Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Large-Margin Multi-Label Causal Feature Learning

Chang Xu[†] Dacheng Tao[‡] Chao Xu[†]
[†]Key Lab. of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China
[‡]Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney 2007, Australia
xuchang@pku.edu.cn dacheng.tao@uts.edu.au xuchao@cis.pku.edu.cn

Abstract

In multi-label learning, an example is represented by a descriptive feature associated with several labels. Simply considering labels as independent or correlated is crude; it would be beneficial to define and exploit the causality between multiple labels. For example, an image label 'lake' implies the label 'water', but not vice versa. Since the original features are a disorderly mixture of the properties originating from different labels, it is intuitive to factorize these raw features to clearly represent each individual label and its causality relationship. Following the large-margin principle, we propose an effective approach to discover the causal features of multiple labels, thus revealing the causality between labels from the perspective of feature. We show theoretically that the proposed approach is a tight approximation of the empirical multi-label classification error, and the causality revealed strengthens the consistency of the algorithm. Extensive experimentations using synthetic and real-world data demonstrate that the proposed algorithm effectively discovers label causality, generates causal features, and improves multi-label learning.

Introduction

In the conventional single-label learning scenario, an example is associated with a single label that characterizes its property. In many real-world applications, however, an example will naturally have several class labels. For example, an image of a natural scene can simultaneously be annotated with 'sky', 'mountains', 'trees', 'lakes', and 'water'. Multi-label learning (Luo et al. 2013b; 2013a; Xu, Yu-Feng, and Zhi-Hua 2013; Bi and Kwok 2014; Doppa et al. 2014) has emerged as a new and increasingly important research topic which has the capacity to handle such tasks.

The most straightforward solution to multi-label learning is to decompose the problem into a series of binary classification problems, one for each label (Boutell et al. 2004). However, this solution is limited because it neglects to take the relationships between labels into account. Learning multiple labels simultaneously has been empirically shown to significantly improve performance relative to independent label learning, especially when there are insufficient training examples for some labels. It is therefore advantageous to exploit the relationships between labels for learning, and some works have already used this strategy. For example, (Cai and Hofmann 2004; Cesa-Bianchi, Gentile, and Zaniboni 2006; Rousu et al. 2005; Hariharan et al. 2010; Bi and Kwok 2011) utilize external knowledge to derive the label relationships, such as knowledge of label hierarchies and label correlation matrices. Since these external knowledge resources are often unavailable for real-world applications, other studies (Sun, Ji, and Ye 2008; Tsoumakas et al. 2009; Petterson and Caetano 2011) have attempted to exploit label relationships by counting the co-occurrence of labels in the training data.

In practice, the label relationship is asymmetric rather than symmetric, as assumed by most of the existing multi-label learning algorithms. For example, an image labeled 'lake' implies a label 'water', but the inverse is not true. Only a small number of works have tried to exploit this asymmetric label relationship. For example, in (Zhang and Zhang 2010), a Bayesian network was used to characterize the dependence structure between multiple labels, and a binary classifier was learned for each label by treating its parental labels in the dependence structure as additional input features. (Huang, Yu, and Zhou 2012) assumed that if two labels are related, the hypothesis generated for one label can be helpful for the other label, and implemented this idea as a boosting approach with a hypothesis reuse mechanism.

The feature of each example in multi-label learning determines the appearance of its labels, thus the feature itself can be seen as a disorderly mixture of the properties originating from diverse labels. To comprehensively understand the asymmetric label relationship, we define the relationship as *causality*, and propose to reveal the causality from the perspective of feature. Moreover, there is a demand for the theoretical results to guarantee that exploiting causality is actually beneficial for multi-label learning.

In this paper, we intend to transform the original features of examples shared by different labels into causal features corresponding to each individual label. Following the large-margin principle, we propose a new algorithm termed Large-margin Multi-label Causal Feature learning (LMCF) to achieve this aim. The discovered causal features will reveal causality between labels from the perspective of feature. Geometrically, the causality is encoded by the 'margins' corresponding to different labels on the hyperplane of causal features. By encouraging the margins to be large while satisfying the causality constraints, the optimal causal features will be discovered. We theoretically show that the proposed approach yields a tight approximation to the empirical multi-label classification error, and the exploited causality will further strengthen the consistency of the algorithm. Lastly, we conduct experiments on synthetic data to explicitly illustrate the discovered causality and show, using real-world datasets, that our approach effectively improves the performance of multi-label learning.

Motivation and Notation

In a multi-label learning problem, we are given training data $\{(x_i, y_i)\}_{i=1}^N$, where each example x_i is randomly drawn from an input space $\mathcal{X} \subseteq \mathbb{R}^D$, and the corresponding $y_i = [y_{i1}, \cdots, y_{iL}]$ in the label space $\mathcal{Y}$ contains its L possible labels. If x_i has the j-th label, y_{ij} is 1 and -1 otherwise. The aim of multi-label learning is then to predict the L possible labels for each example from its unique feature vector. Since the feature is a disorderly mixture of the properties originating from different labels, it is reasonable to assume that the causality between different labels will reflect on the features as well. Hence, to reveal the causality from the perspective of feature, we propose to learn the causal features corresponding to different labels for each example.

Instead of directly learning a function $f : \mathcal{X} \to \mathcal{Y}$ for multi-label prediction, we attempt to make the prediction on a space $\mathcal{T} \subseteq \mathbb{R}^d$, which contains the causal features corresponding to different labels of each example. In particular, for an example x_i and its L labels $\{y_{ij}\}_{j=1}^L$, we use a set of linear transformations $U = [U_1, \cdots, U_L]$ to obtain the causal features for different labels:

$$t_{ij} = U_j x_i, \quad \forall j \in [1, L]. \tag{1}$$

Our prediction for the labels is parametrized as

$$f(t_{ij}; w) = w^T t_{ij}, \quad \forall j \in [1, L], \tag{2}$$

where $w \in \mathbb{R}^d$ is shared by different labels.

Large-margin Multi-label Causal Feature Learning

Given two labels y_i and y_j (e.g., 'lake' and 'water'), if label y_i implies label y_j, we define $y_i \to y_j$; otherwise, $y_i \leftarrow y_j$. To propagate the causality from the labels to the features, we next analyze what properties the causal features in the new space $\mathcal{T}$should have.

For an example x, its two labels y_i (e.g., 'lake') and y_j (e.g., 'water') have the causal features t_i and t_j, respectively. Given the hyperplane w shared by different labels in space $\mathcal{T}$, the causality $(y_i = 1) \to (y_j = 1)$ implies that if $f_i = w^T t_i \geq 0$, then there must exist $f_j = w^T t_j \geq 0$, in other words, we have $0 \leq f_i \leq f_j$. Geometrically, the constraints on $w^T t_i$ and $w^T t_j$ require that both t_i and t_j are in the positive region, and t_j is farther from the hyperplane w than t_i, as shown in Figure 1. On the other hand, if $(y_i = 1) \to (y_j = 1)$ holds, $(y_i = -1) \leftarrow (y_j = -1)$ exists as well. As a result, for this contrapositive causality relationship, both t_i and t_j are in the negative region, and t_i is farther from the hyperplane w than t_j.

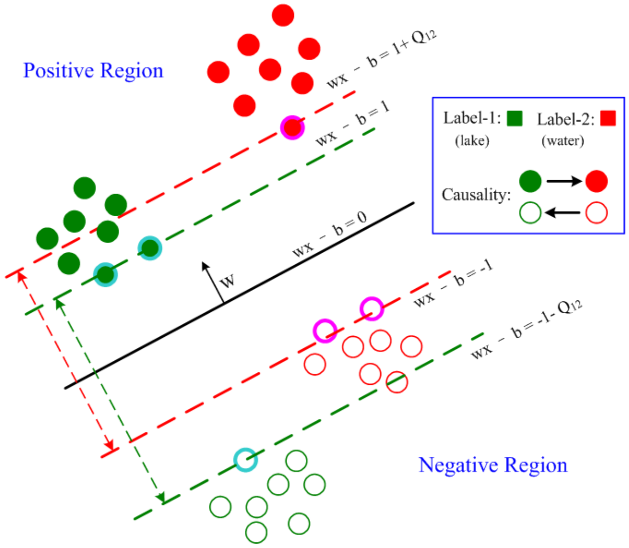

Figure 1: Geometrical illustration of the causal features. Note that we only consider two labels in this example. The marked data points on their corresponding margins are support vectors.

Formally, we employ a non-negative matrix $Q \in \mathbb{R}^{L \times L}$ with diagonal elements as zeros to indicate the causality relationships between labels. In particular, Q_{ij} measures the power that label $y_i = 1$ can infer label $y_j = 1$. Note that Q_{ij} is not forced to be the same with Q_{ji}, since the relationship between two labels maybe asymmetric. For example, an image label 'lake' implies the label 'water', but not vice versa.

By considering the causality between labels, we want to find the maximum-margin hyperplane in space $\mathcal{T}$ that divides the points having $y_i = 1$ from those having $y_i = -1$ for each label y_i. In particular, two hyperplanes can be selected in such a way that they separate the data, with no points between them, and then try to maximize their distance. The region bounded by the two hyperplanes is called "the margin". These hyperplanes can be described by the equations

$$w^T t - b = 1 \quad \texttt{and} \quad w^T t - b = -1,$$

for different labels and features. If we consider the aforementioned geometrical constraints from the causality relationships between labels, the hyperplanes for distinct labels can be different. By counting the influences from other labels, the hyperplanes for label y_i are written as

$$w^T t_i - b = 1 + \sum_{j=1}^{L} Q_{ji} \tag{3}$$

and

$$w^T t_i - b = -1 - \sum_{j=1}^{L} Q_{ij}. \tag{4}$$

By using geometry, maximizing the distance between these two hyperplanes is equivalent to minimizing $\|w\|$. As we also have to prevent data points from falling into the margin as far as possible, we add the following constraint: for the points corresponding to label y_i

$$w^T t_i - b \geq 1 + \sum_{j=1}^{L} Q_{ji} - \xi, \quad \texttt{for} \quad y_i = 1 \tag{5}$$

and

$$w^T t_i - b \geq -1 - \sum_{j=1}^{L} Q_{ij} - \xi, \quad \texttt{for} \quad y_i = -1 \tag{6}$$

where ξ is the non-negative slack variable for each point. This can be rewritten as

$$y_i(w^T t_i - b) \geq 1 + h(y_i, Q, i) - \xi, \tag{7}$$

where $h(y_i, Q, i) = \texttt{cond}(y_i, Q_{(:,i)}, Q_{(i,:)})$, $\texttt{cond}(*)$ is the conditional operator, and $Q_{(:,i)}$ and $Q_{(i,:)}$ are the i-th column summarization and i-th row summarization of matrix Q, respectively. We further define $\widehat{Q} = \{(Q_{ij})^\gamma\}_{i,j}^L$, where γ is the parameter to control the weights distribution.

Following the large-margin principle, we obtain the resulting objective function

$$\begin{aligned} \min_{Q,w,U} \ & \frac{1}{2}\|w\|^2 + \frac{C_1}{NL}\sum_i^N \sum_j^L \xi_{ij} + C_2 \sum_i^L \|U_i\|_F^2 \\ \texttt{s.t.} \ & y_{ij}(w^T U_j x_i - b) \geq 1 + h(y_{ij}, \widehat{Q}, j) - \xi_{ij}, \\ & \xi_{ij} \geq 0, \quad \forall i \in [1,N], \ \forall j \in [1,L], \\ & \|Q\|_1 = L, \ Q \geq 0, \end{aligned} \tag{8}$$

where C_1 and C_2 are non-negative constants that can be determined using cross validation. It is expected that by solving this problem with multi-label examples, the causality between labels can be revealed from the perspective of feature, and the discovered causal features will be beneficial for improving multi-label learning.

Optimization

We solve the optimization Problem 8 in an alternating way. If we focus on the transformation matrix U_j corresponding label-j, while keeping the other transformation matrices and w and Q fixed, we obtain the following sub-problem,

$$\begin{aligned} \min_{U_j} \ & \frac{C_1}{NL}\sum_i^N \xi_{ij} + C_2\|U_i\|_F^2 \\ \texttt{s.t.} \ & y_{ij}(w^T U_j x_i - b) \geq 1 + h(y_{ij}, \widehat{Q}, j) - \xi_{ij}, \\ & \xi_{ij} \geq 0, \quad \forall i \in [1,N]. \end{aligned} \tag{9}$$

The most challenging part arises from the non-smooth hinge loss function. For simplicity, we define

$$s_i = 1 + h(y_{ij}, \widehat{Q}, j) - y_{ij}(w^T U_j x_i - b)). \tag{10}$$

Here we apply the smoothing technique introduced by (Nesterov 2005) to approximate the hinge loss with smooth parameter $\sigma > 0$:

$$\begin{aligned} hinge_\sigma =& z_i s_i - \frac{\sigma}{2}\|x_i\|_\infty z_i^2 \\ \mathcal{Z} =& \{z : 0 \leq z_i \leq 1, z \in \mathbb{R}^n\}, \end{aligned}$$

where z_i can be obtained by setting the gradient of this function as zero and then projecting z_i in $\mathcal{Z}$, i.e.,

$$z_i = \texttt{median}\left\{\frac{s_i}{\sigma\|x_i\|_\infty}, 0, 1\right\}.$$

Therefore, the smoothed hinge loss is a piece-wise approximation of hinge loss according to different choices of z_i,

$$hinge_\sigma = \begin{cases} 0 & z_i = 0, \\ s_i - \dfrac{\sigma}{2}\|x_i\|_\infty & z_i = 1, \\ \dfrac{s_i^2}{2\sigma\|x_i\|_\infty} & else, \end{cases} \tag{11}$$

whose gradient is calculated by

$$\frac{\partial hinge_\sigma}{\partial U_j} = \begin{cases} 0 & z_i = 0, \\ -wx_i^T y_{ij} & z_i = 1, \\ \dfrac{-2s_i(wx_i^T y_{ij})}{2\sigma\|x_i\|_\infty} & else. \end{cases} \tag{12}$$

The gradient is now continuous and gradient descent type methods can be efficiently applied to solve the objective function and find the optimal U_j.

When we fix Q and U, the original problem is reduced to

$$\begin{aligned} \min_w \ & \frac{1}{2}\|w\|^2 + \frac{C_1}{NL}\sum_i^N \sum_j^L \xi_{ij} \\ \texttt{s.t.} \ & y_{ij}(w^T t_{ij} - b) \geq 1 + h(y_{ij}, \widehat{Q}, j) - \xi_{ij}, \\ & \xi_{ij} \geq 0, \quad \forall i \in [1,N], \ \forall j \in [1,L], \end{aligned} \tag{13}$$

which is a SVM problem with adapted margins. We can use the smoothing technique to smooth the loss function in Eq. (13) as well. The gradient descent method can then be straightforwardly applied to solve for w based on this prime problem.

Fixing w and U, Q can be solved by the following Lagrange function:

$$\sum_{i,j}^L \Omega_{ij} Q_{ij}^\gamma - \lambda(\sum_{i,j}^L Q_{ij} - L), \tag{14}$$

where Ω is the constant matrix originating from Eq. (8). To obtain the optimal solution to the above sub-problem, the derivate of Eq. (14) with respect to Q_{ij} is set to zero. We have

$$Q_{ij} = \left(\frac{\lambda}{\gamma\Omega_{ij}}\right)^{\frac{1}{\gamma-1}}. \tag{15}$$

Substituting Eq. (15) into the constraint $\|Q\|_1 = L$, we obtain:

$$Q_{ij} = \frac{L(\gamma\Omega_{ij})^{\frac{1}{1-\gamma}}}{\sum_{i,j}^L(\gamma\Omega_{ij})^{\frac{1}{1-\gamma}}}. \tag{16}$$

The diagonal elements of Q are further set as zeros to complete the update.

Statistical Property

In this section, we provide a statistical interpretation of optimizing Problem 8. Our multi-label learning model is characterized by a distribution $\mathcal{Q}$ on the space of data points and labels $\mathcal{X} \times \{-1, 1\}^L$, where $\mathcal{X} \subseteq \mathbb{R}^D$. We receive N training points $\{(x_i, y_i)\}_{i=1}^N$ sampled i.i.d. from the distribution $\mathcal{Q}$, where $y_i \in \{-1, 1\}^L$ are the ground truth label vectors. Given these training data, we learn feature transformation matrices $\{U_i\}_{i=1}^L$ corresponding to L labels and the weight vector w shared by different labels. Therefore, the aim is to seek a function $f = (f_1, \cdots, f_L) : \mathcal{X} \to \mathbb{R}^L$ such that the prediction error of f given below is as small as possible:

$$L(f(\cdot)) = \mathbb{E}_{(\widetilde{x}, \widetilde{y}) \sim \mathcal{Q}} \left[\sum_{i=1}^{L} I(f_i(\widetilde{x}), \widetilde{y}_i) \right]. \tag{17}$$

Here $I(f_i(\widetilde{x}), \widetilde{y}_i) = \mathbf{1}[\widetilde{y}_i f_i(\widetilde{x}) \leq 0]$. We define $L_k(f_k(\cdot)) = \mathbb{E}_{(\widetilde{x}, \widetilde{y}) \sim \mathcal{Q}} \left[I(f_k(\widetilde{x}), \widetilde{y}_k) \right]$ for k-th label, and represent the loss in Eq. (8) as

$$\phi(x) = \sum_{i=1}^{L} \phi_k(x), \tag{18}$$

where $\phi_k(x) = (1 + h(y_k, Q, k) - y_k f_k(x))_+$.

In the following, we consider the prediction error of label-k for simplicity. Let $\eta_k(x) = \mathbb{P}[\widetilde{y}_k = 1 | \widetilde{x} = x]$. For each x, we seek the minimizer $f_k(x)$ of

$$\begin{aligned} \mathcal{J}(\eta_k, f_k(\cdot)) =& E\left[(1 + h(\widetilde{y}_k, Q, k) - \widetilde{y}_k f_k(x))_+ | \widetilde{x} = x \right] \\ =& \eta_k(x)(1 + Q_{(:,k)} - f_k(x)) \\ &+ (1 - \eta_k(x))(1 + Q_{(k,:)} + f_k(x)). \end{aligned}$$

When $f_k(x) \in [-1 - Q_{(k,:)}, 1 + Q_{(:,k)}]$, we get

$$f_k^*(x) = \begin{cases} 1 + Q_{(:,k)} & \texttt{if } \eta_k(x) > 1/2, \\ -1 - Q_{(k,:)} & \texttt{if } \eta_k(x) < 1/2, \\ 0 & \texttt{if } \eta_k(x) = 1/2, \end{cases} \tag{19}$$

and

$$\mathcal{J}(\eta_k, f_k^*) = (1 - |2\eta_k - 1|) \frac{2 + Q_{(:,k)} + Q_{(k,:)}}{2}. \tag{20}$$

For convenience, we also introduce the notation:

$$\mathcal{J}(\eta_k, f_k) = \eta_k \phi_k(f_k) + (1 - \eta_k) \phi_k(-f_k), \tag{21}$$

$$\Delta\mathcal{J}(\eta_k, f_k) = \mathcal{J}(\eta_k, f_k) - \mathcal{J}(\eta_k, f_k^*). \tag{22}$$

It is then easy to obtain

$$\begin{aligned} \Delta\mathcal{J}(\eta_k, f_k) =& \ \mathcal{J}(\eta_k, f_k) - \mathcal{J}(\eta_k, f_k^*) \\ &= \eta_k(\phi_k(f_k) - \phi_k(f_k^*)) + (1 - \eta_k)(\phi_k(-f_k) - \phi_k^*(-f_k)) \\ &= \eta_k(1 + Q_{(:,k)} - f_k)_+ + (1 - \eta_k)(1 + Q_{(k,:)} + f_k)_+ \\ &- (1 - |2\eta_k - 1|) \frac{2 + Q_{(:,k)} + Q_{(k,:)}}{2}, \end{aligned}$$

which implies that

$$\begin{aligned} \Delta\mathcal{J}(\eta_k, 0) =& 1 + \eta_k(Q_{(:,k)} - Q_{(k,:)}) + Q_{(k,:)} \\ &- (1 - |2\eta_k - 1|) \frac{2 + Q_{(:,k)} + Q_{(k,:)}}{2} \\ &= |2\eta_k - 1| \left(1 + \texttt{cond}(2\eta_k - 1, Q_{(:,k)}, Q_{(k,:)}) \right) \\ &\geq |2\eta_k - 1|. \end{aligned}$$

Hence, we can obtain the following theorem to bound the prediction error of $f_k(\cdot)$ w.r.t. $\phi_k(\cdot)$.

Theorem 1. *For any measurable function $f_k(x)$, we have*

$$\begin{aligned} L_k(f_k(\cdot)) - L_k^* \leq& E_{\widetilde{x}} \Delta\mathcal{J}(\eta_k(\widetilde{x}), f_k(\widetilde{x})) \\ =& E_{\widetilde{x}} \left[\mathcal{J}(\eta_k(\widetilde{x}), f_k(\widetilde{x})) - (1 - |2\eta_k(\widetilde{x}) - 1|) \frac{2 + Q_{(:,k)} + Q_{(k,:)}}{2} \right]. \end{aligned}$$

Proof. By definition of $L(\cdot)$, it is easy to verify that

$$\begin{aligned} L_k(f_k(\cdot)) - L_k(2\eta_k(\cdot) - 1) =& \mathbb{E}_{\eta_k(x) \geq 0.5, f_k(x) < 0}(2\eta_k(x) - 1) \\ &+ \mathbb{E}_{\eta_k(x) < 0.5, f_k(x) \geq 0}(1 - 2\eta_k(x)) \\ &\leq \mathbb{E}_{(2\eta_k(x) - 1) f_k(x) \leq 0} |2\eta_k(x) - 1| \end{aligned}$$

Since $\Delta\mathcal{J}(\eta_k, 0) \geq |2\eta_k - 1|$, we have

$$L_k(f_k(\cdot)) - L_k^* \leq \mathbb{E}_{(2\eta_k(\widetilde{x}) - 1) f(\widetilde{x}) \leq 0} \Delta\mathcal{J}(\eta_k(\widetilde{x}), 0).$$

To complete the proof, since $\Delta\mathcal{J}(\eta_k, f_k) = \mathcal{J}(\eta_k, f_k) - \mathcal{J}(\eta_k, f_k^*)$, it suffices to show that $\mathcal{J}(\eta_k(x), 0) \leq \mathcal{J}(\eta_k(x), f_k(x))$ for all x such that $(2\eta_k(x) - 1) f_k(x) \leq 0$. To see this, we consider the following three cases:

- $\eta_k > 0.5$: From Eq. (19), we have $f_k^*(\eta_k) > 0$. In addition, $(2\eta_k - 1) f_k \leq 0$ implies $f_k \leq 0$. Since $0 \in [f_k, f_k^*(\eta_k)]$ and the convexity of $\mathcal{J}(\eta_k, f_k)$ w.r.t. f_k, we have $\mathcal{J}(\eta_k, 0) \leq \max\{\mathcal{J}(\eta_k, f_k), \mathcal{J}(\eta_k, f_k^*(\eta_k))\} = \mathcal{J}(\eta_k, f_k$.
- $\eta_k < 0.5$: In this case, we have $f_k^*(\eta_k) < 0$ and $f_k \geq 0$, which leads to $0 \in [f_k^*(\eta_k), f_k]$. Thus, $\mathcal{J}(\eta_k, 0) \leq \max\{\mathcal{J}(\eta_k, f_k), \mathcal{J}(\eta_k, f_k^*(\eta_k))\} = \mathcal{J}(\eta_k, f_k)$.
- $\eta_k = 0.5$: Note that $f_k^* = 0$, which implies that $\mathcal{J}(\eta_k, 0) \leq \mathcal{J}(\eta_k, f_k)$ for all f_k.

Given Eq. (22), we then have $\Delta\mathcal{J}(\eta_k, f_k) = \mathcal{J}(\eta_k, f_k) - (1 - |2\eta_k - 1|) \frac{2 + Q_{(:,k)} + Q_{(k,:)}}{2}$. This completes the proof of the theorem. □

Since Theorem 1 holds for any label, we obtain the following corollary.

Corollary 1. *For any measurable function $f = (f_1, \cdots, f_L)$, we have*

$$\begin{aligned} L(f(\cdot)) - L^* \leq& \sum_{k=1}^{L} E_{\widetilde{x}} [\mathcal{J}(\eta_k(\widetilde{x}), f_k(\widetilde{x})) \\ & - (1 - |2\eta_k(\widetilde{x}) - 1|) \frac{2 + Q_{(:,k)} + Q_{(k,:)}}{2}]. \end{aligned}$$

For the N training points $\{(x_i, y_i)\}_{i=1}^N$, the empirical estimation of the bound in Corollary 1 is $\frac{1}{NL} \sum_{i=1}^N \sum_{k=1}^L \phi_k(x_i)$, which implies that optimizing Problem 8 is equivalent to minimizing the empirical bound of the difference between $L(f(\cdot))$ and L^*. Most importantly, the existence of the exploited causality (i.e., $Q_{(:,k)}$ and $Q_{(k,:)}$) in Corollary 1 will tighten this bound, and then strengthen the consistency of the algorithm.

Experiments

In this section, we qualitatively and quantitatively evaluate the proposed LMCF algorithm on synthetic datasets and real-world datasets. The proposed algorithm is compared with RankSVM (Elisseeff and Weston 2001), binary SVM (BSVM) (Boutell et al. 2004), multi-label hypothesis reuse

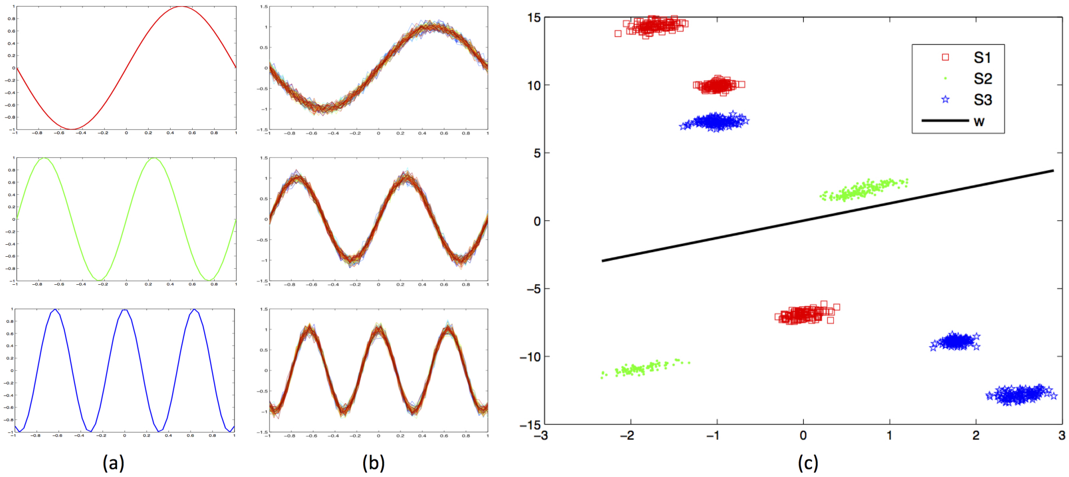

Figure 2: Causal features discovered by LMCF algorithm based on the synthetic data.

(MAHR) (Huang, Yu, and Zhou 2012), multi-label k-nearest neighbors (ML-KNN) (Zhang and Zhou 2007) and ensemble of classifier chains (ECC) (Read et al. 2011). The performances are evaluated through four commonly used multi-label criteria: *Hamming loss*, *One error*, *Ranking loss* and *Average precision*. These criteria measure the performance from different perspectives and their detailed formulations can be found in (Zhou et al. 2012). For the LMCF algorithm, we set $C2 = 0.1$ and $\sigma = 1$, and determine the optimal γ and $C1$ on the validation sets.

Toy Example

We first conducted a toy experiment using synthetic data. This shows our algorithm's ability to correctly discover the causal features corresponding to different labels. In particular, we used the following circular functions to generate the data points of multiple labels,

$$S1 = sin(2\pi t), \quad S2 = 2sin(2\pi t)cos(2\pi t),$$

$$S3 = cos(\pi^2 t),$$

which are displayed in Figure 2 (a). These three kinds of curves (i.e, S1, S2 and S3) are regarded as three labels y_1, y_2 and y_3, respectively. Since S1 is a component of S2, it is reasonable to define the label causality $y_2 \to y_1$. Hence, the data generated from S1 is labeled by $[y_1 = 1, y_2 = -1, y_3 = -1]$, while those from S2 and S3 are $[y_1 = 1, y_2 = 1, y_3 = -1]$ and $[y_1 = -1, y_2 = -1, y_3 = 1]$, respectively. For each curve, we randomly add perturbation, and then uniformly sample 50 points on this perturbed curve in the interval $(-1, 1)$, which leads to a 50-dimensional feature vector. We repeat this procedure 100 times for each curve and finally obtain a synthetic dataset composed of 300 feature vectors with 3 labels, as shown in Figure 2 (b).

Figure 2 (c) depicts the 2-dimensional causal features discovered by the proposed LMCF algorithm. We find that for each label (e.g., y_1), its corresponding features are appropriately clustered. The positive examples (e.g., $y_1 = 1$) and negative examples (e.g., $y_1 = -1$) are separated from each other by the large margin principle. Most importantly, it is instructive to note that the positive features corresponding to $y_2 = 1$ are farther from the hyper-plane w than those positive features corresponding to $y_1 = 1$. Meanwhile, the negative features corresponding to $y_2 = -1$ are closer to the hyper-plane than those negative features corresponding to $y_1 = -1$. This is actually a reflection of label causality on the features. Hence the causality between labels has been effectively propagated to casual features corresponding to different labels by means of the "margin".

Table 2: Average Precision for different low-dimensional causal features on different datasets.

d/L	Yahoo	Enron	Yeast	Scene	Image	Corel5k
5%	0.492	0.522	0.698	0.751	0.698	0.088
10%	0.583	0.576	0.749	0.848	0.763	0.196
20%	0.654	0.604	0.768	0.862	0.812	0.222
40%	0.649	0.629	0.755	0.864	0.818	0.276
80%	0.652	0.635	0.758	0.856	0.798	0.293

Multi-label Classification

Six real-world datasets are used in our experiments. These datasets are extracted from diverse applications: *Yahoo* for web paper categorization, *Enron* for email analysis, *Yeast* for gene function prediction, and *Scene*, *Image* and *Corel5K* for image classification. All these datasets are obtained from the Mulan website.

The comparison results are shown in Table 1. The proposed LMCF algorithm is designed for minimizing the hamming loss. Compared to other methods, LMCF achieves stable performance improvements on hamming loss in most cases; moreover, it obtains comparable performance with that of MAHR, which is designed for optimizing hamming loss in a boosting approach. This reflects the strong discriminative ability of LMCF derived through the large-margin principle. It is instructive to note that LMCF achieves excellent performance for the other three criteria as well, though it does not aim to optimize these criteria.

To examine the influence of the dimensionality of causal features, i.e., the parameter d, we conduct LMCF with different ratios d/L on different datasets. The performances evaluated through hamming loss are presented in Table 2. From this table, we find that the performance of the lower

Table 1: Comparison of LMCF with different multi-label learning approaches on different datasets using several evaluation criteria. '↑ (↓)' indicates the larger (smaller), the better. ●(○) indicates that LMCF is significantly better (worse) than the corresponding method.

	Hamming loss ↓					
	Yahoo	Enron	Yeast	Scene	Image	Corel5k
LMCF	0.042 ± 0.015	0.045 ± 0.004	0.188 ± 0.003	0.075 ± 0.004	0.168 ± 0.005	0.011 ± 0.001
RankSVM	0.042 ± 0.014	0.311 ± 0.367 ●	0.196 ± 0.003 ●	0.251 ± 0.017 ●	0.339 ± 0.021 ●	0.012 ± 0.001
BSVM	0.044 ± 0.016 ●	0.056 ± 0.002 ●	0.189 ± 0.003	0.098 ± 0.002 ●	0.179 ± 0.006 ●	0.009 ± 0.000 ○
MAHR	0.039 ± 0.012 ○	0.047 ± 0.003	0.204 ± 0.004 ●	0.084 ± 0.004 ●	0.169 ± 0.011 ●	0.008 ± 0.002 ○
ML-KNN	0.043 ± 0.014	0.051 ± 0.002 ●	0.196 ± 0.003 ●	0.090 ± 0.003 ●	0.175 ± 0.007 ●	0.009 ± 0.000 ○
ECC	0.049 ± 0.017 ●	0.055 ± 0.002 ●	0.208 ± 0.005 ●	0.095 ± 0.004 ●	0.180 ± 0.010 ●	0.014 ± 0.000 ●

	One error ↓					
	Yahoo	Enron	Yeast	Scene	Image	Corel5k
LMCF	0.389 ± 0.111	0.215 ± 0.035	0.169 ± 0.010	0.155 ± 0.009	0.251 ± 0.020	0.642 ± 0.012
RankSVM	0.412 ± 0.130 ●	0.855 ± 0.020 ●	0.224 ± 0.009 ●	0.457 ± 0.065 ●	0.708 ± 0.052 ●	0.977 ± 0.018 ●
BSVM	0.291 ± 0.016 ○	0.359 ± 0.033 ●	0.217 ± 0.011 ●	0.209 ± 0.014 ●	0.291 ± 0.016 ●	0.768 ± 0.009 ●
MAHR	0.398 ± 0.122 ●	0.234 ± 0.030 ●	0.243 ± 0.011 ●	0.217 ± 0.011 ●	0.301 ± 0.024 ●	0.651 ± 0.013 ●
ML-KNN	0.471 ± 0.157 ●	0.299 ± 0.031 ●	0.235 ± 0.012 ●	0.238 ± 0.012 ●	0.325 ± 0.024 ●	0.740 ± 0.011 ●
ECC	0.391 ± 0.133 ●	0.228 ± 0.036 ●	0.180 ± 0.012	0.232 ± 0.011 ●	0.300 ± 0.022 ●	0.647 ± 0.012

	Ranking loss ↓					
	Yahoo	Enron	Yeast	Scene	Image	Corel5k
LMCF	0.103 ± 0.033	0.113 ± 0.008	0.129 ± 0.009	0.063 ± 0.008	0.156 ± 0.012	0.221 ± 0.007
RankSVM	0.112 ± 0.047 ●	0.267 ± 0.019 ●	0.172 ± 0.006 ●	0.214 ± 0.039 ●	0.463 ± 0.018 ●	0.408 ± 0.035 ●
BSVM	0.100 ± 0.052	0.115 ± 0.008	0.169 ± 0.005 ●	0.070 ± 0.005 ●	0.161 ± 0.009 ●	0.141 ± 0.002 ○
MAHR	0.109 ± 0.046 ●	0.098 ± 0.010 ○	0.184 ± 0.005 ●	0.077 ± 0.006 ●	0.166 ± 0.012 ●	0.310 ± 0.011 ●
ML-KNN	0.102 ± 0.045	0.091 ± 0.008 ○	0.168 ± 0.006 ●	0.083 ± 0.006 ●	0.177 ± 0.013 ●	0.307 ± 0.003 ●
ECC	0.332 ± 0.084 ●	0.246 ± 0.018 ●	0.279 ± 0.011 ●	0.139 ± 0.008 ●	0.247 ± 0.016 ●	0.601 ± 0.006 ●

	Average precision ↑					
	Yahoo	Enron	Yeast	Scene	Image	Corel5k
LMCF	0.665 ± 0.082	0.662 ± 0.018	0.780 ± 0.005	0.871 ± 0.006	0.820 ± 0.005	0.296 ± 0.005
RankSVM	0.658 ± 0.103 ●	0.262 ± 0.017 ●	0.767 ± 0.007 ●	0.698 ± 0.047 ●	0.516 ± 0.011 ●	0.067 ± 0.007 ●
BSVM	0.662 ± 0.089	0.578 ± 0.019 ●	0.771 ± 0.007 ●	0.876 ± 0.008	0.808 ± 0.010	0.214 ± 0.003 ●
MAHR	0.660 ± 0.098 ●	0.678 ± 0.020 ○	0.749 ± 0.007 ●	0.869 ± 0.006 ●	0.804 ± 0.014 ●	0.254 ± 0.003 ●
ML-KNN	0.625 ± 0.117 ●	0.636 ± 0.015 ●	0.762 ± 0.010 ●	0.857 ± 0.007 ●	0.788 ± 0.012 ●	0.242 ± 0.005 ●
ECC	0.616 ± 0.092 ●	0.637 ± 0.021 ●	0.731 ± 0.007 ●	0.846 ± 0.007 ●	0.789 ± 0.014 ●	0.227 ± 0.004 ●

Table 3: Example related labels discovered on Enron and Corel5k datasets.

	Enron			Corel5k	
jubilation	dislike	legal document	water	art	grass
camaraderie	political influence	company business	pool	carvings	sheep
friendship	regulations and regulators	dislike	stream	paintings	meadow
include new text in forwarding	meeting minutes	government report	lake	sculpture	tundra

dimensional causal features is limited, whereas with the increased d, LMCF will discover the effective causal features and achieve stable performance.

We examine the causality discovered on different datasets and show the example causalities in Table 3. It can be seen that the discovered label causality is reasonable. For example, we dislike strict rules and regularizations, and 'grass' is likely to appear with 'sheep'.

Conclusion

In contrast to existing approaches that exploit label correlations in multi-label learning, we assume that the relationship between labels is asymmetric and define this as *causality*. To obtain an in-depth comprehension of the connections between features and labels, we factorize the original features shared by multiple labels into causal features corresponding to different labels. Inspired by the large-margin principle, the causality between labels is interpreted as the margin related to different causal features, which enables us to reveal the label causality from the perspective of feature. The proposed approach is theoretically shown to be a tight approximation of the empirical multi-label classification error, and the exploited causality is beneficial for strengthening the consistency of the algorithm. Experiments on synthetic datasets and real-world datasets demonstrate the effectiveness of the proposed algorithm to discover the causality and improve the performance of multi-label learning.

Acknowledgments

The work was supported in part by Australian Research Council Projects FT-130101457, DP-140102164 and LP-140100569, NSFC 61375026, 2015BAF15B00 and JCYJ 20120614152136201.

References

Bi, W., and Kwok, J. T. 2011. Multi-label classification on tree-and dag-structured hierarchies. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 17–24.

Bi, W., and Kwok, J. T. 2014. Multilabel classification with label correlations and missing labels. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*.

Boutell, M. R.; Luo, J.; Shen, X.; and Brown, C. M. 2004. Learning multi-label scene classification. *Pattern recognition* 37(9):1757–1771.

Cai, L., and Hofmann, T. 2004. Hierarchical document categorization with support vector machines. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 78–87. ACM.

Cesa-Bianchi, N.; Gentile, C.; and Zaniboni, L. 2006. Hierarchical classification: combining bayes with svm. In *Proceedings of the 23rd international conference on Machine learning*, 177–184. ACM.

Doppa, J. R.; Yu, J.; Ma, C.; Fern, A.; and Tadepalli, P. 2014. Hc-search for multi-label prediction: An empirical study. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*.

Elisseeff, A., and Weston, J. 2001. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, 681–687.

Hariharan, B.; Zelnik-Manor, L.; Varma, M.; and Vishwanathan, S. 2010. Large scale max-margin multi-label classification with priors. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 423–430.

Huang, S.-J.; Yu, Y.; and Zhou, Z.-H. 2012. Multi-label hypothesis reuse. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 525–533. ACM.

Luo, Y.; Tao, D.; Xu, C.; Li, D.; and Xu, C. 2013a. Vector-valued multi-view semi-supervsed learning for multi-label image classification. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*.

Luo, Y.; Tao, D.; Xu, C.; Xu, C.; Liu, H.; and Wen, Y. 2013b. Multiview vector-valued manifold regularization for multilabel image classification. *IEEE transactions on neural networks and learning systems* 24(5):709–722.

Nesterov, Y. 2005. Smooth minimization of non-smooth functions. *Mathematical programming* 103(1):127–152.

Petterson, J., and Caetano, T. S. 2011. Submodular multilabel learning. In *NIPS*, 1512–1520.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine learning* 85(3):333–359.

Rousu, J.; Saunders, C.; Szedmak, S.; and Shawe-Taylor, J. 2005. Learning hierarchical multi-category text classification models. In *Proceedings of the 22nd international conference on Machine learning*, 744–751. ACM.

Sun, L.; Ji, S.; and Ye, J. 2008. Hypergraph spectral learning for multi-label classification. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 668–676. ACM.

Tsoumakas, G.; Dimou, A.; Spyromitros, E.; Mezaris, V.; Kompatsiaris, I.; and Vlahavas, I. 2009. Correlation-based pruning of stacked binary relevance models for multi-label learning. In *Proceeding of ECML/PKDD 2009 Workshop on Learning from Multi-Label Data, Bled, Slovenia*, 101–116. Citeseer.

Xu, M.; Yu-Feng, L.; and Zhi-Hua, Z. 2013. Multi-label learning with pro loss. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*.

Zhang, M.-L., and Zhang, K. 2010. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 999–1008. ACM.

Zhang, M.-L., and Zhou, Z.-H. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition* 40(7):2038–2048.

Zhou, Z.-H.; Zhang, M.-L.; Huang, S.-J.; and Li, Y.-F. 2012. Multi-instance multi-label learning. *Artificial Intelligence* 176(1):2291–2320.