\mathbf{x}_i$ and construct a new data set $\tilde{\mathcal{D}}$ with $|\tilde{\mathcal{D}}| = mn$. Consequently, the above reconstruction loss in Eq.(1) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}||\mathbf{x}_i - \mathbf{B}^x\tilde{\mathbf{x}}_{ij}||^2 \tag{2}$$

where $\tilde{\mathbf{x}}_{ij} \sim p(\tilde{\mathbf{x}}_{ij}|\mathbf{x}_i)$.

When $m \to \infty$, we follow the weak law of larger numbers and rewrite the reconstruction loss as its expectation (Duda, Hart, and Stork 2001)

$$\begin{aligned}\mathcal{L}(\mathbf{B}^x) =& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[||\mathbf{x}_i - \mathbf{B}^x\tilde{\mathbf{x}}_i||^2\right]_{p(\tilde{\mathbf{x}}_i|\mathbf{x}_i)}\\ =& \frac{1}{n}\text{trace}\left(\mathbf{B}^x\mathbf{Q}^x\mathbf{B}^{x\top} - 2\mathbf{P}^x\mathbf{B}^{x\top} + \mathbf{X}\mathbf{X}^\top\right)\end{aligned} \tag{3}$$

where $\mathbf{P}^x = \sum_{i=1}^n \mathbf{x}_i\mathbb{E}\left[\tilde{\mathbf{x}}_i\right]^\top$, $\mathbf{Q}^x = \sum_{i=1}^n \mathbb{E}[\tilde{\mathbf{x}}_i]\mathbb{E}\left[\tilde{\mathbf{x}}_i\right]^\top + \mathbf{V}^x\left[\tilde{\mathbf{x}}_i\right]$, and

$$\begin{aligned}[\mathbf{Q}^\mathbf{x}]_{\alpha,\beta} &= \begin{cases}\mathbf{S}^x_{\alpha\beta}q_\alpha q_\beta & \text{if } \alpha \neq \beta\\ \mathbf{S}^x_{\alpha\beta}q_\alpha & \text{if } \alpha = \beta\end{cases}\\ [\mathbf{P}^x]_{\alpha\beta} &= \mathbf{S}^x_{\alpha\beta}q_\beta\end{aligned} \tag{4}$$

where $q_\alpha = q_\beta = 1 - p_x$, the variance matrix $\mathbf{V}^x\left[\tilde{\mathbf{x}}_i\right]_{p(\tilde{\mathbf{x}}_i|\mathbf{x}_i)} = p(1-p)\delta(\mathbf{x}_i\mathbf{x}_i^\top)$, and $\mathbf{S}^x = \mathbf{X}\mathbf{X}^\top$ is the covariance matrix of the uncorrupted data set. Here, $\delta(\cdot)$ denotes a operation that sets up all the entries except the diagonal to zero.

Then the solution of Eq.(3) can be expressed in a closed-form

$$\mathbf{B}^x = \mathbf{P}^x[\mathbf{Q}^x]^{-1} \tag{5}$$

Here, the matrix $\mathbf{B}^x$ encode the weights of the feature network $\mathbb{N}_F$. In particular, $\mathbf{B}^x_{ij}$ represents the similarity relationship of feature i and feature j.

Feature Network Regularization Learning Given labeled training data set $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$, for linear multi-label learning, we obtain the following L_2 norm regularized loss

$$\mathcal{L}\left(\mathbf{W}\right) = \frac{1}{n}\sum_{i}^{n}||\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \lambda||\mathbf{W}||_2^2, \tag{6}$$

where $\mathbf{W} \in \mathbb{R}^{l\times d}$ is the multi-label regression matrix.

Since a common semantic of the network assumes that positive linked features should have similar weight parameters and negative linked features should have dissimilar weight parameters, we penalize each feature's weight parameters by the squared amount it differs from the weighted average of its linked features. This way of network penalization gives us the following regularization term to incorporate in multi-label learning:

$$\begin{aligned}&\sum_{j=1}^{d}||\mathbf{W}_j - \mathbf{W}\mathbf{B}^x_j||^2\\ =&||\mathbf{W} - \mathbf{W}\mathbf{B}^x||^2\\ =&\text{trace}\left(\mathbf{W}\left(\mathbf{I} - \mathbf{B}^x\right)\left(\mathbf{I} - \mathbf{B}^x\right)^\top\mathbf{W}^\top\right)\\ =&\text{trace}\left(\mathbf{W}\mathbf{G}\mathbf{W}^\top\right)\end{aligned} \tag{7}$$

where $\mathbf{I} \in \mathbb{R}^{d\times d}$ and $\mathbf{G} = \left(\mathbf{I} - \mathbf{B}^x\right)\left(\mathbf{I} - \mathbf{B}^x\right)^\top$.

Further, by incorporating the feature network regularization in Eq.(7) into the framework of multi-label learning in Eq.(6), we have the following feature network regularized objective:

$$\begin{aligned}\mathcal{L}(\mathbf{W}) =& \frac{1}{n}\sum_{i=1}^{n}||\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \eta\text{trace}\left(\mathbf{W}\mathbf{G}\mathbf{W}^\top\right)\\ &+ \lambda||\mathbf{W}||_2^2\end{aligned} \tag{8}$$

where the parameters η and λ specify the strength of network and ridge regularization respectively.

Learning with feature network is referred to minimizing the above loss function, which introduces a feature network regularization into multi-label learning.

Learning with Label Network

Regularization using label network is very useful for multi-label learning where the given training labels are usually incomplete or incorrect since the network can be used to enrich the missing labels and correct the incorrect labels. In this section, similar to the part of learning with feature network, we first construct a label network $\mathbb{N}_L$ and then present a framework of label network regularized learning.

Label Network Construction We assume that a label relationship mapping $\mathbf{B}^y \in \mathbb{R}^{l\times l}$ can be learned with a marginalized linear denoising autoencoder to encode the corresponding weights of network $\mathbb{N}_L$. Further, the key idea is also based on a reconstruction mechanism that $\mathbf{B}^y$ should be able to predict the original tag vector $\mathbf{y}$ from the "corrupted version" $\tilde{\mathbf{y}}$. In particular, we first create a corrupted version by removing each tag in $\mathbf{y}$ with probability $p_y \geq 0$, and then $\mathbf{B}^y$ is learned to reconstruct the original tag vector $\mathbf{y}$ from the corrupted version $\tilde{\mathbf{y}}$ by minimizing the squared reconstruction error,

$$\mathcal{L}\left(\mathbf{B}^y\right) = \frac{1}{n}\sum_{i=1}^{n}||\mathbf{y}_i - \mathbf{B}^y\tilde{\mathbf{y}}_i||^2 \tag{9}$$

Here, $\mathbf{B}^y$ can be considered as a network weight matrix which represents the structural relationship among different labels.

Further, to reduce variance in $\mathbf{B}^y$, we take repeated samples of $\tilde{\mathbf{y}}$. In the limit (with infinitely corrupted versions of

$\mathbf{y}$), the expected loss function under the dropout distribution can be expressed as

$$\begin{aligned}\mathcal{R}(\mathbf{B}^y) =& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}^y\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} \\ =& \frac{1}{n}\text{trace}\left(\mathbf{B}^y\mathbf{Q}^y\mathbf{B}^{y\top} - 2\mathbf{P}^y\mathbf{B}^{y\top} + \mathbf{Y}\mathbf{Y}^\top\right)\end{aligned} \tag{10}$$

where $\mathbf{P}^y = \sum_{i=1}^n \mathbf{y}_i\mathbb{E}\left[\tilde{\mathbf{y}}_i\right]^\top$, $\mathbf{Q}^y = \sum_{i=1}^n \mathbb{E}[\tilde{\mathbf{y}}_i]\mathbb{E}\left[\tilde{\mathbf{y}}_i\right]^\top + \mathbf{V}^y\left[\tilde{\mathbf{y}}_i\right]$, and

$$\begin{aligned}\left[\mathbf{Q}^{\mathbf{y}}\right]_{\alpha,\beta} &= \begin{cases}\mathbf{S}^y_{\alpha\beta}q_\alpha q_\beta & \text{if } \alpha \neq \beta \\ \mathbf{S}^y_{\alpha\beta}q_\alpha & \text{if } \alpha = \beta\end{cases} \\ \left[\mathbf{P}^y\right]_{\alpha\beta} &= \mathbf{S}^y_{\alpha\beta}q_\beta\end{aligned} \tag{11}$$

where $q_\alpha = q_\beta = 1 - p_y$, the variance matrix $\mathbf{V}^y\left[\tilde{\mathbf{y}}_i\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} = p(1-p)\delta(\mathbf{y}_i\mathbf{y}_i^\top)$, and $\mathbf{S}^y = \mathbf{Y}\mathbf{Y}^\top$ is the covariance matrix of the uncorrupted tag set.

Then the solution of Eq.(10) can be expressed in a closed-form

$$\mathbf{B}^y = \mathbf{P}^y[\mathbf{Q}^y]^{-1} \tag{12}$$

Here, the matrix $\mathbf{B}^y$ encodes the weights of the network $\mathbb{N}_L$ and its element $\mathbf{B}^y_{ij}$ indicates the relationship between label i and label j.

Label Network Regularization Learning Unlike the way of feature network regularization, we want to enrich the user tags with the network of labels since the given tags are usually incomplete. The original tag vector $\mathbf{y}$ are improved by propagating the original labels with the learned network weight matrix $\mathbf{y} \to \mathbf{B}^y\mathbf{y}$. Consequently, we can reformulate the criterion of linear multi-label learning in Eq.(6) as follows:

$$\frac{1}{n}\sum_{i=1}^{n}||\mathbf{B}^y\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \lambda||\mathbf{W}||_2^2 \tag{13}$$

In fact, the above loss function can be interpreted as a cross-view learning objective when considering training data (samples with incomplete tags) as unlabeled multi-view data. In particular, it can be considered as a cross-view agreement from two sub-tasks: 1) training a classifier $\mathbf{x}_i \to \mathbf{W}\mathbf{x}_i$ that predicts the complete tag set from observations, and 2) enriching the existing incomplete tag vector with the network weight matirx $\mathbf{y}_i \to \mathbf{B}^y\mathbf{y}_i$.

However, the loss function in Eq.(13) only exploits a permanent label network weight matrix $\mathbf{B}^y$ and does not take the learning process of label network into account. Expecting that using the learning process of label network can obtain a better $\mathbf{B}^y$ for the label enrichment, we consider to use the loss of label network learning to guide the label network regularized multi-label learning,

$$\begin{aligned}\mathcal{L}\left(\mathbf{B}^y, \mathbf{W}; \mathbf{x}, \mathbf{y}\right) =& \frac{1}{n}\sum_{i=1}^{n}||\mathbf{B}^y\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \lambda||\mathbf{W}||_2^2 \\ &+ \gamma\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}^y\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)}\end{aligned} \tag{14}$$

where the parameter γ specifies the strength of label network regularization.

Learning with label network is referred to minimizing the above loss function, which enriches the existing tag vector with the label network weight matrix and introduces the label network learning into multi-label learning at the same time.

Learning with Feature Network and Label Network Simultaneously

Based on the above two network regularized learning frameworks presented in Eq.(8) and Eq.(14), we propose a new co-regularized method, minimizing the following objective,

$$\begin{aligned}&\mathcal{J}_{DRML}\left(\mathbf{B}^y, \mathbf{W}; \mathbf{x}, \mathbf{y}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}||\mathbf{B}^y\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \eta\text{trace}\left(\mathbf{W}\mathbf{G}\mathbf{W}^\top\right) \\ &\quad + \lambda||\mathbf{W}||_2^2 + \gamma\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}^y\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)}\end{aligned} \tag{15}$$

where $\lambda, \gamma, \eta \geq 0$ are regularization parameters. Since there are two different network regularizations in the objective, we call it the Doubly Regularized Multi-Label learning (DRML).

Optimization and Extensions The loss function in Eq.(15) can be efficiently optimized using coordinate descent. When $\mathbf{B}^y$ is fixed, the computation of regression matrix $\mathbf{W}$ can be solved in a closed form:

$$\mathbf{W} = \mathbf{B}^y\mathbf{Y}\mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + n\eta\mathbf{G} + n\lambda\mathbf{I}\right)^{-1} \tag{16}$$

where $\mathbf{G}$ can be computed following the definition in Eq.(7).

When $\mathbf{W}$ is fixed, the problem in Eq.(15) with respect to $\mathbf{B}^y$ can be reformulated as

$$\begin{aligned}\mathcal{R}\left(\mathbf{B}^y\right) =& \frac{1}{n}\text{trace}\left(\mathbf{B}^y\mathbf{Y}\mathbf{Y}^\top\mathbf{B}^{y\top} - 2\mathbf{B}^y\mathbf{Y}\mathbf{X}^\top\mathbf{W}^\top\right) \\ &+ \gamma\frac{1}{n}\text{trace}\left(\mathbf{B}^y\mathbf{Q}^y\mathbf{B}^{y\top} - 2\mathbf{P}^y\mathbf{B}^{y\top} + \mathbf{Y}\mathbf{Y}^\top\right)\end{aligned} \tag{17}$$

Similarly, when $\mathbf{W}$ is fixed, the solution of $\mathbf{B}$ in Eq.(17) can be obtained as follows:

$$\mathbf{B}^y = \left(\gamma\mathbf{P}^y + \mathbf{W}\mathbf{X}\mathbf{Y}^\top\right)\left(\gamma\mathbf{Q}^y + \mathbf{Y}\mathbf{Y}^\top\right)^{-1} \tag{18}$$

where $\mathbf{P}^y$ and $\mathbf{Q}^y$ can be computed analytically following Eq.(11).

Further, the loss in Eq.(15) is jointly convex with respect to $\mathbf{B}^y$ and $\mathbf{W}$. Consequently, it is guaranteed that the coordinate descent converges to the global minimum.

Experiments

We evaluate DRML on three standard multi-label benchmark data sets including two biomedical text data sets. All data sets are obtained from http://mulan.sourceforge.net/datasets-mlc.html.

Table 1: Statistics of the three data sets.

Data set	Examples	Labels	Features
Medical	978	45	1449
Bookmarks	10000	208	2150
Yeast	2417	14	103

Experimental Setup

In this section, we give a detailed descriptions of data sets, evaluation metrics, parameter setup, and baselines.

Datasets We have used three multi-label datasets, namely Medical, Bookmarks, and Yeast for experimentation purpose. Their statistics are described in Table 1.

Medical data set is formed by clinical free text, which is collected from the Cincinnati Children's Hospital Medical Center's Department of Radiology. The text is labeled with ICD-9-CM codes. For this experiment, we use a subset of about 978 labeled instances provided by **Mulan**[1].

Bookmarks data set is from Bibsonomy[2]. Bibsonomy is a social bookmarking and publication sharing system. Bookmarks contains metadata for bookmark items such as the URL of a web page and a description of the web page.

Yeast data set is formed by micro-array expression data and phylogenetic profiles. Each gene is associated with a set of functional classes. For this experiment, we use the whole set of 2417 labeled genes.

Evaluation Metric Three metrics, precision, recall, and F1 score, are often used to measure the performance of a tagging algorithm. Here, we also use them as our evaluation metrics. First, all the text data are labeled with the five most relevant tags (i.e., tags with the highest prediction value). Second, precision (P) and recall (R) are computed for each tag. The reported measurements are the average across all the tags. Further, both factors are combined in F1 score ($F1 = 2\frac{P*R}{P+R}$), which is reported separately. In all the metrics a higher value indicates a better performance.

Setup We use cross-validation to estimate the performance of different methods. On the Medical and Yeast data sets, we follow the experimental setup used in **Mulan**. Since there is no fixed split in the Bookmarks data set in **Mulan**, we use a fixed training set of 60% of the data, and evaluate the performance of our predictions on the fixed test set of 40% of the data.

Baselines To demonstrate how DRML improves the tagging performance in comparison with the state-of-the-art tagging methods, we compare it with the following representative tagging methods from the recent literature:

- LeastSquare (Bishop 2006).
- Regularized learning with feature graph Laplacian (referred to FGL) (Li and Li 2008).
- Low rank empirical risk minimization for multi-label learning (referred to LEML) (Yu et al. 2014) .
- FastTag, a model which exploits labels' relationship with marginalized linear denoising autoencoder regularization (Chen, Zheng, and Weinberger 2013).
- FastTag+FGL, which combines the merits of the above two methods by incorporating the feature graph Laplacian regularization into the FastTag method.

In addition, we also study various different configurations of the proposed algorithm:

- DRML ($\mathbf{B}^y = \mathbf{I}$): only using the proposed feature network regularization.
- DRML: using the proposed feature network and label network simultaneously.

In particular, for DRML, when we set up $\mathbf{B}^x = \mathbf{I}$, DRML reduces to FastTag.

Experimental Results

In Table 2, we summarize the precision, recall, and F1 score of the Medical, Bookmarks, and Yeast data sets, for LeastSquare, FGL, LEML, FastTag, FastTag+FGL, DRML ($\mathbf{B}^y = \mathbf{I}$), and DRML, respectively. On the task of multi-label text tagging, compared with DRML, though FGL exploits the feature network based on graph Laplacian, it only considers the positive correlation between features and ignores the negative correlations between features. In particular, FGL does not show obvious advantage comparing with LeastSquare and this suggests that there is a certain limitation with the constructed feature network by graph Laplacian. LEML models multi-label learning as a general empirical risk minimization problem with a low-rank constraint, while it cannot exploit the structure relationship between features. FastTag considers the training set as incomplete tagged data set, but it cannot take advantage of the feature network to mitigate the influence of limited training data. While FastTag+FGL incorporates the feature graph Laplacian regularization into FastTag and performs better than FastTag, it also only considers the positive correlations between features based on feature manifold. Consequently, from Table 2, we see that DRML ($\mathbf{B}^y = \mathbf{I}$) performs better than FGL, since the learned feature network not only considers the positive correlations between features, but also captures the negative correlations. Further, DRML performs better than leastSquare, LEML, FastTag, and FastTag+FGL as the F1 scores achieved by DRML are much higher than those achieved by the competing models in most cases.

Figure 1(a), Figure 1(b), and Figure 1(c) show the test F1 scores of FGL, LEML, FastTag, FastTag+FGL, DRML ($\mathbf{B}^y = \mathbf{I}$), and DRML as a function of the dropout level p of feature network on Medical, Bookmarks, and Yeast data sets, respectively. Herein, dropout level $p = 0$ corresponds to a regular L_2 norm regularization on feature parameters. The results show that DRML improves over the other standard predictors on almost all the cases.

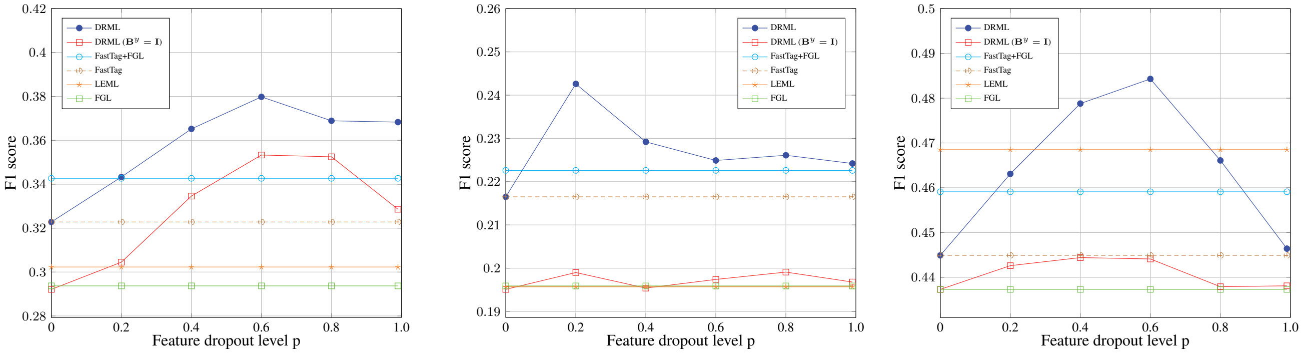
From Figure 1(a), Figure 1(b), and Figure 1(c), we observe that the F1 scores of the testing set for DRML and DRML ($\mathbf{B}^y = \mathbf{I}$) both increase when the dropout level increases at the start, which shows that it is helpful to use the feature network generated by marginalized linear denoising

[1] http://mulan.sourceforge.net/datasets-mlc.html

[2] http://www.bibsonomy.org

Table 2: Comparison of DRML and the competing models in terms of precision, recall, and F1 score on the three data sets.

	Medical			Bookmarks			Yeast		
Methods	precision	recall	F1	precision	recall	F1	precision	recall	F1
LeastSquares	0.1923	0.6077	0.2922	0.1559	0.2612	0.1952	0.4080	0.4712	0.4373
FGL	0.1934	**0.6103**	0.2937	0.1584	0.2565	0.1959	0.4123	**0.4764**	0.4420
LEML	0.2227	0.4703	0.3023	0.1627	0.2454	0.1957	0.4845	0.4535	0.4685
FastTag	0.2231	0.5837	0.3228	0.1861	0.2588	0.2165	0.4258	0.4659	0.4449
FastTag+FGL	0.2468	0.5606	0.3427	0.1941	0.2609	0.2226	0.4611	0.4572	0.4591
DRML ($\mathbf{B}^y = \mathbf{I}$)	0.2620	0.5419	0.3532	0.1875	0.2211	0.2029	0.4262	0.4636	0.4441
DRML	**0.2898**	0.5508	**0.3798**	**0.2334**	**0.2772**	**0.2534**	**0.5285**	0.4468	**0.4843**



(a) Medical.

(b) Bookmarks.

(c) Yeast.

Figure 1: The F1 score for the testing set as a function of feature dropout level p for FGL, LEML, FastTag, FastTag+FGL, DRML ($\mathbf{B}^y = \mathbf{I}$), and DRML, respectively.

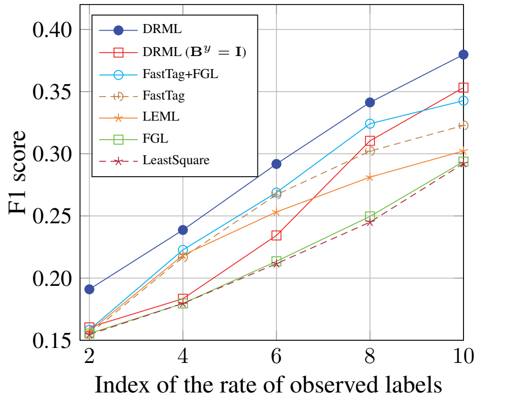


Figure 2: Performance in terms of F1 score as a function of the rate of the observed labels for each training document on the Medical data set. The i-th coordinate on the x-axis corresponds to the observed labels' rate $i \times 10\%$.

autoencoder to improve the tagging performance. It is also noted that after a certain point when the feature dropout level continues to increase, the performance may substantially drop. This is due to the *over dropout* issue, which loses much useful feature information and the leaned network is not accurate. In particular, we can see that the optimal values of dropout level p often exist in $[0.2, 0.6]$.

Figure 2 demonstrates the comparison between DRML and the competing models at different rates of observed labels in the training process. We gradually add the rate of observed labels. As we see from Figure 2, for training set with the label rate $r = 20\%$, DRML outperforms FGL and FastTag with about 4% gain. With the increase of the observed label rate, DRML continues to outperform the competing models with significant margins. In particular, DRML outperforms FastTag+FGL and DRML ($\mathbf{B}^y = \mathbf{I}$) obtains a better performance than FGL. It illustrates that the constructed feature network by marginalized linear denoising autoencoder is more effective than that by graph Laplacian. Although FastTag performs similarly to LeastSquare when the observed label rate is small, its performance improves fast with the increase of the observed label rate and it outperforms LeastSquare when the observed label rate is larger. This also shows the importance of learning with label network.

Conclusion

For many supervised learning problems, limited training samples and incomplete labels are two difficult challenges, which usually lead to degenerated performance on label prediction. To improve the generalization performance, in this paper, we first construct a feature network and a label network with marginalized linear denoising autoencoder in data feature set and label set, respectively, and then propose the

Doubly Regularized Multi-Label learning (DRML) by exploiting feature network and label network regularization simultaneously. Extensive evaluations on three benchmark data sets demonstrate that DRML outstands with a superior performance in comparison with some existing multi-label learning methods.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (NSFC No. 61672456) and Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis. Z. Xu was also supported by NSFC (No. 61572111, 61433014, 61440036), a 985 Project of UESTC (No.A1098531023601041) and Basic Research Projects of China Central Universities (No. ZYGX2014J058, A03012023601042).

References

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 585–591.

Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15(6):1373–1396.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* 7:2399–2434.

Biggio, B.; Nelson, B.; and Laskov, P. 2011. Support vector machines under adversarial label noise. In *ACML*, 97–112.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. Springer.

Cai, D.; Mei, Q.; Han, J.; and Zhai, C. 2008. Modeling hidden topics on document manifold. In *CIKM*, 911–920.

Chen, M.; Xu, Z. E.; Weinberger, K. Q.; and Sha, F. 2012. Marginalized denoising autoencoders for domain adaptation. In *ICML*.

Chen, M.; Zheng, A. X.; and Weinberger, K. Q. 2013. Fast image tagging. In *ICML*, 1274–1282.

Duda, R. O.; Hart, P. E.; and Stork, D. G. 2001. *Pattern Classification (2nd Ed)*. Wiley.

Elisseeff, A., and Weston, J. 2001. A kernel method for multi-labelled classification. In *NIPS*, 681–687.

Fei, H.; Quanz, B.; and Huan, J. 2010. Regularization and feature selection for networked features. In *CIKM*, 1893–1896.

Gu, Q., and Zhou, J. 2009. Co-clustering on manifolds. In *KDD*, 359–368.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* abs/1207.0580.

Hsu, D.; Kakade, S.; Langford, J.; and Zhang, T. 2009. Multi-label prediction via compressed sensing. In *AAAI*, 772–780.

Li, C., and Li, H. 2008. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 24(9):1175–1182.

Li, Y.; Yang, M.; and Zhang, Z. M. 2016. Multi-view representation learning: A survey from shallow methods to deep methods. *CoRR* abs/1610.01206.

Lin, C. F., and de Wang, S. 2004. Training algorithms for fuzzy support vector machines with noisy data. *Pattern Recognition Letters* 25:1647–1656.

Liu, W., and Tsang, I. W. 2015. Large margin metric learning for multi-label prediction. In *AAAI*, 2800–2806.

Liu, Y.; Jin, R.; and Yang, L. 2006. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*, 421–426.

Nigam, K.; McCallum, A.; Thrun, S.; and Mitchell, T. M. 1998. Learning to classify text from labeled and unlabeled documents. In *AAAI/IAAI*, 792–799.

Sandler, T.; Blitzer, J.; Talukdar, P. P.; and Ungar, L. H. 2008. Regularized learning with networks of features. In *NIPS*, 1401–1408.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Van Hulse, J., and Khoshgoftaar, T. M. 2006. Class noise detection using frequent itemsets. *Intelligent Data Analysis* 10:487–507.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML*, 1096–1103.

Wang, F., and Zhang, C. 2006. Label propagation through linear neighborhoods. In *ICML*, 985–992.

Wang, C.; Jing, F.; Zhang, L.; and Zhang, H.-J. 2007. Content-based image annotation refinement. In *CVPR*, 1–8.

Yu, H.-F.; Jain, P.; Kar, P.; and Dhillon, I. S. 2014. Large-scale multi-label learning with missing labels. In *ICML*, 593–601.

Yu, K.; Yu, S.; and Tresp, V. 2005. Multi-label informed latent semantic indexing. In *SIGIR*, 258–265.

Zhang, M., and Zhou, Z. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 40(7):2038–2048.

Zhang, M., and Zhou, Z. 2014. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* 26(8):1819–1837.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Schölkopf, B. 2003. Learning with local and global consistency. In *NIPS*, 321–328.