Probabilistic Multi-label Classification with Sparse Feature Learning

Yuhong Guo and **Wei Xue**
Department of Computer and Information Sciences
Temple University
Philadelphia, PA 19122, USA

Abstract

Multi-label classification is a critical problem in many areas of data analysis such as image labeling and text categorization. In this paper we propose a probabilistic multi-label classification model based on novel sparse feature learning. By employing an individual sparsity inducing ℓ_1-norm and a group sparsity inducing $\ell_{2,1}$-norm, the proposed model has the capacity of capturing both label interdependencies and common predictive model structures. We formulate this sparse norm regularized learning problem as a non-smooth convex optimization problem, and develop a fast proximal gradient algorithm to solve it for an optimal solution. Our empirical study demonstrates the efficacy of the proposed method on a set of multi-label tasks given a limited number of labeled training instances.

1 Introduction

Multi-label classification is a critical problem in many areas of data analysis, where each data instance can be assigned into multiple categories. For example, in image labeling [Zhou and Zhang, 2006] or video annotation [Qi *et al.*, 2007], a given scene usually contains multiple objects of interests. In text categorization [Schapire and Singer, 2000], a given article or webpage can be assigned into multiple topic categories. In gene and protein function prediction [Elisseeff and Weston, 2002], multiple functions are typically associated with each gene and protein. Due to its complex nature, the labeling process of a multi-label data set is typically more expensive or time-consuming comparing to single-label cases, since the annotator needs to evaluate each class label even when the positive labels appear in a very sparse pattern. To mitigate the needs hence the cost of collecting labeled data, learning effective multi-label classifiers from a small number of training instances thus is important to be investigated.

One straightforward approach for multi-label classification is to cast the multi-label learning problem as a set of independent single label classification problems [Lewis *et al.*, 2004; Chen *et al.*, 2007]. This simple method has the obvious drawback of ignoring useful correlation information between the predictions of multiple labels. Developing methods to exploit the label dependency information and capture shared prediction structures among the multiple labels is critical in multi-label learning. In the literature, many approaches have been proposed to address multi-label learning by either exploiting label dependencies [Elisseeff and Weston, 2002; Godbole and Sarawagi, 2004; Guo and Gu, 2011; Schapire and Singer, 2000; Petterson and Caetano, 2011], or capturing the common prediction structures of the multiple binary prediction tasks associated with the individual classes [Yan *et al.*, 2007; Zhang and Zhou, 2008; Yu *et al.*, 2005; Ji *et al.*, 2010]. But very few have taken both aspects into account for multi-label learning.

In this paper we propose a novel probabilistic multi-label classification model to simultaneously exploit both label dependency knowledge and shared prediction structures across labels based on sparse feature learning. Sparse feature learning has been effectively exploited in simultaneous multi-task learning problems by enforcing an $\ell_{2,1}$ norm [Argyriou *et al.*, 2006; Liu *et al.*, 2009; Obozinski *et al.*, 2006]. Different from these works which consider only common input features, our model contains two types of features: *structural label dependency features* associated with each individual single-label prediction task, and *common input features* which are shared across the multiple single-label prediction tasks. We first propose to learn sparse label dependency structures by associating an ℓ_1-norm regularization with the label dependency features, aiming to overcome possible overfitting issues. It induces a sparse conditional dependency network under probabilistic multi-label predictors. Then by adding another $\ell_{2,1}$-norm regularization over the input features, we formulate the overall probabilistic multi-label learning problem as a joint convex optimization problem with combined sparse norm regularizations, where an ℓ_1-norm is used for the sparse structural feature selection, and an $\ell_{2,1}$-norm is used for selecting the discriminative input features shared across multiple binary predictors. We develop a fast proximal gradient algorithm to solve the proposed optimization problem for an optimal solution. Our empirical results on a number of multi-label data sets demonstrate the efficacy of the proposed approach when the number of training instances is small, comparing to a few related probabilistic methods.

The remainder of the paper is organized as follows. We introduce the related work on multi-label learning in Section 2. The proposed sparse multi-label learning method is presented in Section 3. We then report the experimental results in Section 4 and conclude the paper in Section 5.

2 Related work

Multi-label classification has received increasing attention from machine learning community in recent years, due to its wide applications in practice. There is a rich body of work on multi-label learning in the literature. We provide a review to the most related methods in this section.

One simple approach for multi-label classification is to cast the multi-label learning problem as a set of independent single label classification problems [Lewis *et al.*, 2004; Chen *et al.*, 2007]. Such an approach however is unsatisfactory, since the different labels occurring in a multi-label classification problem are not independent. On the contrary, they often exhibit strong correlations or dependencies. Capturing these correlations in different manners have led to many advanced developments in multi-label classification.

A significant number of multi-label learning approaches have been proposed to exploit label dependencies in classification model formulation, including ranking based methods [Elisseeff and Weston, 2002; Schapire and Singer, 2000; Shalev-Shwartz and Singer, 2006; Fuernkranz *et al.*, 2008], pairwise label dependency methods [Zhu *et al.*, 2005; Petterson and Caetano, 2011], probabilistic classifier chains [Dembczynski *et al.*, 2010], large-margin methods [Guo and Schuurmans, 2011; Godbole and Sarawagi, 2004; Hariharan *et al.*, 2010], and probabilistic graphical models [Ghamrawi and McCallum, 2005; de Waal and van der Gaag, 2007; Bielza *et al.*, 2011; Zaragoza *et al.*, 2011; Guo and Gu, 2011]. Most of these methods however involve resource-consuming optimization procedures or extensive model-structure learning processes. On the other hand, another set of methods attempt to exploit the relationships between multiple binary classification models in multi-label learning by capturing their common prediction structures [Yan *et al.*, 2007; Zhang and Zhou, 2008; Yu *et al.*, 2005; Ji *et al.*, 2010]. These two types of multi-label prediction approaches have in general all achieved good empirical performance. However, very few methods have taken both aspects of label dependencies and shared model structures into account for multi-label learning. The Bayesian network models for multi-label learning [de Waal and van der Gaag, 2007; Bielza *et al.*, 2011] although take steps in this direction by learning separate feature subnetwork, class subnetwork, and feature-class bridge subnetwork, they nevertheless are limited to problems with a small number of discrete feature variables.

The proposed approach in this paper aims to integrate the strengths of label dependency based methods and common prediction structure based methods within a novel convex sparse feature learning framework. From the perspective of capturing label dependency, our work is closely related to the simple multi-label learning methods in [Godbole and Sarawagi, 2004; Guo and Gu, 2011]. The work in [Godbole and Sarawagi, 2004] uses a very intuitive and simple procedure to exploit multi-label dependency information. It first trains a set of binary SVM classifiers, one for each of the K classes. Then it uses the K binary classifiers to produce K label features to augment the original features of each instance. Finally another set of K binary SVM classifiers are trained on the augmented instances. Its testing process follows a corresponding procedure. This work is simple and straightforward, but lacks of principled explanation. The work of [Guo and Gu, 2011] generalizes this intuitive idea into a principled probabilistic framework based on directed conditional dependency networks (CDNs), where each label variable takes all the other label variables as its parents. In a CDN model, the training of multiple binary prediction models can be interpreted as maximizing the approximated joint conditional distributions of the label variables. The training process is even simpler than the SVM method in [Godbole and Sarawagi, 2004]. It only requires training one set of K independent binary probabilistic classifiers, and a simple Gibbs sampling procedure is used for conducting inference in the testing phase. Nevertheless, the CDN model in [Guo and Gu, 2011] uses a fully connected directed graph as the label dependency structure, which can easily fall into the trap of overfitting, especially when there are a limited number of training instances. With sparse feature learning, the proposed approach in this paper aims to integrate the strength of the CDN model but overcome its drawbacks.

3 Multi-label Classification with Sparse Feature Learning

Assume we have a set of N multi-label training instances $D = \{(\mathbf{x}^i, y_1^i, \cdots, y_K^i)\}_{i=1}^N$, where $\mathbf{x}^i \in I\!R^d$ is a column feature vector, y_k^i has value $+1$ when the kth label is assigned to instance i, and has value -1 otherwise. In this section, we present a sparse feature learning model for multi-label classification. We propose to first learn sparse label dependency features with ℓ_1 norm under a probabilistic conditional dependency network framework, and then learn sparse discriminative common features shared among the multiple single-label predictors by conducting multi-task style learning. Together, these two steps of sparse feature learning contribute to a novel sparse multi-label optimization problem. We develop a fast proximal gradient algorithm to solve the formulated problem for a global solution.

3.1 Learning Sparse Label Dependency Features

Inspired by the CDN model, we plan to capture the label dependency information using the structural *label dependency features* within a probabilistic framework of conditional dependency networks. The fixed fully connected dependency structure used in [Guo and Gu, 2011] can capture arbitrary label interdependency relationships, but is prone to overfitting given a limited number of labeled training instances. It is important to identify the most informative structural label dependency features to use.

Let $\mathbf{X}$ denote all the feature variables and $\mathbf{Y} = \{Y_1, \cdots, Y_K\}$ denote the K label variables. The fully connected conditional dependency network assumes that each label variable Y_k takes all the other label variables $\mathbf{Y}_{\overline{k}} = \{Y_1, \cdots, Y_{k-1}, Y_{k+1}, \cdots, Y_K\}$ as its augmented parent features in addition to the input features $\mathbf{X}$. Given the observation feature variables $\mathbf{X}$ and all label dependency feature variables $\mathbf{Y}_{\overline{k}}$, a probabilistic prediction function is associated with each label variable Y_k to define its local conditional probability distributions $p(Y_k|\mathbf{X}, \mathbf{Y}_{\overline{k}})$. Following the work [Guo and Gu, 2011], we use L2-norm regularized binary logistic regression classifiers as probabilistic predictors. For an instance $(\mathbf{x}, y_1, \cdots, y_K)$, the conditional probability of each y_k given all other variable values can be computed as follows using a logistic regression predictor

$$p(y_k|\mathbf{x}, \mathbf{y}_{\overline{k}}, \boldsymbol{\theta}_k) = \frac{1}{1 + \exp\big(-y_k(\mathbf{w}_k^\top \mathbf{x} + \mathbf{v}_k^\top \mathbf{y}_{\overline{k}} + b_k)\big)} \quad (1)$$

where $\boldsymbol{\theta}_k = [\mathbf{w}_k; \mathbf{v}_k; b_k]$ are the model parameters, $\mathbf{w}_k$ is the weight vector for the input features $\mathbf{x}$, $\mathbf{v}_k$ is the class-dependent weight vector for augmenting features $\mathbf{y}_{\overline{k}}$, and b_k is a bias term. Given the labeled training data D, K logistic regression predictors can be trained independently by minimizing the regularized log-loss of the training instances

$$\min_{\boldsymbol{\theta}_k} \; -\sum_{i=1}^{N} \log p(y_k^i|\mathbf{x}^i, \mathbf{y}_{\overline{k}}^i, \boldsymbol{\theta}_k) + \frac{\lambda_1}{2}\|\mathbf{w}_k\|^2 + \frac{\lambda_2}{2}\|\mathbf{v}_k\|^2 \quad (2)$$

for all $k = 1 \cdots K$. Here $\|\cdot\|$ denotes the Euclidean norm of a vector, λ_1 and λ_2 are tradeoff parameters that control the degree of regularization.

The objective function in (2) assumes each label variable Y_k depends on all the other label dependency feature variables $\mathbf{Y}_{\overline{k}}$. It corresponds to a fully connected directed graph over all the label variables $Y_1, \cdots, Y_K$. However, in practice, label dependencies often exhibit sparse patterns, and each Y_k may only have dependency relationships with a subset label variables $\mathbf{Y}_{\pi_k} \subseteq \mathbf{Y}_{\overline{k}}$. Thus a fully connected structural dependency can easily capture spontaneous label correlation relationships and suffer from overfitting, especially when the training set is small. To tackle this problem and expose the sparse structure of conditional dependency networks, we add an ℓ_1-norm regularization on parameters $\mathbf{v}_k$ in each independent logistic regression training to learn sparse label dependency features:

$$\begin{aligned}\min_{\boldsymbol{\theta}_k} \quad & -\sum_{i=1}^{N} \log p(y_k^i|\mathbf{x}^i, \mathbf{y}_{\overline{k}}^i, \boldsymbol{\theta}_k) + \qquad (3)\\ & \frac{\lambda_1}{2}\|\mathbf{w}_k\|^2 + \frac{\lambda_2}{2}\|\mathbf{v}_k\|^2 + \gamma\|\mathbf{v}_k\|_1\end{aligned}$$

As one of the simplest sparsity inducing norm, ℓ_1-norm has been widely used for feature selection [Ng, 2004; Zou and Hastie, 2005]. The ℓ_1-norm regularization here can push many elements of $\mathbf{v}_k$ to zeros, which correspondingly will drop some $\mathbf{Y}_t \subset \mathbf{Y}_{\overline{k}}$ features and remove the directed edges from $\mathbf{Y}_t$ to Y_k. Moreover, with both ℓ_1-norm and squared ℓ_2-norm over $\mathbf{v}_k$, it forms an elastic-net regularization [Zou and Hastie, 2005], which can lead to stabilized feature selection even when the data exhibits strong correlations among $\mathbf{Y}_{\overline{k}}$.

3.2 Learning Common Predictive Features Shared Across Prediction Models

The sparse learning with ℓ_1-norm above captures sparse structural label dependency features. We next consider the input features $\mathbf{X}$. With a large set of input features available for use in a prediction task, it is often beneficial to use only an informative small subset of input features to reduce overfitting and improve generalization performance, especially when the labeled training set is small. Moreover, the multiple binary prediction tasks, one for each label variable prediction, are closely related within the same input feature space. It is thus essential to select the common predictive features shared across the multiple binary prediction models, with the expectation of improving the generalization performance of multi-label classification. This naturally forms a group feature selection problem over the model parameters associated with the common input features in the multiple binary prediction models. We employ a group sparsity inducing $\ell_{2,1}$-norm to conduct group feature selection and train K logistic regression classifiers in one joint minimization problem

$$\begin{aligned}\min_{W,V,\mathbf{b}} \; & -\sum_{k=1}^{K}\sum_{i=1}^{N} \log p(y_k^i|\mathbf{x}^i, \mathbf{y}_{\overline{k}}^i, W_{:k}, V_{:k}, b_k) + \qquad (4)\\ & \frac{\lambda_1}{2}\|W\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2 + \gamma\sum_{k=1}^{K}\|V_{:k}\|_1 + \mu\|W\|_{2,1}\end{aligned}$$

where $W = [\mathbf{w}_1, \cdots, \mathbf{w}_K]$ is a $d \times K$ matrix, $W_{:k}$ denotes its kth column containing the parameter vector $\mathbf{w}_k$ used for the kth logistic regression classifier; $V = [\mathbf{v}_1, \cdots, \mathbf{v}_K]$ is a $(K-1) \times K$ matrix, $V_{:k}$ denotes its kth column containing the parameter vector $\mathbf{v}_k$ used for the kth logistic regression classifier; $\mathbf{b} = [b_1, \cdots, b_K]^\top$ is a $K \times 1$ vector containing the bias terms for the K logistic regression classifiers; $\|\cdot\|_F$ denotes the Frobenius norm, and the $\ell_{2,1}$-norm is defined as $\|W\|_{2,1} = \sum_{j=1}^{d}\|W_{j:}\|$. This optimization problem encodes both sparse label dependency feature learning by the ℓ_1-norm regularizers, $\sum_{k=1}^{K}\|V_{:k}\|_1$, and common predictive feature learning by the $\ell_{2,1}$-norm regularizer, $\|W\|_{2,1}$.

$\ell_{2,1}$-norm has been frequently used for robust group feature selections in multi-task learning scenarios [Obozinski *et al.*, 2006; Liu *et al.*, 2009; Nie *et al.*, 2010; Yang *et al.*, 2011]. The multi-label learning problem can be viewed as a very special multi-task problem, where each task is the binary prediction problem associated with each class label. It is thus reasonable to use $\ell_{2,1}$-norm to induce robust common feature selections here. Although optimization problems with $\ell_{2,1}$-norm regularizations have been addressed in different ways in the literature, the joint optimization problem we formulated above in (4) is unique from any previous work by containing three different types of norms: ℓ_2-norm, ℓ_1-norm, and $\ell_{2,1}$-norm. We will present a fast proximal gradient descent algorithm to solve this novel optimization problem.

3.3 Fast Proximal Gradient Descent

Although the minimization problem (4) is a convex optimization problem, the objective function is non-smooth due to the non-smoothness of the ℓ_1-norm regularization terms and the $\ell_{2,1}$-norm regularization term. We develop a fast proximal gradient algorithm to solve this non-smooth optimization problem for a global optimal solution.

For simplicity of presentation, we use $\boldsymbol{\theta}$ to denote the parameter column vector formed by combing all parameters

$\{W, V, \mathbf{b}\}$ together in a vector form, and use the notations $\boldsymbol{\theta}$ and $\{W, V, \mathbf{b}\}$ in an interchangeable manner. Then (4) can be re-expressed as

$$\min_{\boldsymbol{\theta}} \qquad F(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + g(\boldsymbol{\theta}) \tag{5}$$

where

$$\begin{aligned} f(\boldsymbol{\theta}) = & -\sum_{k=1}^{K}\sum_{i=1}^{N} \log p(y_k^i|\mathbf{x}^i, \mathbf{y}_{\bar{k}}^i, W_{:k}, V_{:k}, b_k) \qquad (6) \\ & + \frac{\lambda_1}{2}\|W\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2 \end{aligned}$$

$$g(\boldsymbol{\theta}) = \gamma \sum_{k=1}^{K} \|V_{:k}\|_1 + \mu\|W\|_{2,1} \tag{7}$$

The objective function $F(\boldsymbol{\theta})$ is a sum of two functions, where the first function $f(\boldsymbol{\theta})$ is a convex and smooth function, and the second function as a sum of ℓ_1-norms and $\ell_{2,1}$ norm is convex but non-smooth. To develop a proximal optimization method, we consider the second order approximation of the objective function $F(\boldsymbol{\theta})$. For any $L > 0$, at a given point $\boldsymbol{\theta}^{(t)}$, we define

$$\begin{aligned} Q_L(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) \;=\; & f(\boldsymbol{\theta}^{(t)}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^\top \nabla f(\boldsymbol{\theta}^{(t)}) \qquad (8) \\ & + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\|^2 + g(\boldsymbol{\theta}) \end{aligned}$$

where $\nabla f(\boldsymbol{\theta}^{(t)})$ denotes the gradient function of $f(\cdot)$ regarding $\boldsymbol{\theta}$ at point $\boldsymbol{\theta}^{(t)}$.

Proposition 1 *[Beck and Teboulle, 2009] Let $L(f)$ denote the Lipschitz constant of $\nabla f(\boldsymbol{\theta})$. Then for any $L \geq L(f)$, one has*

$$Q_L(\boldsymbol{\theta}, \boldsymbol{\theta}') \geq F(\boldsymbol{\theta}) \tag{9}$$

It is straightforward to prove the Proposition based on the definition of Lipschitz constant and the convexity and smoothness of $f(\boldsymbol{\theta})$. From now on, we assume $L \geq L(f)$. According to Proposition 1, we have $Q_L(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) \geq F(\boldsymbol{\theta})$. By minimizing $Q_L(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$, a unique solution can be returned,

$$\begin{aligned} p_L(\boldsymbol{\theta}^{(t)}) &= \arg\min_{\boldsymbol{\theta}} Q_L(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) \\ &= \arg\min_{\boldsymbol{\theta}} \left\{ \frac{L}{2}\left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\|^2 + g(\boldsymbol{\theta}) \right\}, \qquad (10) \end{aligned}$$

where $\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{\theta}^{(t)} - \frac{1}{L}\nabla f(\boldsymbol{\theta}^{(t)})\right)$. The minimization in (10) can be equivalently decomposed into a set of separate minimization problems over $W, \{V_{:k}\}, b$ respectively, such as

$$p_L(W^{(t)}) = \arg\min_{W} \left\{ \frac{L}{2}\left\|W - \widehat{W}\right\|^2 + \mu\|W\|_{2,1} \right\}, \tag{11}$$

$$p_L(V_{:k}^{(t)}) = \arg\min_{\mathbf{v}} \left\{ \frac{L}{2}\left\|\mathbf{v} - \widehat{V_{:k}}\right\|^2 + \gamma\|\mathbf{v}\|_1 \right\}, \tag{12}$$

$$p_L(\mathbf{b}^{(t)}) = \arg\min_{\mathbf{b}} \left\{ \frac{L}{2}\left\|\mathbf{b} - \widehat{\mathbf{b}}\right\|^2 \right\}, \tag{13}$$

Algorithm 1 Fast Proximal Gradient Algorithm

```
Initialize: $L_0 > 0, \eta > 1, \mathbf{z}^{(1)} = \boldsymbol{\theta}^{(0)}, \alpha_1 = 1, t = 0.$
Repeat
1. Set $t = t + 1, L = L_{t-1}$
2. While $F(p_L(\mathbf{z}^{(t)})) > Q_L(p_L(\mathbf{z}^{(t)}), \mathbf{z}^{(t)})$
       $L = \eta L$
3. Set $L_t = L$ and update
       $\boldsymbol{\theta}^{(t)} = p_{L_t}(\mathbf{z}^{(t)}),$
       $\alpha_{t+1} = \frac{1+\sqrt{1+4\alpha_t^2}}{2},$
       $\mathbf{z}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \left(\frac{\alpha_t - 1}{\alpha_{t+1}}\right)(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)})$
Until Convergence is achieved
```

which have closed-form solutions. The closed-form solutions of the convex optimization problems with sparsity-inducing norms in (11) and (12) can be computed by applying group soft-thresholding and direct soft-thresholding operators [Bach *et al.*, 2011] respectively, such as

$$p_L(W^{(t)}) \;=\; \widehat{W_{j:}}\left(1 - \frac{\mu}{L\|\widehat{W_{j:}}\|}\right)_+, \;\forall j; \tag{14}$$

$$p_L(V_{:k}^{(t)}) \;=\; sign(\widehat{V_{:k}}) \circ \left(|\widehat{V_{:k}}| - \frac{\gamma}{L}\right)_+, \;\forall k; \tag{15}$$

where $\circ$ denotes elementwise vector product and the operator $(\cdot)_+ = \max(\cdot, 0)$. The closed-form solution of (13) is

$$p_L(\mathbf{b}^{(t)}) \;=\; \widehat{\mathbf{b}} \tag{16}$$

By conducting iterative updates according to Equation (10), we can obtain a sequence of points, $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots$, to minimize the objective function $F(\boldsymbol{\theta})$. But to achieve an optimal convergence rate, we further adopt a fast convergence update scheme of FISTA [Beck and Teboulle, 2009]. The overall fast proximal gradient algorithm is given in Algorithm 1. With the fast update scheme, the iterative algorithm has a convergence rate of $O(\frac{1}{t^2})$, under the condition of $Q_L(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) \geq F(\boldsymbol{\theta}^{(t+1)})$. To find L values in each iteration guaranteeing $Q_L(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) \geq F(\boldsymbol{\theta}^{(t+1)})$, we use a step search procedure.

Testing After the training process, K logistic regression classifiers will be returned to define the conditional probabilistic distributions associated with each label variable given its parents. The sparse structure of the conditional dependency network is determined according to the V parameter matrix. Then given new instances, the prediction can be conducted using a Gibbs sampling inference algorithm, which is similar but more efficient than in the CDN model of [Guo and Gu, 2011].

4 Experiments

To evaluate the effectiveness and robustness of the proposed method on learning multi-label classification models from a small set of labeled training instances, we conducted experiments on six multi-label data sets, comparing the proposed approach to a number of alternative approaches. In this section, we report our experimental setting and results.

Table 1: The number of instances, label cardinality (LC) and dimensions of the six constructed data sets.

Data set	No. of inst.	LC	Dimensions
Corel 5k	1654	2.27	1000
IAPR TC12	9494	2.63	1000
MIR Flickr	1396	2.23	1000
Enron	1302	2.94	1001
Medical	785	1.19	1449
Bibtex	3400	1.13	1836

4.1 Experimental Setting

Data Sets We conducted experiments on six commonly used multi-label data sets, *Corel 5k, IAPR TC12, MIR Flickr, Enron, Medical*, and *Bibtex*. The first three data sets come from INRIA feature data sets, [1] which are benchmarks for computer vision tasks. The rest three data sets are for text categorizations, and can be found at mulan website. [2] For the three image data sets, we used SIFT descriptors as features. Each instance in these data sets is represented as a 1000-dimensional SIFT feature vector. The three text data sets, are represented in terms of 0-1 indicator features, in the size of 1001, 1449 and 1836 respectively. One key measure of multi-label data sets is their label cardinality [Tsoumakas and Katakis, 2007], which is defined as the average number labels assigned to each instance. Thus it should be a value between 1 and K, where K is the number of classes. The label cardinality of most existing multi-label data sets is small. However, if the label cardinality of a data set is close to 1, the multi-label classification task will be close to a standard single labeled multi-class task, and it does not make much sense to capture label dependencies on such a data set. The effectiveness of multi-label learning can most likely be demonstrated on data sets whose label cardinality is reasonably large. We thus chose ten most popular labels, i.e., labels most frequently appeared, from each of the six data sets to use, aiming to produce multi-label data sets with reasonable label cardinality values. According to these labels, we extract subsets from these data sets by eliminating instances not labeled with any one of the selected ten labels. Table 1 shows the statistics, data size and label cardinality information of these constructed data sets.

Methods We compared the following six probabilistic multi-label learning methods in the experiments:

(1) *LR*, the baseline method that decomposes multi-label classification into a set of binary classification problems using 1-vs-all scheme, and uses binary logistic regression as classifiers;

(2) *PCC*, the probabilistic multi-label chain classifiers developed in [Dembczynski *et al.*, 2010];

(3) *CDN*, the conditional dependency network method developed in [Guo and Gu, 2011];

(4) ℓ_1-*CDN*, the sparse CDN with an ℓ_1-norm regularizer. This is the optimization problem we constructed in Eq. (3), which can be solved using a modified version of the proximal gradient algorithm in Algorithm 1;

(5) $\ell_{2,1}$-*LR*, a variant of our proposed method by dropping the structural dependency features, which captures only the shared features among the multiple binary LR classifiers using a $\ell_{2,1}$-norm regularizer;

(6) *Sparse*, the proposed sparse feature learning method, which is a combination of ℓ_1-*CDN* and $\ell_{2,1}$-*LR* and captures both label interdependency and shared predictive features of multi-label learning.

Evaluation Measures In our experiments, we used two $F1$ measures, Micro-F1 measure and Macro-F1 measure, to evaluate the multi-label prediction performance. The two $F1$ measures take both the precision and recall into accounts in different ways. The precision and recall, P_k and R_k, for each label k, are defined as follows:

$$P_k = \frac{\sum_i (y_k^i)_+ (\hat{y}_k^i)_+}{\sum_i (\hat{y}_k^i)_+}, \quad R_k = \frac{\sum_i (y_k^i)_+ (\hat{y}_k^i)_+}{\sum_i (y_k^i)_+}$$

where y_k^i and $\hat{y}_k^i$ are the true label value and predicted label value respectively for the ith instance on the k-th class; and $(\cdot)_+ = \max(\cdot, 0)$. Then we have Micro-F1 and Macro-F1 measure as

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K} \frac{2P_kR_k}{P_k + R_k} = \frac{1}{K}\sum_{k=1}^{K} \frac{2\sum_i (y_k^i)_+ (\hat{y}_k^i)_+}{\sum_i (y_k^i)_+ + \sum_i (\hat{y}_k^i)_+}$$

$$\text{Micro-F1} = \frac{2\sum_{k=1}^{K}\sum_{i=1}^{N} (y_k^i)_+ (\hat{y}_k^i)_+}{\sum_{k=1}^{K}\sum_{i=1}^{N} (y_k^i)_+ + \sum_{k=1}^{K}\sum_{i=1}^{N} (\hat{y}_k^i)_+}$$

4.2 Experimental Results

When there are sufficiently large number of training instances, good classification performance can be achieved even with simple binary classifiers. However, it is costly and time consuming to obtain labeled data and thus it is beneficial to develop methods that can learn good multi-label classification models with a small number of training instances. The purpose of our experiments is to investigate the performance of the proposed methods when the number of training instances is small. In particular, we conducted two sets of experiments on the constructed data sets using two different training sizes, $|D| = \{200, 400\}$.

In each set of experiments, with the given training size, we randomly divided each data set into partitions containing training set and test set, with a test size of 1000 when possible. The trade-off parameters for the comparison methods are pre-selected using one random training and test partition. The parameters are selected to maximize the average value of the two F1 measures. The methods, *LR, CDN*, ℓ_1-*CDN*, $\ell_{2,1}$-*LR*, and *Sparse*, all involve some or all of the following trade-off parameters: $\lambda_1, \lambda_2, \gamma$, and μ. We set $\lambda_1 = \lambda_2 = 0.1$. γ and μ are selected from the following sets of values, $\{1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}\}$ and $\{1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}\}$, respectively.

[1] http://lear.inrialpes.fr/people/guillaumin/data.php

[2] http://mulan.sourceforge.net/datasets.html

Table 2: Performance in Macro-F1 measure (%)

#Train.	Method	Corel 5k	Mir Flickr	IAPR TC12	Enron	Medical	Bibtex
200	LR	34.9±1.0	28.9±0.8	38.8±0.5	44.5±1.2	62.9±2.5	60.0±4.6
	PCC	38.5±1.4	30.2±1.0	39.7±0.6	50.0±1.6	79.9±4.6	57.0±5.4
	CDN	36.7±0.9	32.1±0.8	41.5±0.8	49.5±1.9	71.4±6.7	46.1±4.2
	ℓ_1-CDN	38.9±0.9	31.5±1.6	42.4±0.2	50.0±1.8	75.8±4.7	52.3±5.0
	$\ell_{2,1}$-LR	40.8±0.2	33.5±0.7	42.4±0.6	50.6±2.3	78.9±4.2	53.4±4.5
	Sparse	**44.5±1.3**	**34.4±0.4**	**44.3±0.8**	**51.2±1.7**	**83.1±4.6**	**62.3±3.8**
400	LR	33.0±1.0	30.7±0.7	39.5±0.5	45.2±0.8	65.0±3.7	68.8±1.5
	PCC	35.7±1.6	31.4±0.8	40.0±0.4	51.1±1.6	86.1±1.2	69.6±1.2
	CDN	37.6±0.5	33.7±1.9	41.0±0.8	53.6±1.5	83.4±1.2	60.6±2.8
	ℓ_1-CDN	38.4±0.9	32.2±1.4	42.1±0.8	54.4±2.0	85.3±1.6	65.1±1.9
	$\ell_{2,1}-$LR	42.8±1.0	33.2±1.2	43.5±0.8	54.2±1.9	**87.0±2.2**	65.8±1.5
	Sparse	**45.1±1.0**	**34.4±1.0**	**44.8±0.7**	**55.0±1.5**	86.1±1.1	**72.2±1.2**

Table 3: Performance in Micro-F1 measure (%)

#Train	Method	Corel 5k	Mir Flickr	IAPR TC12	Enron	Medical	Bibtex
200	LR	40.0±1.2	30.9±1.2	41.3±0.8	53.9±1.2	70.5±2.9	76.4±3.6
	PCC	45.8±0.9	31.9±0.8	42.5±0.2	61.8±2.6	85.5±2.0	74.5±4.9
	CDN	39.4±1.1	33.3±0.8	41.5±0.8	62.7±1.6	80.6±1.7	63.0±2.7
	ℓ_1-CDN	42.0±1.4	32.3±1.5	42.7±0.2	64.4±1.0	83.7±1.6	71.9±2.6
	$\ell_{2,1}$-LR	48.5±1.1	34.4±1.3	44.4±0.7	64.3±0.8	84.9±0.7	72.5±1.1
	Sparse	**48.8±1.9**	**35.1±1.6**	**45.6±1.0**	**64.3±2.4**	**86.6±1.6**	**77.2±4.5**
400	LR	36.7±1.2	32.0±0.2	41.4±1.3	53.5±1.6	69.8±0.9	80.1±2.7
	PCC	39.4±1.5	32.5±1.0	41.7±0.8	61.0±0.6	87.9±1.3	80.1±1.1
	CDN	40.5±1.2	34.5±1.4	41.3±1.1	64.7±1.6	85.7±2.7	74.2±3.8
	ℓ_1-CDN	41.4±0.7	32.8±0.5	42.6±1.2	65.6±2.1	87.8±0.9	78.9±2.4
	$\ell_{2,1}$-LR	50.8±1.3	34.6±0.3	45.8±0.7	65.8±0.5	**88.8±1.8**	79.4±2.4
	Sparse	**51.2±1.1**	**35.3±1.0**	**46.3±0.7**	**65.9±0.5**	87.1±1.3	**82.4±0.8**

With the selected parameters, we then repeatedly run each method 10 times on 10 sets of randomly partitioned training and test data. The average results for the six methods on the six data sets are presented in Table 2 and Table 3 in terms of Macro-F1 measure and Micro-F1 measure respectively. We can see that the *PCC* method, with an ensemble of different label interdependencies, has a clear win over *LR* except on the *Bibtex* data. With the small training sizes, although the *CDN* method still outperforms the baseline *LR* method in many cases, it has inferior performance on two data sets, *Corel 5k* and *Bibtex*. With an additional ℓ_1-norm to induce sparse dependency structure, the resulted ℓ_1-*CDN* can improve *CDN* in all cases except on the *Mir Flicker* data set. On the other hand, with only $\ell_{2,1}$-norm regularization, the $\ell_{2,1}$-*LR* outperforms all the previous four methods on *Corel 5k*, *Mir Flicker*, and *IAPR TC12* for both training sizes, and on *Medical* for training size 400. But on *Enron*, it demonstrates similar performance as ℓ_1-*CDN*, and on *Bibtex*, it produces inferior performance even comparing to the baseline *LR*. By integrating the two types of sparse feature learning together, the proposed *Sparse* method, with both ℓ_1-norm and $\ell_{2,1}$-norm regularizations, outperforms all the other five methods on almost every data set of different training sizes, except only on the *Medical* with training size 400 where its performance is still competitive to the best. These results suggest that the two types of sparse feature learning over label dependency features and common input features respectively can complement each other in most cases, and our proposed sparse learning method provides an effective framework for integrating their strengths on multi-label classification.

5 Conclusions

In this paper we proposed a novel probabilistic multi-label classification model based on sparse feature learning. By employing an individual sparsity inducing ℓ_1-norm and a group sparsity inducing $\ell_{2,1}$-norm together, the proposed model can induce both a sparse dependency network structure over the label variables, and a common set of predictive features across the multiple binary predictors associated with the multiple classes. The resulted optimization problem is a convex but non-smooth optimization problem. We developed a fast proximal gradient algorithm to solve the proposed problem for a global optimal solution. Our empirical results on a number of multi-label data sets show the proposed sparse learning approach can outperform a set of related probabilistic multi-label learning methods and produce effective results even when the number of training instances is small.

References

[Argyriou *et al.*, 2006] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Proc. of NIPS*, 2006.

[Bach *et al.*, 2011] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

[Beck and Teboulle, 2009] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. on Imaging Sciences*, 2(1):183–202, 2009.

[Bielza *et al.*, 2011] C. Bielza, G. Li, and P. Larraòaga. Multi-dimensional classification with bayesian networks. *Int. J. Approx. Reasoning*, 52:705–727, 2011.

[Chen *et al.*, 2007] W. Chen, J. Yan, B. Zhang, Z. Chen, and Q. Yang. Document transformation for multi-label feature selection in text categorization. In *Proc. of ICDM*, 2007.

[de Waal and van der Gaag, 2007] P. de Waal and L. van der Gaag. Inference and learning in multi-dimensional bayesian network classifiers. In *Proc. of Euro. Conf. on Symb. and Quant. Appro. to Reasoning with Uncertainty*, pages 501–511, 2007.

[Dembczynski *et al.*, 2010] K. Dembczynski, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proc. of ICML*, 2010.

[Elisseeff and Weston, 2002] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Proc. of NIPS*, 2002.

[Fuernkranz *et al.*, 2008] J. Fuernkranz, E. Huellermeier, E. Mencia, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2), 2008.

[Ghamrawi and McCallum, 2005] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *Proc. of CIKM*, 2005.

[Godbole and Sarawagi, 2004] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Proc. of the Pacific-Asia Conference on KDDM*, 2004.

[Guo and Gu, 2011] Y. Guo and S. Gu. Multi-label classification using conditional dependency networks. In *Proc. of IJCAI*, 2011.

[Guo and Schuurmans, 2011] Y. Guo and D. Schuurmans. Adaptive large margin training for multilabel classification. In *Proc. of AAAI*, 2011.

[Hariharan *et al.*, 2010] B. Hariharan, L. Zelnik-Manor, SVN Vishwanathan, and M. Varma. Large scale max-margin multi-label classification with priors. In *Proc. of ICML*, 2010.

[Ji *et al.*, 2010] S. Ji, L. Tang, S. Yu, and J. Ye. A shared-subspace learning framework for multi-label classification. *ACM Transactions on Knowledge Discovery from Data*, 4, No. 2:1–29, 2010.

[Lewis *et al.*, 2004] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *JMLR*, 5:361–397, 2004.

[Liu *et al.*, 2009] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient L2,1-norm minimization. In *Proc. of UAI*, 2009.

[Ng, 2004] A. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proc. of ICML*, 2004.

[Nie *et al.*, 2010] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint L2,1-norms minimization. In *Proc. of NIPS*, 2010.

[Obozinski *et al.*, 2006] G. Obozinski, B. Taskar, and M.I. Jordan. Multi-task feature selection. Technical report, Statistics Department, UC Berkeley, Tech. Rep, 2006.

[Petterson and Caetano, 2011] J. Petterson and T. Caetano. Submodular multi-label learning. In *Proc. of NIPS*, 2011.

[Qi *et al.*, 2007] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, and H. Zhang. Correlative multi-label video annotation. In *Proc. of Multimedia*, 2007.

[Schapire and Singer, 2000] R. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2):135–168, 2000.

[Shalev-Shwartz and Singer, 2006] S. Shalev-Shwartz and Y. Singer. Efficient learning of label ranking by soft projections onto polyhedra. *JMLR*, 7:1567–1599, 2006.

[Tsoumakas and Katakis, 2007] G. Tsoumakas and I. Katakis. Multi-label classification: an overview. *Inter. J. of Data Warehous. and Mining*, 3(3):1–13, 2007.

[Yan *et al.*, 2007] R. Yan, J. Tesic, and J. Smith. Model-shared subspace boosting for multi-label classification. In *SIGKDD*, 2007.

[Yang *et al.*, 2011] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou. L2,1-norm regularized discriminative feature selection for unsupervised learning. In *Proc. of IJCAI*, 2011.

[Yu *et al.*, 2005] K. Yu, S. Yu, and V. Tresp. Multi-label informed latent semantic indexing. In *Proc. of the Annual ACM SIGIR Conference*, 2005.

[Zaragoza *et al.*, 2011] J. Zaragoza, L. Sucar, E. Morales, C. Bielza, and P. Larranaga. Bayesian chain classifiers for multidimensional classification. In *Proc. of IJCAI*, 2011.

[Zhang and Zhou, 2008] M. Zhang and Z. Zhou. Multi-label dimensionality reduction via dependency maximization. In *Proc. of AAAI*, 2008.

[Zhou and Zhang, 2006] Z. Zhou and M. Zhang. Multi-instance multi-label learning with application to scene classification. In *Proc. of NIPS*, 2006.

[Zhu *et al.*, 2005] S. Zhu, X. Ji, W. Xu, and Y. Gong. Multi-labelled classification using maximum entropy method. In *Proc. of the Annual ACM SIGIR Conference*, 2005.

[Zou and Hastie, 2005] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.