Constrained Submodular Minimization for Missing Labels and Class Imbalance in Multi-Label Learning

Baoyuan Wu  
KAUST, Saudi Arabia

Siwei Lyu  
SUNY-Albany, NY USA

Bernard Ghanem  
KAUST, Saudi Arabia

Abstract

In multi-label learning, there are two main challenges: missing labels and class imbalance (CIB). The former assumes that only a partial set of labels are provided for each training instance while other labels are missing. CIB is observed from two perspectives: first, the number of negative labels of each instance is much larger than its positive labels; second, the rate of positive instances (i.e. the number of positive instances divided by the total number of instances) of different classes are significantly different. Both missing labels and CIB lead to significant performance degradation. In this work, we propose a new method to handle these two challenges simultaneously. We formulate the problem as a constrained submodular minimization that is composed of a submodular objective function that encourages label consistency and smoothness, as well as, class cardinality bound constraints to handle class imbalance. We further present a convex approximation based on the Lovasz extension of submodular functions, leading to a linear program, which can be efficiently solved by the alternative direction method of multipliers (ADMM). Experimental results on several benchmark datasets demonstrate the improved performance of our method over several state-of-the-art methods.

1 Introduction

Multi-label learning (ML) assumes that one instance can be assigned to multiple classes simultaneously. For example, one image can be annotated with several tags, and one document can be associated with multiple topics. Although many multi-label learning methods have been proposed in recent years, a main challenge remains for this problem, i.e., the lack of completely labeled training instances. This is important because in many real life applications, most training instances are only partially labeled, while other labels are not provided or missing. One such example is image annotation, a human labeling can only feasibly annotate each training image with a subset of tags, especially when the number of classes/tags is large. Learning from such partially labeled instances is referred to as the multi-label learning with missing labels (MLML) problem (Wu et al. 2014; Yu et al. 2014).

Several previous works have tried to handle the MLML problem, such as (Goldberg et al. 2010; Cabral et al. 2011; Kapoor, Viswanathan, and Jain 2012; Xu, Jin, and Zhou 2013; Wu et al. 2014; Yu et al. 2014; Wu et al. 2015; Chen et al. 2015). However, most of them disregard another important challenge in multi-label learning, i.e., class imbalance (CIB), which has two different phenomena. Firstly, each instance is assigned with only a few positive labels, while most other labels are negative. We refer to this type of class imbalance as CIB-1 in this work. Secondly, the proportions of positive instances of different classes may be significantly different. We refer to this type of CIB as CIB-2 in this work. CIB-1 often occurs in binary classification problems, while CIB-2 is more widely encountered in multi-label classification problems. It has been observed (Sahare and Gupta 2012; Zhang and Hu 2014) that both CIB-1 and CIB-2 are likely to lead to the performance degradation of many popular models, such as SVM and neural networks.

The more challenging case is missing labels and two types of class imbalances co-exist in a multi-label learning problem. This is because the bias between the positive and negative labels becomes larger than where there is no missing labels. Thus the missing labels tend to be negative labels with the larger degree, as the amount of missing labels increases. To date, there only exist a few works (Zhang, Li, and Liu 2015; Petterson and Caetano 2010) specifically consider CIB-1, but no existing work can handle CIB-1 and CIB-2 simultaneously, in particular in the context of MLML.

The goal of this work is to develop a unified model to handle missing labels and class imbalance jointly. We formulate the problem as a transductive learning problem that include five components that are label consistency, instance-level and class-level label smoothness, and two types of class cardinality (lower and upper) bounds. The first three components are used to propagate the label information from the provided labels to missing labels, and the latter two components are included to handle two types of the class imbalance problem. We first formulate a unified model that combines these components as a constrained submodular minimization problem (CSM). However, due to the class cardinality constraint, it is a NP-hard problem. Utilizing the Lovasz extension of submodular function, we approximate CSM by a continuous linear programming (LP) with linear constraints. The LP problem is efficiently solved by alternative direction method of multipliers (ADMM). As the final output, the signs of the continuous labels are adopted as the discrete
class labels. A graphical illustration of the overall framework is presented in Figure 1.

The main contributions of this work include: (1) we propose a unified model to jointly handle missing labels and class imbalance in multi-label learning by incorporating several types of prior knowledge; (2) the unified framework is formulated as a constrained submodular minimization problem, of which a convex approximation is also provided based on the Lovasz extension of the submodular function; (3) experiments on several benchmark multi-label datasets verify the efficacy of the proposed method, on handling both missing labels and CIB, as well as its improvements over the state-of-the-art methods.

2 Related Work

We briefly review multi-label learning methods that handle missing labels and handle class imbalance.

Multi-label learning methods handling missing labels can be generally partitioned into four categories. (i) Missing labels can be treated as negative labels, which will bring in undesired label bias, such as (Chen et al. 2008; Sun, Zhang, and Zhou 2010; Bucak, Jin, and Jain 2011; Chen, Zheng, and Weinberger 2013; Wang et al. 2014; Wang, Si, and Zhang 2014; Chen et al. 2015). When massive missing labels exist, many ground-truth positive labels will be incorrectly initialized as negative labels, leading to significant performance degradation. (ii) Missing labels can be augmented into the label set as a special type of labels. Wu et al. (Wu et al. 2014; 2015) propose to use three different labels, including positive labels +1, negative labels −1, and missing labels 0 to model the learning problem. A similar setting of using {1, 0, 1/2} is also used (Wu, Lyu, and Ghanem 2015). The label bias is avoided in these three works. (iii) Techniques in matrix completion (MC) (Johnson 1990) is borrowed to handle missing labels, as in (Goldberg et al. 2010; Cabral et al. 2011; Xu, Jin, and Zhou 2013; Yu et al. 2014). In (Goldberg et al. 2010; Cabral et al. 2011; Xu, Jin, and Zhou 2013; Yu et al. 2014), a recent work known as LEML (Yu et al. 2014) proposes an empirical risk minimization (ERM) framework to handle missing labels. Both MC-based methods and LEML exploit the mask matrix to avoid the label bias and utilize the low rank assumption to explicitly embed the label dependencies. (iv) Missing labels can be treated as latent variables in probabilistic models, including Bayesian networks (Kapoor, Viswanathan, and Jain 2012; Vasisht et al. 2014; Bi and Kwok 2014) and conditional RBMs (Li, Zhao, and Guo 2015). However, all the aforementioned work assumes balanced positive and negative training data labels, though some of them simply set different costs to positive and negative labels in the loss function. Class imbalance has not been well handled in these models.

On the other hand, class imbalance (CIB) has not been extensively studied in the multi-label learning context. There are two main categories of methods here. One category is to directly maximize the imbalance-specific metric in multi-label learning, such as the \( F_{\text{macro}} \) in (Pettersson and Cai-tano 2010; Dembczynski et al. 2013). Another category is recently proposed in (Zhang, Li, and Liu 2015), where the binary-class imbalance classifier for the current class and the multi-class imbalance classifier for other classes are aggregated to make final predictions. However, it only considers CIB-1 in supervised multi-label learning scenarios, while CIB-2 and missing labels are ignored. In contrast, we handle CIB by embedding linear constraints, which are independent of any specific metric or base-learner. Moreover, to the best of our knowledge, no previous work has been developed to handle missing labels and the two CIB challenges jointly.

3 Proposed Model

Given a multi-dimensional dataset \( X = (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n} \), each instance \( x_i \) corresponds to a multi-dimensional data vector and is associated with \( m \) different classes \( \{c_1, \ldots, c_m\} \). An incomplete label matrix \( Y \in \{-1, 0, +1\}^{m \times n} \) is also provided, where \( Y_{ji} = +1 \) means \( x_i \) is associated with \( c_j \) (i.e. the positive label), \( Y_{ji} = -1 \) means \( c_j \) does not exist in \( x_i \) (i.e. the negative label), and \( Y_{ji} = 0 \) denotes the missing label. The missing label proportion in the training label matrix is denoted as \( \varepsilon = \frac{|Y_{\text{fr}} = 0|}{N_{\text{fr}}} \), with \( |Y_{\text{fr}} = 0| \) indicating the number of zero entries in \( Y_{\text{fr}} \). The training label matrix \( Y_{\text{fr}} \) corresponds to the instances with at least 1 provided label (−1 or +1) and \( N_{\text{fr}} \) denotes the number of training instances. The positive label rate in the whole label matrix is denoted as \( \eta = \frac{|Y_{\text{fr}} = 1|}{n \times m} \). Our goal is to obtain a complete label matrix \( Z \in \{-1, +1\}^{m \times n} \), based on \( X \) and \( Y \). To this end, we use label dependencies among the instances and the classes to propagate the label information from the provided labels to the missing labels, as well as adopt cardinality constraints to avoid degenerate results due to CIB. We adopt three types of information, including label consistency, label smoothness, and class cardinality bounds, of which the definitions will be presented in the following sections.
Label Consistency

Label consistency serves as the loss function. For $Y_P = P \circ Y$ (‘$\circ$’ denotes Hadamard product), we have

$$\ell(Z, Y) = \sum_{i,j} P_{ij} Y_{ij}(Y_{ij} - Z_{ij}) = tr(Y_P^\top(Y - Z)).$$

When $Y_{ij} = \pm 1$, if $Z_{ij} \neq Y_{ij}$, then $\ell(Z_{ij}, Y_{ij}) = 2P_{ij}$, while $\ell(Z_{ij}, Y_{ij}) = 0$ if $Z_{ij} = Y_{ij}$. Besides, $\ell(Z_{ij}, Y_{ij}) = 0$ indicates the missing labels do not contribute to the loss function. The label cost matrix $P$ is defined as follows: if $Y_{ij} = +1$, $P_{ij} = \tau_1^+$; if $Y_{ij} = -1$, $P_{ij} = \tau_1^-$; if $Y_{ij} = 0$, $P_{ij} = 0$. $\tau_1^+ / \tau_1^-$ is the ratio between the number of negative and positive instances in class $c_i$.

Label Smoothness

Instance-level label smoothness (Wu et al. 2014) is

$$\text{tr}(ZL_X Z^\top) = \sum_{i,j} \frac{W_X(i,j)}{2} \left( \frac{Z_{ki}}{d_X(i)} - \frac{Z_{kj}}{d_X(j)} \right)^2,$$

where the instance similarity is $W_X(i,j) = W_X(j,i) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma_i \sigma_j})$, $i \neq j$, and $W_X(i,i) = 0$. The kernel size $\sigma_i$ and $\sigma_j$ are determined by following the setting in (Wu et al. 2014). The normalized Laplacian matrix is $L_X = I - D_X^{-\frac{1}{2}} W_X D_X^{-\frac{1}{2}}$ with $D_X = \text{diag}(d_X(1), \ldots, d_X(n))$. $W_X$ is viewed as a weighted graph $G_X = \{V_X, E_X\}$ among instances, with $V_X = \{1, \ldots, n\}$ and $E_X = \{(i,n)\}_{i=1}^{\frac{n(n-1)}{2}}$. $n_{eX}$ indicates the number of edges in $W_X$, which is half the number of non-zero entries in $W_X$. The edge between nodes $e_1$ and $e_2$ is indexed as $e \in \{1, \ldots, n_{eX}\}$, with its weight being $W_X(e_1, e_2)$.

Class-level label smoothness (Wu et al. 2014) is

$$\text{tr}(Z^\top L_C Z) = \sum_{i,j} \frac{W_C(i,j)}{2} \left( \frac{Z_{ik}}{d_C(i)} - \frac{Z_{jk}}{d_C(j)} \right)^2,$$

where the class co-occurrence is $W_C(i,j) = N_{i,j} / \|Y_i\|_0 \|Y_j\|_0$ when $i \neq j$ and $W_C(i,i) = 0$. $Y_i = (Y_{i,1} = 1, \ldots, Y_{i,n})$. $d_C(i) = \sum_j W_C(i,j)$.

$$L_C = I - D_C^{-\frac{1}{2}} W_C D_C^{-\frac{1}{2}}$$

Similar to $W_X$, $W_C$ constitutes a weighted graph $G_C$ among classes, where $V_C = \{1, \ldots, n\}$ and $E_C = \{1, \ldots, n_{eC}\}$. $n_{eC}$ equals to the half number of the non-zero entries in $W_C$, and the weight of edge $e$ is $W_C(e_1, e_2)$.

Note that, to the best of our knowledge, above two smoothness assumptions were firstly proposed in (Chen et al. 2008), and they have been borrowed and cited in some more recent works, such as (Wu et al. 2014; 2015) etc. Due to the space limit, we cannot cover all work that also use the similar smoothness assumptions.

Class Cardinality Bounds

We introduce two types of class cardinality bounds to handle class imbalance. The first type constrains the number of positive labels of each instance in range $[v_1^+, v_1^-] \subset [1, m]$. The second type requires the number of positive instances of class $c_i \in [v_2^+, v_2^-] \subset [1, n]$. We formulate them as linear constraints:

$$\text{CIB-1: } v_1^+ \leq N_{1,n}^i \leq v_1^- \leq n_{1,n} \leq v_2^+.$$

$$\text{CIB-2: } v_2^+ \leq N_{1,n} \leq v_2^-.$$

Determine $\{v_1^+, v_1^-\}$ of CIB-1. As shown in Figure 1, we firstly calculate the vector $r_1$ with $r_1(j) = |Y_{tr}(\cdot, j) = 1|$. Then a histogram $h$ is plotted, with $x$-coordinate being the number of positive labels and $y$-coordinate being the number of corresponding instances. Denoting the $5^{th}$ and $95^{th}$ percentiles of $h$ as $h_{0.05}$ and $h_{0.95}$ respectively, we define the lower and upper bounds as: $v_1^+ = \max\{1, h_{0.05} \times \min\{1 + \varepsilon / 3, 1.2\}\}$ and $v_1^- = \min\{0.8m, h_{0.95} \times \min\{1 + \varepsilon / 1.5\}\}$. When $\varepsilon = 0$, we enforce that the number of positive labels of each instance to be in $[h_{0.05}, h_{0.95}]$. Although there is a range bias for about 10% of the instances, the rest will have satisfactory prediction. When $\varepsilon > 0$, both $h_{0.05}$ and $h_{0.95}$ will be smaller than their ground-truth values (i.e., in the case of $\varepsilon = 0$). Thus, we amplify them with a reasonable rate according to $\varepsilon$. In experiments, we observe that $h_{0.05}$ varies in a very small range as $\varepsilon$ increases. The reason is $h_{0.05} = 0$ is always a small value in most experiments, and the range $[1, h_{0.05}]$ to which $h_{0.05}$ belongs is very narrow. Thus we increase $h_{0.05}$ with a small rate, by multiplying $\min\{1 + \varepsilon / 3, 1.2\}$. Compared to $h_{0.05}$, the variation range of $h_{0.95}$ becomes larger. But $h_{0.95}$ still varies smoothly w.r.t. $\varepsilon$. Thus we adjust it by multiplying it by $\min\{1 + \varepsilon / 1.5\}$.

Determine $\{v_2^+, v_2^-\}$ of CIB-2. As shown in Figure 1, we calculate $r_2$ with $r_2(i) = |Y_{tr}(i, \cdot) = 1|$, which is the number of positive instances of class $c_i$. The corresponding number in the complete label matrix is estimated as $r_2(i) = r_2(i) \times n_{eX} / n_{eC}$. Then, the bounds are: $v_2^+ = \max\{1, f_2(i) \times \zeta_1\}$ and $v_2^- = \min\{0.8n, f_2(i) \times \zeta_2\}$, where we choose $\zeta_1$ and $\zeta_2$ from the sets $\{0.8, 1.2, 1.5\}$ and $\{1.5, 2, 3\}$ respectively.

Objective Function

Combining the previous terms, we formulate MLML as a discrete optimization problem with linear constraints,

$$\min_{Z} \text{tr}(Y_P^\top(Y - Z)) + \beta \text{tr}(ZL_X Z^\top) + \gamma \text{tr}(Z^\top L_C Z),$$

s.t. $Z \in \{-1, 1\}$, $v_1^+ 1_n^1 \leq 1_n^1 Z \leq v_1^- 1_n^1$, $v_2^+ \leq Z 1_n \leq v_2^-$.  

$$\beta$$ and $$\gamma$$ control the trade-off between label consistency and label smoothness. They can be tuned by cross validation. For clarity, we denote the objective function of (4) as $F(Z)$, and define the constraint space as $\Omega_0 = \{Z | v_1^+ 1_n^1 \leq 1_n^1 Z \leq v_1^- 1_n^1, v_2^+ \leq Z 1_n \leq v_2^- \}$, $\Omega_1 = \{-1, 1\} \cap \Omega_0$. Problem (4) is simplified as $\min_{Z \in \Omega_0} F(Z)$.
Proposition 1 Problem (4) is a constrained submodular min-
imization (CSM) problem, and it is NP-hard.

Proposition 2 The Lovasz extension of objective function
(4) is formulated as follows:
\[
f(Z) = -\text{tr}(Y^T_p Z) + \frac{1}{2} \sum_{k} \sum_{i,j} \tilde{W}_k(i,j)|Z_{ki} - Z_{kj}| + \frac{1}{2} \sum_{k} \sum_{i,j} \tilde{W}_k(i,j)|Z_{ik} - Z_{jk}| + \text{const},
\]
where \( \tilde{W}_k(i,j) = 2\gamma W_k(i,j)\sqrt{d_k(i)d_k(j)} \) and
\[
\tilde{W}_C(i,j) = 2\gamma W_C(i,j)\sqrt{d_C(i)d_C(j)}.\]
\[
\text{const} = \text{tr}(Y^T_p Y) + \frac{\beta}{2} \sum_{j} \sum_{i,j} (d_C(i) - d_X(j))^2 + \frac{\beta}{2} \sum_{j} \sum_{i,j} (d_C(i) - d_X(j))^2.
\]

Proposition 3 \( F(Z) \) and \( f(Z) \) satisfy the conditions:
\[
\min_{Z \in \Omega_1} f(Z) \geq \min_{Z \in \Omega_2} f(Z) = \min_{Z \in \{-1,1\}} f(Z) = \min_{Z \in \{-1,1\}} F(Z).
\]
A Convex Approximation to (4)

We will focus on three different approximation methods to
solve (4). The first one is to simply drop the cardinality con-
straints, thus, becoming a submodular optimization problem
\( \min_{Z \in \{-1,1\}} F(Z) \), which can be exactly solved using
the st-cuts algorithm (Boykov and Kolmogorov 2004). The
second approximation is a LP relaxation of the first model
based on Lovasz extension of submodular functions, as
\( \min_{Z \in [-1,1]} f(Z) \) (see Proposition 2). The third method
adds the cardinality constraints to the second one. For sub-
sequent description, the corresponding algorithms of the
second and third methods are referred to as MMB-0 and
MMB, respectively.

Utilizing Proposition 2, we relax (4) to a continuous one:
\[
\min_{Z} f(Z), \quad \text{s.t. } Z \in \Omega_2 = \{-1,1\}^{m \times n} \cap \Omega_0.
\]
As shown in Proposition 3, the original discrete problem (4)
and its approximated continuous problem (6) share the same
lower and upper bounds.

4 ADMM Solution to (6)

To handle the \( \ell_1 \) terms in Problem (6), we introduce two
auxiliary variables \( U_X \) and \( U_C \), as follows:
\[
\min_{Z \in \Omega_2, U_X \geq 0, U_C \geq 0} \text{tr}(U_X W_X^T) + \text{tr}(U_C W_C^T) - \text{tr}(Y^T_p Z),
\]
\[
\text{s.t. } ZA \leq U_X, -ZA \leq U_X, BZ \leq U_C, -BZ \leq U_C,
\]
where \( U_X \in R^{m \times n_x} \) with \( U_X(k, e) = |Z_{ke_1} - Z_{ke_2}| \). Two
nodes \( e_1 \) and \( e_2 \) are connected by the edge \( e \). \( \tilde{W}_X = \)
\[
(w_X, \ldots, w_X)^T \in R^{m \times n_x} \text{ with } w_X(e) = \tilde{W}_X(e_1, e_2), \quad \forall e \in E_X. \]
\( U_C \in R^{n_c \times n} \) with \( U_C(e, k) = |Z_{ke_1} - Z_{ke_2}| \). \( \tilde{W}_C = (w_C, \ldots, w_C) \in R^{n_c \times n} \text{ with } w_C(e) = \tilde{W}_C(e_1, e_2), \quad \forall e \in E_C. \]
As \( A \in \{-1,0\}^{n_c \times n} \), \( A(e_1, e) = 1, A(e_2, e) = -1 \) for \( e \in E_X \), while other entries being 0. \( B \in \{-1,0\}^{n_c \times n} \text{ with } B(e_1) = 1, B(e_2, e) = -1 \) for \( e \in E_C \), while other entries being 0.

The LP problem (7) can be solved by many standard algo-
rithms using off-the-shelf solvers. However, most existing
solvers are designed for vector variables. Although we can
vectorize (7), that will lead to a large LP problem that is inef-
sufficient to solve with standard solvers. Instead, we adopt the
alternating direction method of multipliers (ADMM) (Boyd et
al. 2011), which is known to have good convergence pro-
erties and can take advantage of special structural properties
of our problem.

Following the procedure of the conventional ADMM, we
firstly present the augmented Lagrange function of Problem
(7), by introducing two slack variables \( \Phi_1 \) and \( \Phi_2 \).
\[
L_{\rho_1, \rho_2}(Z, U_X, U_C, \Phi_1, \Phi_2, A_1, A_2) = -\text{tr}(Y^T_p Z) + \text{tr}(W_X^T U_X) + \text{tr}(W_C^T U_C) - \text{tr}(A_1^T (ZA - U_X C) - \Phi_1) - \text{tr}(A_2^T (BZ - D U_C - C_2 + \Phi_2)) + \frac{\rho_1}{2} ||ZA - U_X C - \Phi_1||^2_F + \frac{\rho_2}{2} ||BZ - D U_C - C_2 + \Phi_2||^2_F,
\]
where \( A_1 = [A - A, -1_n, 1_n] \in \{-1,1\}^{n_c \times 2(n_x + 1)} \), \( C_1 = [I, I, 0_n, 0_n] \in \{0,1\}^{n_c \times 2(n+x+1)}, \)
\( C_1 = [0_m \times 2n_x, -1_{m \times n_c}, 0_{1 \times n_c}], \) \( \Phi_1 \in \mathbb{R}^{m \times 2(n_x + 1)} \), and \( \Phi_2 \in \mathbb{R}^{m \times 2(n_x + 1)} \).
\( B = [B; -B; -1_m; 1_m] \in \mathbb{R}^{(2n_c + 2) \times m}, \)
\( D = [I; I; 0_{2 \times n_c}], \) \( D \in \mathbb{R}^{(2n_c + 2) \times (2n_c + 2)}, \) \( C_2 = [0_{2 \times n_c}, -\tau_1(T^T 1_m^T), \tau_1(T^T 1_m^T)], \) \( \Phi_2 \in \mathbb{R}^{(2n_c + 2) \times n}, \)
where \( A_1 \in \mathbb{R}^{m \times 2(n_x + 1)} \) and \( A_2 \in \mathbb{R}^{(2n_c + 2) \times n} \)
denote two Lagrangian parameter matrices, while \( \rho_1, \rho_2 > 0 \)
indicate the trade-off parameters of two augmented terms.
Based on (8), Problem (7) could be solved by following iterative
updates:
\[
Z_{t+1} = \arg \min_{Z \in \{-1,1\}} \text{tr}(M_0^T Z) + \frac{\rho_1}{2} \text{tr}(ZM_1 Z^T) + \frac{\rho_2}{2} \text{tr}(Z^T M_2 Z),
\]
\[
U_X^{t+1} = \max(0, -\frac{1}{2\rho_1} M_X),
\]
\[
U_C^{t+1} = \max(0, -\frac{1}{2\rho_2} M_C),
\]
\[
\Phi_1^{t+1} = \max(0, U_X^{t+1} C_1 + \Phi_1 - \frac{\rho_1}{2} A_1 - Z_{t+1} A_1),
\]
\[
\Phi_2^{t+1} = \max(0, D U_C^{t+1} + C_2 - \frac{\rho_2}{2} A_2 - BZ_{t+1}),
\]
\[
A_1^{t+1} = A_1^T + \rho_1 (Z_{t+1} A_1 - U_X^{t+1} C_1 - \Phi_1^{t+1}),
\]
\[
A_2^{t+1} = A_2^T + \rho_2 (BZ_{t+1} - D U_C^{t+1} - C_2 - \Phi_2^{t+1}).
\]
larger impact of CIB. Moreover, we create $Y_{tr}$ by varying $\varepsilon$ from 0% to 80%. In each case, the missing labels are randomly chosen and this process is conducted 5 times to obtain different missing labels.

**Comparison methods.** As mentioned earlier, our proposed model can be approximated by three optimization problems, which are solved by st-cut, MMIB-0 and MMIB. We compare these methods in our experiments. We obtain st-cut from a publicly available MATLAB toolbox BK.matlab4 and we implement MMIB using MATLAB. Several state-of-the-art multi-label methods that can handle missing labels are compared, including MLR-GL (Bucak, Jin, and Jain 2011), MC-Pos (Cabras et al. 2011), FastTag (Chen, Zheng, and Weinberger 2013), MLML-exact (Wu et al. 2014) and LEML (Yu et al. 2014). We also compare with COCOA (Zhang, Li, and Liu 2015), which is the latest class-imbalance-aware multi-label method. Note that we use the MATLAB code made available for all the aforementioned methods. The binary weighted SVM (W-SVM) is also trained using the LIBSVM toolbox (Chang and Lin 2011), as a baseline classifier, which only trains on labeled instances of each class separately. The weight of each class is set to be the ratio between the number of negative and positive instances in this class. The predicted continuous labels of above methods are finally rounded to discrete labels by setting the threshold as their middle values (0 or $\frac{1}{2}$), while COCOA sets the threshold by maximizing $F_1$ score.

**Evaluation metrics.** Three widely used metrics, example-based $F_1$, $F_1-macro$ and $F_1-micro$, are adopted to evaluate the quality of the predicted label matrix from different perspectives. Their formal definitions can be found in (Sorower 2010). In our experiments, we observe that no single metric is enough to reflect prediction quality. Thus, we also define a new metric: $F_1-mean = \frac{1}{3}(F_1 + F_1-macro + F_1-micro)$.

**Comparison on Handling Missing Labels**

Figure 2 presents the results of all compared methods on handling missing labels.

**MMIB-0 v.s. st-cut.** According to Proposition 3, their objective functions will be equal at the global optimum. However, their solutions tend to be very different. When $\varepsilon = 0\%$, st-cut gives good results on the first three datasets, where the CIB degrees are not very high. On CAL500 and Corel, due to the high CIB, st-cut shows poor performance. This demonstrates the performance of st-cut is significantly influenced by CIB, and it is likely to give a poor solution when the CIB degree is large. Moreover, when $\varepsilon > 0$, the perfor-
Figure 2: Results on all datasets with different missing label proportions.

MIB vs. MMIB. The main difference between them is the class cardinality bounds. Comparing them will evaluate the contribution of the linear constraints that enforce these bounds. MMIB gives better results than MMIB-0 in all cases. Specifically, following the sequence $\varepsilon = (0, 20, 40, 60, 80)$%, the improvements of $F_{1-\text{mean}}$ values are: Emotions $[1.49, 2.2, 2.65, 3.6, 6.54]$%; Scene $[3.9, 4.36, 4.5, 2.08, 1.41]$%; Yeast $[2.3, 2.4, 1.85, 1.59, 3.12]$%; CAL500 $[5.64, 5.01, 4.86, 5.32, 6.15]$%; Corel5k $[4.62, 4.89, 4.42, 5.21, 4.32]$%. These results show that the bounds (embedded as linear constraints) are beneficial to the model performance.

MMIB vs. MMIB-0. MMIB-0 gives consistent improvements over most compared methods in most datasets, while MMIB shows further improvements over MMIB-0. This is due to two main reasons. (1) From the model perspective, both class-level and instance-level label smoothness are used to propagate the label information among different classes and instances. Also, there is no label bias since missing labels are treated as 0. In contrast, other methods except MLML-exact ignore the correlations among different instances, and label bias exists in MLR-GL and FastTag. Note that MLML-exact adopts a similar model. However, there are significant differences between MMIB-0 and MLML-exact. The objective function of MMIB-0 is based on Lovasz extension, which leads to the same objective value as the original discrete objective function, while MLML-exact directly relaxes $\{-1, 1\}$ to $[-1, 1]$. The specific label weights are assigned to different classes and different instances in MMIB-0 through $Y_P$, while uniform weights are adopted in MLML-exact. (2) The linear constraints enforcing class cardinality bounds play an important role in MMIB to alleviate the negative impact of CIB. In contrast, other methods...
ignore this unavoidable challenge, and when $\varepsilon > 0$, the impact of CIB is further observed. Since MLR-GL and FastTag treat missing labels as negative, the imbalance between positive and negative labels will be further amplified. Thus, it is not strange that some methods show very poor (even degenerate) results when $\varepsilon > 0$. Note that there is an exception in the CAL500 dataset, where the $F_{1\text{-}macro}$ values of MLR-GL and FastTag are higher than those of MMIB, but the values of the other metrics are much lower. On this dataset, we see that MLR-GL and FastTag actually give degenerate results (i.e. classes with relatively high rates of positive instances in training have predictions that are always positive in the test). Thus the $F_1$ scores for these classes are very high, leading to a relatively high $F_{1\text{-}macro}$ value. This demonstrates that separate $F_1$ metrics are not enough to reflect prediction quality properly and that the proposed $F_{1\text{-}mean}$ is a more comprehensive metric.

**Comparison on Handling Class Imbalance**

The results on handling CIB are shown in Table 2. As COCOA cannot handle missing labels, we only compare them in the case of $\varepsilon = 0$. The data format of Corel5k does not satisfy the requirement of COCOA, thus it is not tested. COCOA gives better results than W-SVM, while MMIB shows significant improvements over COCOA, up to $[6.1, 3.4, 2.7, 8.3]%$ in $F_{1\text{-}mean}$ on the four datasets, respectively. We believe such an improvement is due to two main reasons. First, COCOA is built on base classifiers (C4.5 decision tree with undersampling), whose performance dictates that of COCOA. In contrast, the class cardinality bounds adopted in our model are independent of any classifier. Second, COCOA only considers CIB-1, while MMIB enforces both CIB-1 and CIB-2.

**Sensitivity Analysis of Cardinality Bounds**

The estimated class bounds $\{v_1^l, v_1^p, v_2^l, v_2^p\}$ based on $Y_{tr}$ may not be exactly the ground-truth values. However, the proposed model is robust to large variations in these bounds. To verify this, we test bound sensitivity on the challenging dataset CAL500. When $\varepsilon$ changes, $v_1^l, v_1^p$ do not vary much, while $v_2^l, v_2^p$ may vary in a range. Thus, we focus on the sensitivity test of $v_2^l, v_2^p$. For simplicity, we fix $v_2^l = 0, v_2^p = m$, i.e., CIB-1 is not embedded. We choose the rates $\zeta_1$ and $\zeta_2$ from $\{0, 0.8, 1.2, 1.5, 2\}$ and $\{\infty, 3, 2.5, 2, 1.5\}$ respectively, with $\zeta_1 < \zeta_2$. Using each pair $(\zeta_1, \zeta_2)$, we run our model several times to report the average $F_{1\text{-}mean}$ value. The test results are summarized in Table 3. In every $\varepsilon$, our model can always give satisfactory results in a relatively wide and stable range, e.g., $\zeta_1 \in \{1.2, 1.5, 2\}$ and $\zeta_2 \in \{3, 2.5, 2\}$. Note that the pair $(\zeta_1 = 0, \zeta_2 = \infty)$ means there is no CIB-2 constraint. Compared to other pairs $(\zeta_1, \zeta_2)$, embedding CIB-2 constraints leads to significant improvements in most cases.

**6 Conclusion**

In this work, we propose a unified method to jointly handle missing labels and class imbalance in multi-label learning. To handle missing labels, our method propagates label information from the labeled instances to the unlabeled ones using label consistency and smoothness. The class imbalance problem is solved by the introduction of cardinality bounds over each instance and each class. We provide efficient numerical algorithms that demonstrate the improved performance over state-of-the-art methods on benchmark datasets. Moreover, more useful prior knowledge, such as semantic hierarchy and mutual exclusion, can be naturally incorporated into the proposed linear programming framework as linear constraints, without any changes of the current algorithm. This will be explored in our future work.

**7 Acknowledgments**

This work is supported supported by competitive research funding from King Abdullah University of Science and Technology (KAUST). The participation of Siwei Lyu in this work is partly supported by US National Science Foundation Research Grant (CCF-1319800) and National Science Foundation Early Faculty Career Development (CAREER) Award (IIS-0953373). We thank reviewers for their constructive comments.
References


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