To Avoid the Pitfall of Missing Labels in Feature Selection: A Generative Model Gives the Answer

Yuanyuan Xu,[1] **Jun Wang,**[2] **Jinmao Wei**[1*]
[1]College of Computer Science, Nankai University, Tianjin 300071, China
[2]College of Mathematics and Statistics Science, Ludong University, Shandong 264025, China
{xuyuanyuan, junwang}@mail.nankai.edu.cn, weijm@nankai.edu.cn

Abstract

In multi-label learning, instances have a large number of noisy and irrelevant features, and each instance is associated with a set of class labels wherein label information is generally incomplete. These missing labels possess two sides like a coin; people cannot predict whether their provided information for feature selection is favorable (relevant) or not (irrelevant) during tossing. Existing approaches either superficially consider the missing labels as negative or indiscreetly impute them with some predicted values, which may either overestimate unobserved labels or introduce new noises in selecting discriminative features. To avoid the pitfall of missing labels, a novel unified framework of selecting discriminative features and modeling incomplete label matrix is proposed from a generative point of view in this paper. Concretely, we relax *Smoothness Assumption* to infer the label observability, which can reveal the positions of unobserved labels, and employ the spike-and-slab prior to perform feature selection by excluding unobserved labels. Using a data-augmentation strategy leads to full local conjugacy in our model, facilitating simple and efficient Expectation Maximization (EM) algorithm for inference. Quantitative and qualitative experimental results demonstrate the superiority of the proposed approach under various evaluation metrics.

Introduction

Feature selection is an important part in machine learning, which aims at selecting the most discriminative features to reduce dimensionality and computation costs, improve the learning performance of models, and better understand the inherent regularities in data (Guyon and Elisseeff 2003). In the feature selection for multi-label cases, most approaches (Gu, Li, and Han 2011; Ma et al. 2012a; 2012b) capture label correlations to help better select the most discriminative features across multiple labels, which can improve the learning performance in prediction phrase. Nevertheless, these approaches have all assumed training data with complete label assignments.

In practice, however, it is difficult to attain complete labels; a considerable number of labels are left aside by labelers, intentionally or accidentally. Taking an intentional example, in image or text learning, human labelers tend to only annotate a few keyword labels that describe the most obvious visual or semantic contents, a.k.a. “positive labels”. Those rare or ambiguous contents are commonly omitted without being labeled, which we call “unobserved labels” in this work. From the accidental aspect, labelers could hardly be exposed to all of the samples in massive scale, which also leads to the existence of numerous unobserved labels.

A critical issue is that the candidate label set is typically huge, and the labels that are irrelevant with the described items are also abandoned by labelers, a.k.a. “negative labels”. For example, the label “iceberg” could hardly be associated with a “desert” image. This comes to a problem that, negative labels and unobserved labels are compounded in the pool of “missing label”. This issue has not received enough attention by existing feature selection approaches; yet, it is crucial because the nature of two kinds of labels is entirely disparate. They make different contributions in determining whether a feature is discriminative for the learning prototype; the negative labels provide definite negative feedback information, while the unobserved labels give nothing back. This is the core motivation of our study, that is, we attempt to accurately trace the unobserved labels and preclude them from the process of feature selection.

Existing multi-label feature selection to deal with the incomplete label cases includes the label embedding line and the imputation line. The former family of approaches simply treats all of the missing labels as the negative ones, and the latter family randomly assumes the positions of the unobserved labels in the incomplete label matrix and predicts their positive or negative values. The manipulation of the label embedding family is apparently inconsistent with the reality, and that of the imputation family seems superficially complicated, because (1) randomly positioning the unobserved labels is unreasonable; and (2) no method could promise an absolutely perfect prediction for the unobserved labels, and an indiscreet prediction inevitably introduces new noises and further deteriorates the performance of feature selection.

In this study, we concentrate on tackling the following challenges to prevent the participation of the unobserved la-

[*]Corresponding Author

bels in feature selection, considering the possible negative effects caused by their ambiguous nature: (1) accurately discriminating the unobserved labels in the missing label pool; (2) finding a compromise way to both adequately take advantage of the known label information and avoid the introduction of the label noises; and (3) with the support of the credible label information obtained by solving (1) and (2), selectively shrinking the weights of features to facilitate the discriminative features prevail.

In this paper, we propose a novel generative probabilistic framework to conduct multi-label feature selection with incomplete labels, named GMFS (**G**enerative **M**ulti-label **F**eature **S**election). Concretely, the binary latent indicator variables are incorporated into the generative model to trace the positions of the unobserved labels. To precisely estimate these indicator variables, we relax the traditional *Smooth Assumption* for explicit observations of labels. Then based on the indication of these variables, label correlations are captured entirely on the credible labels and employed to guide feature selection, which helps GMFS avoid the pitfall of the unobserved labels. In addition, the golden standard for sparse learning, i.e., spike-and-slab prior, is incorporated into GMFS, to separately model discriminative features and irrelevant features with different priors. This strategy leads to the aggressive shrinkage of irrelevant features as well as preserves discriminative features for rare labels.

In summary, the main contributions are as follows:

- We firstly make an in-depth view into the disparity of missing labels, and differentially treat the unobserved ones and the negative ones in feature selection through a generative probabilistic model;
- A fusion of both terms, i.e., label observability inference by relaxing *Smooth Assumption* and sparse learning via spike-and-slab prior, contributes to ambiguity-free guide and selective shrinkage in feature selection;
- An extensive empirical evaluation is conducted to quantitatively and qualitatively assess the selection performance of GMFS, validating its superiority in the incomplete learning scenarios.

Background

Smoothness Assumption has been widely applied in multi-label learning for label completion, which describes that *similar instances share similar labels* (Ma et al. 2012b; Zhu et al. 2018). Taking the text annotation for an example, we can find some similar description words (labels) in the Olympic reports respectively about swimming and diving, such as "champion", "aquatics", and "athlete". While we observe that similar instances also possess their personalized labels, e.g., the swimming reports may have words "breaststroke" and "relay", which could hardly appear in the diving reports, and vice versa. Thus we argue that *Smoothness Assumption* is too strong to be directly applied in the incomplete learning scenarios in this work, because it emphasizes the uniformity of similar instances, while neglects their distinctiveness. In order to accurately trace the unobserved labels, we firstly relax the *Smoothness Assumption* as that *similar instances have similar probabilities to achieve the same observed labels*, and employ it to handle the label observability rather than predict the missing label values, which is less restrictive and seems closer to reality.

Spike-and-Slab Prior (SSP) is a golden standard in the Bayesian variable selection (Ishwaran, Rao, and others 2005), which takes the marginal posterior of a variable as its selection probability. A variable ω following SSP is sampled from a linear combination of two distributions:

$$\omega \sim \pi \mathcal{N}(\mu, \sigma^2) + (1-\pi)\delta_0, \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ is the slab prior, which is modeled using a Gaussian distribution with mean μ and variance σ^2, and δ_0 is the spike prior, which is modeled using a Dirac delta mass function centered at zero. SSP can assign the non-zero probability for the event $\omega = 0$ as $p(\omega = 0) = 1 - \pi$. Therefore, SSP is an ideal distribution for variable selection. However, the Dirac delta function in SSP complicates its inference. In this work, we present a variant of SSP to perform sparse feature selection with a fast inference.

The Proposed Framework

Preliminaries

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \in \mathbb{R}^{n \times d}$ denote the instance matrix, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the i-th instance and n is the total number. $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\} \in \{0,1\}^{n \times l}$ denotes the label matrix, where l is the number of labels. $y_{ij} = 1$ represents a positive label, while $y_{ij} = 0$ offers two possibilities, i.e., negative or unobserved (Jain, Modhe, and Rai 2017).

Formulation

GMFS is a generative model with fusing the unobserved label inference and sparse feature selection. Fig. 1 shows our model in the graphical representation. In this section, we first design a latent factor model coupled with indicator variables to trace the unobserved labels, and relax the *Smoothness Assumption* to estimate these variables by modeling the label observability. Next, based on the guide of these indicator variables, we accomplish the sparse feature selection through the spike-and-slab prior. Finally, inference is performed via the expectation maximization algorithm.

Modeling Label Matrix. We assume that each instance $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ is associated with a latent semantic factor $\boldsymbol{v_i} \in \mathbb{R}^c$. It can be interpreted as clustering the original l labels into c different clusters, and each cluster has a specific semantic meaning. For example, the labels "horse" and "cattle" are more likely to be categorized into the same group because of encoding similar semantic information. Each label $j = 1, \ldots, l$ is associated with a label latent coefficient factor $\boldsymbol{b}_j \in \mathbb{R}^c$, which can be interpreted as its coefficient w.r.t. these c semantic clusters. Inspired by the exploration of the *exposure* variable in recommendation system (Liang et al. 2016), we define a binary latent indicator variable ρ_{ij} to infer the observability of the label y_{ij} from data. Then, the prior distributions of the latent factors $\boldsymbol{v_i}, i = 1, \ldots, n$, $\boldsymbol{b}_j, j = 1, \ldots, l$, and binary latent variable ρ_{ij} are given as

$$\boldsymbol{v}_i | \boldsymbol{x}_i, \boldsymbol{W} \sim \mathcal{N}(\boldsymbol{v}_i | \boldsymbol{W}^T \boldsymbol{x}_i, \lambda_v^{-1} \mathbf{I}_c), \tag{2}$$

$$\boldsymbol{b}_j \sim \mathcal{N}(\boldsymbol{b}_j | \mathbf{0}, \lambda_b^{-1} \mathbf{I}_c), \tag{3}$$

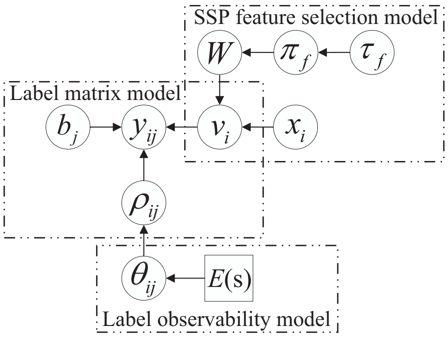


Figure 1: Graphical model of the proposed GMFS.

$$\rho_{ij} \sim \text{Bernoulli}(\theta_{ij}), \tag{4}$$

where θ_{ij} is the prior probability of the latent indicator variable ρ_{ij}, $\mathbf{I}_c$ is the identity matrix, and λ_v and λ_b are hyperparameters that denote the precision of distributions and inverse variances. $\boldsymbol{W} \in \mathbb{R}^{d \times c}$ is the feature coefficient matrix whose rows measure the importance of features in approximating the latent semantics $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^{n}$. We condition $\boldsymbol{v}_i$ on $\boldsymbol{W}$ and $\boldsymbol{x}_i$ by assuming its prior distribution depends on the distribution of $\boldsymbol{x}_i$ in the reduced feature space. Conditioned on these latent factors, y_{ij} is estimated as

$$y_{ij}|\boldsymbol{v}_i, \boldsymbol{b}_j \sim \begin{cases} \text{Bernoulli}(y_{ij}|\sigma(\boldsymbol{v}_i^T \boldsymbol{b}_j)), & \text{if} \quad \rho_{ij} = 1 \\ \delta_0, & \text{if} \quad \rho_{ij} = 0 \end{cases}, \tag{5}$$

where $\sigma(\cdot)$ denotes the logistic function; δ_0 denotes the unobserved cases, that is, $p(y_{ij} = 0|\rho_{ij} = 0) = 1$. $\rho_{ij} = 1$ indicates that y_{ij} is an observed label, and y_{ij} could be 1 or 0, depending on the outcome of the Bernoulli draw. Formally, y_{ij} is modeled by fusing two distributions as

$$y_{ij} \sim \rho_{ij}\text{Bernoulli}(y_{ij}|\sigma(\boldsymbol{v}_i^T \boldsymbol{b}_j)) + (1-\rho_{ij})\mathbb{I}[y_{ij} = 0], \tag{6}$$

where $\mathbb{I}[\cdot]$ is the indicator function. As shown in Eq. (6), the latent indicator variable ρ_{ij} decides which distribution from this mixture generates y_{ij}. ρ_{ij} for the positive label (i.e., $y_{ij} = 1$) is equal to 1 with probability 1; in other words, we only need to infer entry whose corresponding y_{ij} is equal to 0, that is, the missing label. By inferring ρ, we can get an explicit indication of where the unobserved labels are.

Modeling Label Observability. To accurately trace the unobserved labels in the missing label pool, the probability parameter θ_{ij} in Eq. (4) is critical, and we can regard it as the observation probability of the j-th label for the i-th instance. Here, we propose a general method, which takes advantage of instance similarity information to capture the label observability, and this strategy can be augmented with freely available external knowledge in real applications. Concretely, we relax the *Smoothness Assumption* to learn θ_{ij}, because the traditional *Smoothness Assumption* emphasizes the uniformity of two similar instances, while neglects their distinctiveness. A negative example is that a user can obtain the recommendation of a product from her friend, while whether she will buy this product is finally determined by her personal preference. In addition, we choose the Beta distribution Beta(α, β) as the label-dependent conjugate prior for θ_{ij}, which is defined as follows:

$$\theta_{ij} = o_{ij} + E(s), E(s) = \sum_{e \in N_k(i)} s \cdot \theta_{ej}, \tag{7}$$

where o_{ij} is the inner observance of the instance $\boldsymbol{x}_i$ toward the j-th label, s is the coefficient of effects from neighbors, and $N_k(i)$ denotes the k-nearest neighbors of $\boldsymbol{x}_i$ computed by RBF kernel. We employ instance correlations to infer the positions of unobserved labels, facilitating excluding them from the subsequent spare feature selection. Here, credible label correlations can be explored via latent semantics $\boldsymbol{V}$ to guide feature selection process.

Modeling Sparse Feature Selection. The generative model specified in the label matrix modeling module controls features through the feature coefficient matrix $\boldsymbol{W}$ and the features that are most related to the latent semantic matrix $\boldsymbol{V}$ are highly scored. In this situation, however, the features discriminative to rare labels are dominated by the features for common labels, and their weights may iteratively shrink to zero and finally they will be lost in a single sparse regularization. In this study, we tackle this issue by placing a spike-and-slab prior over $\boldsymbol{W}$ for selective shrinkage. Concretely, a set of latent selection variables $\{\pi_f\}, f = 1, \ldots, d$ are introduced to indicate the feature selection: $\pi_f = 1$ means the f-th feature is selected; otherwise, it is unselected. The spike-and-slab prior over $\boldsymbol{W}$ is assigned as

$$\boldsymbol{w}_f|\pi_f \sim (1-\pi_f)\mathcal{N}(\boldsymbol{w}_f|0, \sigma_0) + \pi_f\mathcal{N}(\boldsymbol{w}_f|0, \sigma_1), \tag{8}$$

$$\pi_f|\tau_f \sim \text{Bernoulli}(\tau_f), \tag{9}$$

where σ_0 and σ_1 are the variances of two Gaussian components (i.e., $\sigma_0 \ll \sigma_1$), $\mathcal{N}(\boldsymbol{w}_f|0, \sigma_0)$ is similar to point-mass at zero yet is more robust, and $\tau_f \in [0, 1]$ represents the selection probability of the f-th feature. If the f-th feature is selected, the slab prior over $\boldsymbol{w}_f$ has large variance σ_1 that determines the discrepancy between nonzero weights and, if not, the spike prior has very small variance σ_0, leading to aggressive shrinkage of the irrelevant or noisy features. This strategy promises the discriminative features to be maximally preserved, which will be further demonstrated in inference. Here, we treat τ_f as a random variable with the noninformative Jeffreys prior (Grazian and Robert 2018), dispensing with any hyperparameters. To quantify the selection uncertainty, we marginalize the latent selection variables $\{\pi_f\}$, and sort features in a descending order according to $\|\boldsymbol{w}_f\|_2$ and select the top-ranked ones (Jian et al. 2016).

Inference

We use EM to find the maximum a *posteriori* estimates of the unknown parameters of the model. To develop efficient algorithm for doing inference, we leverage the recently developed Pólya-gamma augmentation techniques (Polson, Scott, and Windle 2013) to handle these non-conjugate likelihoods due to the presence of the logistic-Bernoulli likelihood and the Gaussian prior for the model parameters and are able to transform these likelihoods into Gaussian likelihoods, when conditioned on auxiliary variables.

The pólya-gamma augmentation technique (Polson, Scott, and Windle 2013) is based on the following identity

$$\frac{(\exp(\psi))^{\varsigma_0}}{(1+\exp(\psi))^{\varsigma_1}} = 2^{-\varsigma_1}\exp(\kappa\psi)\int_0^\infty \exp(-\eta\psi^2/2)p(\eta)d\eta, \quad (10)$$

where $\kappa = \varsigma_0 - \varsigma_1/2$ and $p(\eta) = \mathrm{PG}(\varsigma_1, 0)$ denotes the pólya-gamma distribution.

Specially, using PG augmentation, we can write the logistic-Bernoulli likelihood from Eq. (6) as a Gaussian when conditioned on $\eta_{ij} \sim \mathrm{PG}(1, \boldsymbol{v}_i^T\boldsymbol{b}_j)$. In particular, $\psi_{ij} = \boldsymbol{v}_i^T\boldsymbol{b}_j$, conditioned on η_{ij}, becomes a Gaussian

$$p(\psi_{ij}|\eta_{ij}) \propto \exp(\kappa_{ij}\psi_{ij} - \frac{1}{2}\eta_{ij}\psi_{ij}^2), \quad (11)$$

where $\kappa_{ij} = y_{ij} - 0.5$. This likelihood with the Gaussian priors on the latent factors $\boldsymbol{v}_i$ and $\boldsymbol{b}_j$ results in Gaussian posteriors on $\boldsymbol{v}_i$ and $\boldsymbol{b}_j$. When doing EM, this also leads to subproblems that are like least square regression problems.

The EM algorithm for our model alternates between computing the expectations of the local latent variables, namely the pólya-gamma variables $\{\eta_{ij}\}$ and the binary latent indicator variables $\{\rho_{ij}\}$ in the E step, and then using these expectations to estimate the latent semantic factor $\boldsymbol{V}$, latent label coefficient factor $\boldsymbol{B}$, observation probabilities $\{\theta_{ij}\}$, and selection probabilities $\{\tau_f\}$ in the M step.

The E step: The **E** step involves computing the expectations of the latent variables $\{\eta_{ij}\}$ and $\{\rho_{ij}\}$, given the current values of other model parameters estimated in the previous M step. The E step update equations are given below:

- Expectations of Pólya-gamma variables $\{\eta_{ij}\}, \forall i, j$ are known to be available in closed form (Scott and Sun 2013), and are given by

$$\zeta_{ij} = \mathbb{E}[\eta_{ij}|\psi_{ij}] = \frac{1}{2\psi_{ij}}\tanh(\frac{\psi_{ij}}{2}), \quad (12)$$

where $\psi_{ij} = \boldsymbol{v}_i^T\boldsymbol{b}_j$ is computed using the estimates of $\boldsymbol{v}_i$ and $\boldsymbol{b}_j$ from the previous M step.

- Expectations of each of the binary latent indicator variables $\{\rho_{ij}\}, \forall i, j$, are given by

$$\gamma_{ij} = \mathbb{E}[\rho_{ij}|\psi_{ij}] = \frac{\theta_{ij}\sigma(-\psi_{ij})}{\theta_{ij}\sigma(-\psi_{ij}) + (1-\theta_{ij})}. \quad (13)$$

Note that if $y_{ij} = 1$, then $\gamma_{ij} = 1$. In this paper, $\{\gamma_{ij}\}$ are only imposed to distinguish between unobserved labels and negative labels.

The M step: Given the expectations of the latent variables computed in the E step and marginalized the selection variables $\{\pi_f\}$, the log posteriori objective function is optimized in M step, which denotes as $\mathcal{Q}(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{\theta})$

$$\begin{aligned}\mathcal{Q}(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{\theta}) &= -\frac{1}{2}\sum\nolimits_{i,j}\gamma_{ij}\frac{(\kappa_{ij} - \zeta_{ij}\boldsymbol{v}_i^T\boldsymbol{b}_j)^2}{\zeta_{ij}} \\ &+ \sum\nolimits_{i,j}\log\mathrm{Bernoulli}(\gamma_{ij}|\theta_{ij}) - \lambda_v\sum\nolimits_{i=1}^n\|\boldsymbol{v}_i - \boldsymbol{W}^T\boldsymbol{x}_i\|^2 \\ &+ \sum\nolimits_{f=1}^d\log((1-\tau_f)\mathcal{N}(\boldsymbol{w}_f|0,\sigma_0) + \tau_f\mathcal{N}(\boldsymbol{w}_f|0,\sigma_1)) \\ &+ \sum\nolimits_{i,j}\log\mathrm{Beta}(\theta_{ij}|\alpha,\beta) - \lambda_b\sum\nolimits_{j=1}^l\|\boldsymbol{b}_j\|^2.\end{aligned} \quad (14)$$

Note that the first term given in Eq. (14) is due to the logistic likelihood transformed into a Gaussian (using PG augmentation). The term is akin to a weighted least square objective where each label being associated with a weight γ_{ij}. Intuitively, the contribution of each label y_{ij} to the log-likelihood gets modulated based on its expressed label observability. When label y_{ij} is unobserved label (i.e., $\gamma_{ij} = 0$), this label does not contribute to numerical computation in M step. Maximizing $\mathcal{Q}(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{\theta})$ respectively w.r.t. $\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{\theta}$, yield closed-form updates for each of these. The updates are as follows:

- Estimating each of the latent semantic factors $\{\boldsymbol{v}_i\}_{i=1}^n$ is a weighted ridge-regression problem with solution

$$\boldsymbol{v}_i = (\sum_{j=1}^l \gamma_{ij}\zeta_{ij}\boldsymbol{b}_j\boldsymbol{b}_j^T + \lambda_v\mathbf{I}_c)^{-1}(\sum_{j=1}^l \gamma_{ij}\kappa_{ij}\boldsymbol{b}_j + \lambda_v\boldsymbol{W}^T\boldsymbol{x}_i). \quad (15)$$

Note that if the j-th label is unobserved or unobserved with high probability for the i-th instance, i.e., γ_{ij} is zero or small value, it hardly contributes to the update of $\boldsymbol{v}_i$, which available label correlations are captured only depending on credible label information to help better steer feature selection process. And the updates for $\{\boldsymbol{v}_i\}_{i=1}^n$ are all independent of each other and are easily parallelized.

- Estimating each of the latent coefficient factors $\{\boldsymbol{b}_j\}_{j=1}^l$ is a weighted ridge-regression problem with solution

$$\boldsymbol{b}_j = (\sum_{i=1}^n \gamma_{ij}\zeta_{ij}\boldsymbol{v}_i\boldsymbol{v}_i^T + \lambda_b\mathbf{I}_c)^{-1}(\sum_{i=1}^n \gamma_{ij}\kappa_{ij}\boldsymbol{v}_i). \quad (16)$$

Note that if the j-th label is unobserved or unobserved with high probability for the i-th instance, it hardly contributes to the update of $\boldsymbol{b}_j$. And the updates for $\{\boldsymbol{b}_j\}_{j=1}^l$ are all independent of each other and are easily parallelized.

- In Eq. (14), we marginalize out $\{\pi_f\}$ and jointly update over $\boldsymbol{W}$ and the selection probabilities $\{\tau_f\}$. Given $\boldsymbol{W}$ as fixed, the update for τ_f is that $\tau_f = 1$ if $|\boldsymbol{w}_f| \geq \xi$, and $\tau_f = 0$ otherwise, where $\xi = \sqrt{(\frac{2\sigma_0\sigma_1}{\sigma_1 - \sigma_0})\log\sqrt{\frac{\sigma_1}{\sigma_0}}}$. Here if $\tau_f = 1$, we can infer $\pi_f = 1$ based on Bernoulli draw that f-th feature is relevant, slab prior is imposed to model it that this feature information can be preserved; if $\tau_f = 0$, we can infer $\pi_f = 0$ that f-th feature is irrelevant or noisy, spike prior leads to aggressive shrinkage of f-th feature. Moreover, based on a moderate ξ, informative features discriminative to rare labels can be selected to improve the feature selection performance.

- Given $\boldsymbol{\tau}$, the update of $\boldsymbol{W}$ has a closed form solution for regression that is a special case of generalized ridge-regression (Hoerl and Kennard 1970):

$$\boldsymbol{W} = (\lambda_v\boldsymbol{X}^T\boldsymbol{X} + \mathrm{diag}(\boldsymbol{A}))^{-1}\lambda_v\boldsymbol{X}^T\boldsymbol{V}, \quad (17)$$

where $\boldsymbol{A}$ is such that $a_f = (\frac{1}{\sigma_1})^{\tau_f}(\frac{1}{\sigma_0})^{(1-\tau_f)}, f = 1, \ldots, d$. The update of $\boldsymbol{W}$ has a time complexity of $O(d^3)$. This is prohibitively expensive at higher dimensions. However, the EM algorithm does not require solving for $\boldsymbol{W}$ exactly in each M step. Therefore, we solve for $\boldsymbol{W}$ efficient using gradient based methods, such as conjugate-gradient (CG) method (Bertsekas 1997), which also allows us to leverage the sparsity in the feature matrix. Typically, a small number of CG iterations are sufficient in practice.

• Given γ_{ij} from E step, maximizing the objective function w.r.t. θ_{ij} is equivalent to finding the mode of the complete conditional Beta($\alpha + \sum_{u=1}^{n} \gamma_{uj} + (s-1)\sum_{e \in N_k(i)} \gamma_{ej}, \beta + n - \sum_{e \in N_k(i)} \gamma_{ej}$), which is

$$\theta_{ij} \longleftarrow \frac{\alpha + \sum_{u=1}^{n} \gamma_{uj} + (s-1)\sum_{e \in N_k(i)} \gamma_{ej} - 1}{\alpha + \beta + n + (s-1)\sum_{e \in N_k(i)} \gamma_{ej} - 2}, \tag{18}$$

where $s \geq 1$ is the coefficient of instance similarity effects.

In conclusion, we iteratively optimize the variables and infer them to reach the following goals: (1) the latent indicator variables $\{\rho_{ij}\}$ are learned to trace unobserved labels; (2) based on $\{\rho_{ij}\}$, the latent semantic factor $\boldsymbol{V}$ is learned to encode the label correlations; (3) the selection probabilities $\{\tau_f\}$ are learned to provide a pronounced selective shrinkage property for features; and (4) the feature coefficient matrix $\boldsymbol{W}$ is learned to remove irrelevant and noisy features on the basis of the $\boldsymbol{V}$ and $\boldsymbol{\tau}$. Our objective function in Eq. (14) is maximized to comprehensively model incomplete labels and sparse feature selection in the multi-label scenarios.

Complexity Analysis

The cost of the inference consists of two parts: (1) In E step, the variables $\{\eta_{ij}\}$ and $\{\rho_{ij}\}$ are computed that costs $O(Tnlc)$, where T is the number of iterations; and (2) In M step, when inferring the latent factor $\boldsymbol{V}$ and $\boldsymbol{B}$, the complexity is at least $\mathcal{O}(Tnc^2 + Tndc + Tlc^2)$. As c is smaller than l and n, the complexity can be calculated as $\mathcal{O}(Tnd + Tl)$, where they are easily parallelized. We infer the parameter $\boldsymbol{W}$ using CG method, the time complexity achieves $O(Td)$ speedup. The cost of updating variables $\{\theta_{ij}\}$ becomes $\mathcal{O}(Tnlk + n^2)$, where k is very small than n. Hence, the overall complexity is $O(Tnl + Tnd + n^2)$, where similarity computational cost $\mathcal{O}(n^2)$ can be avoided via employing additional knowledge in practice.

Related Work

The existing feature selection methods to handle the missing labels can be mainly divided into two groups: the label embedding line and the imputation line. The label embedding methods decompose the incomplete labels and project them to a low-dimensional space. For instance, Jian et al. (Jian et al. 2016) imposed Latent Semantic Index to decompose the multi-label output space, and Braytee et al. (Braytee et al. 2017) employed non-negative matrix factorization to simultaneously decompose the original data and label matrix. These approaches regard all missing labels as the negative ones, which is inconsistent with common sense, while we consider two sides of missing labels, exclude its negative effects and explore its positive potential in feature selection.

On the other hand, the imputation approaches randomly assume the positions of unobserved labels and complete them with pre-assumed values. For example, Zhu et al. (Zhu et al. 2018) employed robust linear regression combined with *Smoothness Assumption* to simultaneously recover unobserved labels and select informative features, which may introduce label noises and degrade feature selection performance. In contrast, we trace the positions of unobserved labels via a generative model and disregard their recoveries, to avoid introducing noises in feature selection.

Another related field to our study is the sparse learning (Bradley and Mangasarian 1998; Nie et al. 2010; Liu et al. 2013; Jian et al. 2016; Xu et al. 2018). In the regularization theory literature, sparse feature selection is generally tackled via l_1 or $l_{2,1}$ regularization, which requires expensive cross-validation and restricts discriminative and irrelevant features to the same regularization level (Rendle 2010). Our model employs the spike-and-slab prior to provide selective shrinkage property, which can effectively capture the sparse feature structure.

Experimental Study

We evaluate our approach on seven groups of multi-label data sets fetched from Mulan library [1] and the "yahoo.com", as shown in Table 1.

Table 1: Data sets description

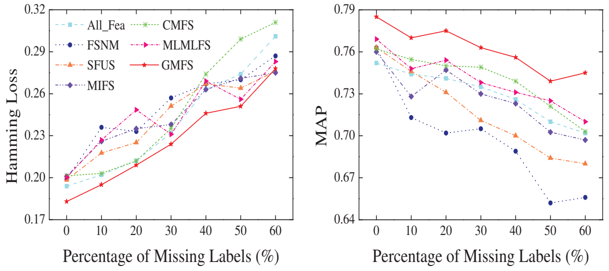
Data sets	Instances	Features	Labels	Domain
Emotions	593	72	6	music
Yeast	2417	103	14	biology
Science	5000	743	40	text
RCV1S1	6000	944	101	text
RCV1S2	6000	944	101	text
Bibtex	7395	1836	159	text
Delicious	16105	500	983	text

Experimental Settings

In this section, we compare GMFS with the following approaches: two state-of-the-art sparse feature selection methods (i.e., FSNM (Nie et al. 2010) and SFUS (Ma et al. 2012b)), two label embedding methods (i.e., MIFS (Jian et al. 2016) and CMFS (Braytee et al. 2017)), an imputation method (i.e., MLMLFS (Zhu et al. 2018)), and All$-$Fea without any feature selection.

For our model, we fix the parameters λ_v, λ_b, α, β, and s to 10^{-3}, 10^{-3}, 0.5, 0.5, and 5, which works well on most data sets we experimented with. We tune the parameters c (number of latent factors), k (number of nearest neighbors), and the variances for spike-and-slab components (i.e., σ_0 and σ_1) using cross-validation. The grids of c, σ_0, and σ_1 are $c = [2, \frac{l}{4}, \frac{2l}{4}, \frac{3l}{4}, l]$, $\sigma_0 = [10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}]$, and $\sigma_1 = [1 : 1 : 5]$. For conjugate gradient (CG) method used in the M step of our inference algorithm, we run five iterations for sufficient valuation. To simulate the situation of label incompleteness, the missing label ratio is set to 20% and 40% by randomly dropping the observed labels from the training data (Bucak, Jin, and Jain 2011). Multi-label libSVM (Chang and Lin 2011) with RBF kernel is chosen

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



(a) Hamming Loss (b) MAP

Figure 2: Variations of Hamming Loss and MAP with increasing the percentage of missing labels on Yeast.

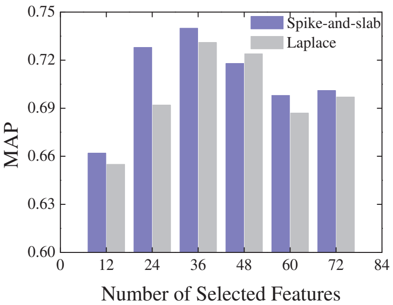


Figure 3: Comparison of two sparse priors.

as the classifier due to its effectiveness verified in many state-of-the-art works (Ma et al. 2012a; Jian et al. 2016; Braytee et al. 2017). Each compared approach respectively selects $\{\frac{d}{6}, \frac{2d}{6}, \frac{3d}{6}, \frac{4d}{6}, \frac{5d}{6}\}$ features to build the multi-label libSVM classifier, where d is the number of the original features. We report the mean average precision (MAP) and standard deviation averaged across five independent runs on each dataset with different size of features.

Classification Performance

The feature selection performance respectively with 20% and 40% missing labels are recorded in Table 2 and Table 3. Here, we examine pairwise t-test on the experimental results to show whether the performance of GMFS is significantly different to the baselines over benchmarks.

According to the experimental results, we observe that (1) GMFS achieves approximately $2\% - 5\%$ improvements over All-Fea across the benchmarks, which means that a discriminative reduced space by removing noises and irrelevant features is beneficial for learning prototypes; (2) GMFS yields better performance than the label embedding approaches (i.e., MIFS and CMFS), which demonstrates that the low-dimensional semantic space of GMFS wherein the unobserved labels are excluded can encode inherent label correlations to help find discriminative features; (3) GMFS is generally better than the imputation approach (i.e., MLMLFS), which indicates that the imputed value may be imperfect and bring additional noises for feature selection, which is also one of the major issues addressed by GMFS; (4) SFUS and FSNM yield relatively inferior performance, since they are incapable of handling missing labels, which are pervasive in practice while beyond their discussions.

Effects of the Size of Missing Labels

To evaluate the effects of label incompleteness on feature selection, we vary the missing label ratio on the Yeast benchmark as $\{10\%, 20\%, 30\%, 40\%, 50\%, 60\%\}$ and show the performance of the compared approaches under Hamming Loss and MAP in Fig. 2. The followings can be observed. (1) GMFS yields better performance than baselines in most learning cases, which indicates that credible label correlations that are captured based on available positive and negative labels are beneficial for improving feature selection performance; (2) With increasing the scale of missing labels, the selection performance of GMFS tends to deteriorate, since the available label information gradually decreases and further weakens its ability of guiding feature selection; and (3) The sparse feature selection approaches (i.e., FSNM and SFUS) perform relatively inferior, especially when the missing label ratio is high, since these approaches are weak of handling missing labels; meanwhile, this also points out the significance of appropriately recognizing and employing missing labels to guide feature selection.

Effects of Sparse Feature Selection

In this section, we compare GMFS with spike-and-slab prior and GMFS with Laplace prior (i.e., l_1 regularization and regularization parameter is decided using cross validation), to evaluate their impacts when selecting different sizes of features. Emotions with 20% missing labels is used as the benchmark. Fig. 3 shows the performance of GMFS with two different priors when increasing the selected features from 12 to 72 (i.e., the number of the original features). We observe that (1) GMFS with spike-and-slab prior generally outperforms GMFS with Laplace prior, which attributes to the selective shrinkage property that facilitates the informative features and especially those discriminative to rare labels to prevail, while conducting single l_1 regularization may lose these partial informative features due to its identical shrinkage on all features; (2) the performance of GMFS is improved with expanding the size of selected feature subset from 12 to 36, due to more excellent features are selected and included in this subset, and the highest MAP score is achieved when the selected number is equal to 36; and (3) the performance of GMFS exhibits a declining trend when the selected number continuously rises, till all of the original features are selected, where a large number of redundant features appear in the selected subset. Thus, we can conclude that it is difficult to find a definite optimal number of selected features at which the best performance can be achieved by different approaches across different data sets. In consideration of this issue and for a fair comparison, we vary the number of selected features and report the average selection performance for each baseline throughout our experiments.

Table 2: MAP score(±standard deviation) when missing label ratio is 20%. The best result and those not significantly worse than it are highlighted in bold (pairwise t-test at 5% significance level).

Data sets	Approaches						
	All−Fea	FSNM	SFUS	MIFS	CMFS	MLMLFS	GMFS
Emotions	0.685±0.023	0.639±0.043	0.679±0.006	0.693±0.016	0.680±0.022	0.707±0.066	**0.732±0.019**
Yeast	0.741±0.011	0.702±0.029	0.731±0.023	0.747±0.027	0.750±0.010	0.754±0.009	**0.775±0.007**
Science	0.558±0.010	0.526±0.090	0.540±0.047	0.543±0.053	0.544±0.074	0.562±0.038	**0.589±0.008**
RCV1S1	0.519±0.006	0.500±0.041	0.467±0.012	0.501±0.006	0.521±0.006	0.525±0.003	**0.554±0.006**
RCV1S2	0.538±0.010	0.463±0.070	0.514±0.021	0.527±0.017	0.537±0.005	0.542±0.004	**0.579±0.011**
Bibtex	0.547±0.009	0.453±0.012	0.478±0.017	0.476±0.064	0.533±0.003	0.557±0.016	**0.578±0.008**
Delicious	0.496±0.007	0.473±0.016	0.499±0.026	0.516±0.005	0.524±0.033	0.510±0.030	**0.547±0.013**

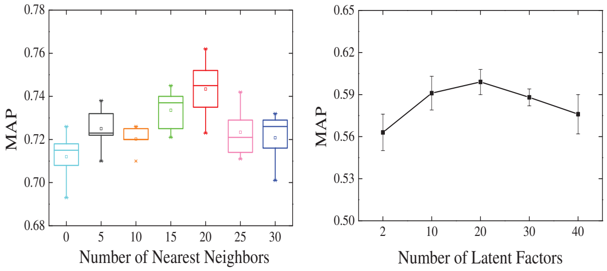
Table 3: MAP score(±standard deviation) when missing label ratio is 40%.

Data sets	Approaches						
	All−Fea	FSNM	SFUS	MIFS	CMFS	MLMLFS	GMFS
Emotions	0.632±0.030	0.626±0.067	0.649±0.009	0.628±0.032	0.623±0.028	0.646±0.029	**0.697±0.027**
Yeast	0.726±0.009	0.689±0.033	0.700±0.006	0.723±0.013	0.739±0.014	0.731±0.027	**0.756±0.010**
Science	0.527±0.013	0.493±0.121	0.516±0.012	0.520±0.021	0.527±0.009	0.526±0.059	**0.550±0.010**
RCV1S1	0.483±0.012	0.451±0.035	0.443±0.013	0.487±0.029	0.482±0.004	0.475±0.005	**0.512±0.009**
RCV1S2	0.504±0.010	0.414±0.081	0.480±0.015	0.471±0.011	0.508±0.012	0.507 ±0.009	**0.531±0.007**
Bibtex	0.506±0.006	0.424±0.079	0.443±0.031	0.464±0.014	0.483±0.006	0.490±0.015	**0.523±0.004**
Delicious	0.461±0.016	0.442±0.028	0.459±0.033	0.478±0.020	0.491±0.022	0.483±0.019	**0.514±0.017**

Parameter Analysis

The scale of instance correlations in Eq. (7) (i.e., $N_k(i)$) plays an important role in estimating the label observability. In this subsection, we evaluate its effects by varying the number of nearest neighbors k from 0 to 30, utilizing Emotions with 20% missing labels as the benchmark. Fig. 4a shows that when k is set to 0, we cannot extract instance correlations to infer label observability and GMFS performs relatively inferior. This further reveals that GMFS relies on the credible labels to guide its feature selection, which is consistent with the observations in the above experiments. When $k > 0$, the instance correlations facilitate effective inference of label observability, which exposes the positions of unobserved labels and paves the way for ambiguity-free feature selection. A small or large scale of nearest neighbors may hamper label observability, because of missing useful neighbor information or incorporating noisy information. The results of GMFS reported throughout the paper are obtained with a moderate value of k as 20.

We also conduct experiments to evaluate the effects of the number of latent factors by varying c from 2 to 40, and demonstrate the MAP score of GMFS on the Science data set with 20% missing labels. The results in Fig. 4b indicate that a moderate number of semantic clusters contributes to capturing available label correlations, which facilitates the selection of relevant features. The optimal cluster number for the Science benchmark is 20, and a smaller or larger scale of semantic clusters may disturb the extraction of true label correlations. In our experiments, the optimal number of latent factors is captured by cross validation.



(a) Optimize k parameter (b) Optimize c parameter

Figure 4: Variations of MAP on GMFS w.r.t. different parameters

Discussion

Conjecturing the intentions of human labelers is a complicated issue, because it is complex to decide whether the labels are intentionally missed or accidentally missed by labelers. We endeavor to address this issue via a generative framework from the probabilistic view in this study.

The missing labels construct an unknown world, and we can expect to distill favorable information from this world and also should be alert of its pitfall. Dozens of approaches have stressed its positive side and devoted efforts to missing information completion, while we focus on excluding its negative effects and exploring its potential in feature selection. In other words, it is unnecessary to recover missing labels in the learning process, which is essentially different with missing label learning. Inevitably, our selection performance would decay to some extent when facing a large ratio of missing labels, while it still performed better than the imputation method, as shown in the experiments.

We fuse sparse feature selection and label observability exploration in a generative framework, which makes it fundamentally different from a few generative multi-label learning models (Jain, Modhe, and Rai 2017; Gaure et al. 2017). Some of the latest works on generative models (Wei, Cao, and Philip 2016; Guan, Dy, and Jordan 2011) solved feature selection problems in unsupervised learning scenarios. However, these methods did not have a mechanism to provide a selective shrinkage and did not approach missing label problems in multi-label feature selection.

Conclusion and Future work

We propose a generative probabilistic framework for multi-label feature selection with incomplete labels. We incorporate a set of latent indicator variables into a latent factor model to trace the unobserved labels in the missing label pool, and relax the *Smoothness Assumption* to infer label observability. Then, the spike-and-slab prior is employed to select features based on the credible label correlations. Finally, a fast and efficient EM algorithm is developed for inference.

In terms of the scalability, we can easily extend our model to a *mixture* of latent factor model, to more adequately capture the label structure. Another interesting and possible extension would be the zero-shot learning (Mensink, Gavves, and Snoek 2014), which can help our model handle the new labels at the test phrase. We have tried to perform in a mini-batch fashion by using an online EM algorithm (Cappe and Moulines 2009), which helps our model scale to massive data sets, and we will improve it in the future.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China(61772288), the Natural Science Foundation of Tianjin City(18JCZDJC30900) and the Shandong Provincial Natural Science Foundation(ZR2019MA049).

References

Bertsekas, D. P. 1997. Nonlinear programming. *Journal of the Operational Research Society* 48(3):334–334.

Bradley, P. S., and Mangasarian, O. L. 1998. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, 82–90.

Braytee, A.; Liu, W.; Catchpoole, D. R.; and Kennedy, P. J. 2017. Multi-label feature selection using correlation information. In *CIKM*, 1649–1656.

Bucak, S. S.; Jin, R.; and Jain, A. K. 2011. Multi-label learning with incomplete class assignments. In *CVPR*, 2801–2808.

Cappe, O., and Moulines, E. 2009. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(3):593–613.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3):27.

Gaure, A.; Gupta, A.; Verma, V. K.; and Rai, P. 2017. A probabilistic framework for zero-shot multi-label learning. In *UAI*, volume 1, 3.

Grazian, C., and Robert, C. P. 2018. Jeffreys priors for mixture estimation: Properties and alternatives. *Computational Statistics & Data Analysis* 121:149–163.

Gu, Q.; Li, Z.; and Han, J. 2011. Correlated multi-label feature selection. In *CIKM*, 1087–1096.

Guan, Y.; Dy, J. G.; and Jordan, M. I. 2011. A unified probabilistic model for global and local unsupervised feature selection. In *ICML*, 1073–1080.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3(Mar):1157–1182.

Hoerl, A. E., and Kennard, R. W. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67.

Ishwaran, H.; Rao, J. S.; et al. 2005. Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics* 33(2):730–773.

Jain, V.; Modhe, N.; and Rai, P. 2017. Scalable generative models for multi-label learning with missing labels. In *ICML*, 1636–1644.

Jian, L.; Li, J.; Shu, K.; and Liu, H. 2016. Multi-label informed feature selection. In *IJCAI*, 1627–1633.

Liang, D.; Charlin, L.; McInerney, J.; and Blei, D. M. 2016. Modeling user exposure in recommendation. In *WWW*, 951–961.

Liu, X.; Wang, L.; Zhang, J.; Yin, J.; and Liu, H. 2013. Global and local structure preservation for feature selection. *IEEE Transactions on Neural Networks and Learning Systems* 25(6):1083–1095.

Ma, Z.; Nie, F.; Yang, Y.; Uijlings, J. R.; Sebe, N.; and Hauptmann, A. G. 2012a. Discriminating joint feature analysis for multimedia data understanding. *IEEE Transactions on Multimedia* 14(6):1662–1672.

Ma, Z.; Nie, F.; Yang, Y.; Uijlings, J. R.; and Sebe, N. 2012b. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Transactions on Multimedia* 14(4):1021–1030.

Mensink, T.; Gavves, E.; and Snoek, C. G. 2014. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2441–2448.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint l2,1-norms minimization. In *NIPS*, 1813–1821.

Polson, N. G.; Scott, J. G.; and Windle, J. 2013. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association* 108(504):1339–1349.

Rendle, S. 2010. Factorization machines. In *ICDM*, 995–1000.

Scott, J. G., and Sun, L. 2013. Expectation-maximization for logistic regression. *arXiv preprint arXiv:1306.0040*.

Wei, X.; Cao, B.; and Philip, S. Y. 2016. Unsupervised feature selection on networks: a generative view. In *AAAI*.

Xu, Y.; Wang, J.; An, S.; Wei, J.; and Ruan, J. 2018. Semi-supervised multi-label feature selection by preserving feature-label space consistency. In *CIKM*, 783–792.

Zhu, P.; Xu, Q.; Hu, Q.; Zhang, C.; and Zhao, H. 2018. Multi-label feature selection with missing labels. *Pattern Recognition* 74:488–502.