Label-Sensitive Task Grouping by Bayesian Nonparametric Approach for Multi-Task Multi-Label Learning

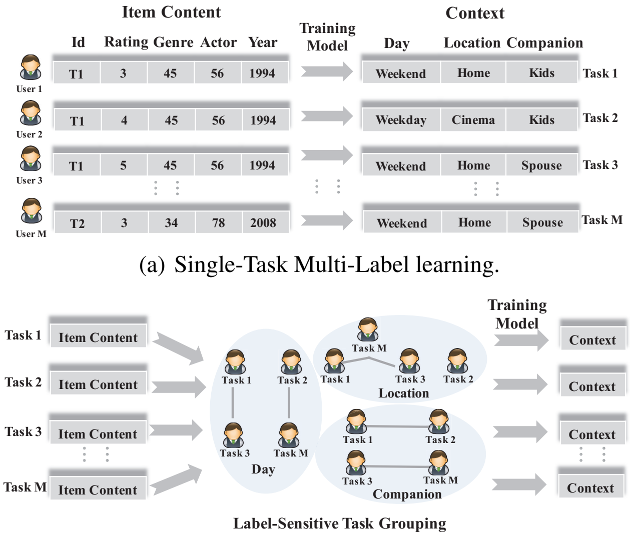
Xiao Zhang[1], **Wenzhong Li**[1], **Vu Nguyen**[2], **Fuzhen Zhuang**[3], **Hui Xiong**[4], **Sanglu Lu**[1]
[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] Center for Pattern Recognition and Data Analytics, Deakin University, Australia
[3] Key Lab of IIP of CAS, Institute of Computing Technology, CAS Beijing, China
[4] Management Science & Information Systems, Rutgers University, USA

Abstract

Multi-label learning is widely applied in many real-world applications, such as image and gene annotation. While most of the existing multi-label learning models focus on the single-task learning problem, there are always some tasks that share some commonalities, which can help each other to improve the learning performances if the knowledge in the similar tasks can be smartly shared. In this paper, we propose a *LAB*el-sensitive *TA*sk *G*rouping framework, named LABTAG, based on Bayesian nonparametric approach for multi-task multi-label classification. The proposed framework explores the label correlations to capture feature-label patterns, and clusters similar tasks into groups with shared knowledge, which are learned jointly to produce a strengthened multi-task multi-label model. We evaluate the model performance on three public multi-task multi-label data sets, and the results show that LABTAG outperforms the compared baselines with a significant margin.

1 Introduction

Recent years have witnessed the vast amount of interest and research in multi-label learning for various kinds of applications, such as image annotation [Gong *et al.*, 2013], gene annotation [Li *et al.*, 2012], sentiment classification [Liu and Chen, 2015] and so on. A common approach to multi-label classification is to transform the multi-label data set to one multi-class data set or multiple single-label data sets. In this faction, single-label classifiers can be applied independently, such as the binary relevance method [Boutell *et al.*, 2004], label powerset method [Tsoumakas *et al.*, 2011], etc. Several works were proposed to exploit label correlations for multi-label classification. For example, the Classifier Chain model and its variants are utilized to depict the label correlations [Dembczynski *et al.*, 2010; Read *et al.*, 2011]. In addition, the label dependencies can be further represented based on Bayesian graphical model or conditional random field [Ghamrawi and McCallum, 2005; Zhang and Zhang, 2010; Guo and Gu, 2011]. Moreover, a Bayesian nonparametric approach was introduced to capture the feature-label correlation patterns for multi-label classification [Nguyen *et al.*, 2016].



(a) Single-Task Multi-Label learning.

(b) Multi-Task Multi-Label learning.

Figure 1: Example of context recommendation.

However, most existing multi-label learning models focus on single-task learning problems. They assume different tasks are independent and learn a classification model for each task individually. In practice, tasks could be highly correlated, and the performance of multiple classification tasks can be improved by learning them jointly. For example, in the spam-filter problem, each user receives only a portion of spams, and learning a model across multiple users (tasks) can obtain a stronger filter for spam detection. The Multi-task learning ap-

This work was partially supported by the National Key R&D Program of China (Grant No. 2017YFB1001801), the National Natural Science Foundation of China (Grant Nos. 61672278, 61473273, 61773361, 61373128, 61321491), the science and technology project from State Grid Corporation of China (Contract No. SGSNXT00YJJS1800031), the NSFC 91746301, the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Sino-German Institutes of Social Computing. The corresponding authors are Wenzhong Li (lwz@nju.edu.cn) and Fuzhen Zhuang (zhuangfuzhen@ict.ac.cn). Fuzhen Zhuang is also with University of Chinese Academy of Sciences, Beijing, 100049, China.

proach provides a good solution to explore similarities among multiple tasks [Bakker and Heskes, 2003; Xue *et al.*, 2007; Kim and Xing, 2010; Chen *et al.*, 2012], which can be incorporated with multi-label learning to learn knowledge from similar tasks to enhance the overall performance.

Taking the Context Recommendation Problem [Zheng *et al.*, 2014] as an example: given a movie item information, we want to recommend suitable contexts (time, location, companion, etc) for a user to watch the movie. As shown in Fig. 1(a): the problem corresponds to learn a mapping $\{$User, Item$\} \rightarrow \{$Context$\}$ that takes user(s) and movie item(s) as input to suggest a list of contexts for the user. Here user information includes the user ID, the user profile (e.g., gender), and movie item information includes the movie ID, the actors, the director, the ratings, etc. The suggested contexts to watch the movie could be the Day (e.g., "Weekend" or "Weekday"), the Location (e.g., "Cinema" or "Home"), the Companion (e.g., "Kids" or "Spouse"), etc. Traditionally, we can apply single-task learning to training a multi-label classification model for each user individually. However, when the data set is sparse (e.g., each user has very few items and contexts), the performance could be degraded due to the shortage of training instances. To overcome the drawback, we adopt the idea of multi-task learning to explore the correlations among similar tasks. Fig. 1 shows the idea of treating context recommendation as a multi-task multi-label learning problem. The tasks are latently related label-sensitively, which means the tasks can be clustered into different groups with respect to different labels, e.g., in Fig. 1(b), Task 1 and 3 are clustered into a group and Task 2 and M are grouped together according to the context of "Day", while Task 1, 3 and M are grouped together with respect to the context of "Location", etc. Intuitively, by grouping similar users/tasks together and learning their models jointly using a shared representation, the training data for each task is strengthened and the overall performance of multiple classification tasks can be improved.

Along this line, in this paper, we propose a *LAB*el-sensitive *TA*sk *G*rouping framework (LABTAG for short) by Bayesian nonparametric approach for multi-task multi-label classification. The reason for adopting the Bayesian nonparametric approach is that it can estimate the unknown number of feature-label patterns and the different unknown number of task clusters under different labels. To this end, we place one Dirichlet process (DP) prior on the distributions of features and labels to capture the feature-label patterns, and another DP prior on parameters of features to clustering similar tasks under different labels, respectively. Therefore, the proposed LABTAG model not only explores the label correlations but also takes the advantage of label-sensitive task grouping. For the model solution, we utilize the variational inference to estimate the parameters in the graphical model to form a learning algorithm. Finally, we conduct extensive experiments on three public data sets to demonstrate the effectiveness of the proposed model compared with various of baseline approaches.

2 Related Work

Multi-label Learning. A straightforward approach to multi-label learning is to transform the multi-label data set to one or multiple single label data sets, which is called the problem transformation methods [Zhang and Zhou, 2014]. Normally, label powerset method [Tsoumakas *et al.*, 2011] transforms the multi-label data set to a multi-class classification problem, in which each class is a unique set of labels that exist in a multi-label training set. Apparently, label powerset method could face the problem of exponential explosion when the label space is too large. In contrast, binary relevance [Boutell *et al.*, 2004] is to transform the multi-label data sets to k binary classification data sets, where k is the number of labels. However, it ignores the fact that some labels are more likely to co-exist in the instances. To parameterize the label co-occurrences, Ghamrawi et al. [Ghamrawi and McCallum, 2005] proposed a conditional random field based multi-label classification model. Zhang et al. [Zhang and Zhang, 2010] utilized a Bayesian network structure to encode the conditional dependencies of the labels and the features. Nguyen et al. [Nguyen *et al.*, 2016] proposed a Bayesian nonparametric approach to learn the number of label-feature correlation patterns automatically. However, the most previous multi-label classification models focus on single-task learning problems while ignoring modeling the similarity among tasks.

Multi-task Learning. The multi-task learning approach try to solve multiple learning tasks at the same time, while exploiting commonalities and differences across tasks. The main purpose of the multi-task learning is to capture the similarity information or shared structures among multiple tasks to enhance the learning performance [Zhou *et al.*, 2012]. Typical shared structures include fully connected structure (all tasks are related) [Lawrence and Platt, 2004; Evgeniou and Pontil, 2004], clustered structure [Bakker and Heskes, 2003], tree structure [Kim and Xing, 2010], network structure [Chen *et al.*, 2012], etc. Particularly, Xue et al. [Xue *et al.*, 2007] proposed a multi-task learning model with Dirichlet process prior to identify groups of related tasks automatically. Different from their works, in this paper, we adopt the multi-task learning to jointly learn several related multi-label classification tasks based on the Bayesian nonparametric approach in a generalized framework.

3 Multi-task Multi-label Classification Model

3.1 Problem Description

Given M tasks with multi-label classification problem, let $X = \{X_1, X_2, \cdots, X_M\}$ denote all domains of data, $Y = (0,1)^C$ be the labels, C be the number of labels, N_m be the number of instances in the m-th task, where $X_m = \{x_{mn}, y_{mn}\}|_{n=1}^{N_m}$ and $y_{mn} \in Y$, the goal of the multi-task multi-label learning is to explore the label correlation and task correlation to learn a function that maps x_{mn} to multi-label vector y_{mn}: $f(x_{mn}) \rightarrow y_{mn}$.

3.2 Model Overview

We build a multi-task multi-label model with Dirichlet process priors to jointly learn the multiple multi-label classification problems. The graphical model is shown in Fig. 2, which consists of two parts:

- Firstly, a Dirichlet process prior is placed on parameters ϕ and ψ. $\{\phi_{mk}, \psi_{mk}\}$ are feature-label pattern pairs, which is similar with literature [Nguyen *et al.*, 2016].

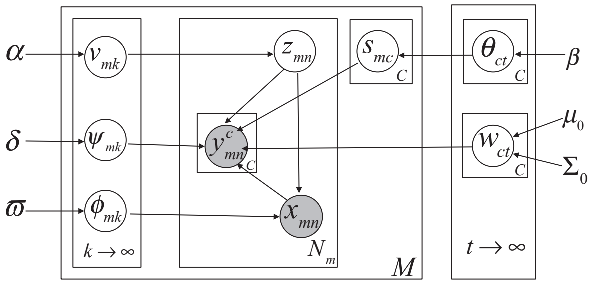


Figure 2: Graphical representation of the model.

For different tasks, different feature-label patterns are captured. The number of feature-label patterns can be determined automatically after training due to the non-parametric setting;

- Then we place another Dirichlet process prior on the parameters $\{w_{ct}\}$ to implement label-sensitive task grouping. Under different labels, the tasks can be clustered into different groups according to the between-task similarity. The variable $\{s_{mc}\}$ indicates which cluster task m belongs to under label c.

3.3 Generative Process

The generative process of the proposed multi-task multi-label model is shown as follows. Given the context recommendation as an example. For each task (e.g., user) m, we utilize the stick-breaking view [Ishwaran and James, 2001] of the Dirichlet process to capture the feature-label patterns firstly. Therefore, We sample v_{mk} from the Beta distribution $Be(1, \alpha)$ and calculate $\pi_{mk} = v_{mk} \prod_{i=1}^{k-1} (1 - v_{ik})$. Then we sample ψ_{mk}, ϕ_{mk} from the base Dirichlet distribution $Dir(\delta)$, $Dir(\varpi)$ respectively, $k = 1, \cdots, \infty$. In addition, the tasks are latently related label-sensitively, which means the parameters may be shared between different tasks under different labels. Therefore, we place another Dirichlet process prior on the parameter $\{w_{ct}\}$ for each label c. Similarly, we draw θ_{ct} from the Beta distribution $Beta(1, \beta)$, calculate $\pi^*_{ct} = \theta_{ct} \prod_{i=1}^{t-1} (1 - \theta_{ci})$, and then draw w_{ct} from Gaussian distribution $N(\mu_0, \Sigma_0)$, $t = 1, \cdots, \infty, c = 1, \cdots, C$. In the next step, we sample the indicator variable $s_{mc} \sim Mult(1; \pi^*_{c1}, \cdots \pi_{c\infty^*})$. Then we record the parameters $w_m = \{w_{c,s_{mc}}\}|_{c=1}^{C}$. Finally, for each instance n in task m, we draw the indicator variable $z_{mn} \sim Mult(1; \pi_{m1}, \cdots \pi_{m\infty})$. Next we generate the feature x_{mn} (e.g., the item content information): $x_{mn} \sim Mult(\phi_{z_{m,mn}})$, and the corresponding label y_{mn} (e.g., the contexts information) : $y_{mn} \sim Mult(\psi_{m,z_{mn}} .* \sigma(x_{mn}^T * w_m))$, where $\sigma(\cdot)$ is the sigmoid function. The details of the generative process can be found in Alg. 1.

3.4 Learning Algorithm

In the Bayesian approach, the calculation of the posterior distribution of the latent variables given the observed variable and the hyper-parameters is critical. However, the posterior distribution does not have an analytic form in most cases [Xue *et al.*, 2007]. Variational inference and Monte Carlo Markov chain (MCMC) sampling are two widely used methods for Bayesian inference. However, when faced with Dirichlet process prior, MCMC method is slow and difficult to converge while variational inference methods are deterministic compared with MCMC sampling methods [Ishwaran and James, 2001; Blei *et al.*, 2006].

Algorithm 1 The generative process of the proposed LABTAG model

```
for each task $m$ do
    Draw $v_{mk}$ independently from Beta distribution $Beta(1, \alpha), k = 1, \cdots, \infty, m = 1, \cdots, M$.
    $\pi_{mk} = v_{mk} \prod_{i=1}^{k-1} (1 - v_{ik}), k = 1, \cdots, \infty$.
    Draw $\psi_{mk}$ independently from Dirichlet distribution $Dir(\delta), k = 1, \cdots, \infty, m = 1, \cdots, M$.
    Draw $\phi_{mk}$ independently from Dirichlet distribution $Dir(\varpi), k = 1, \cdots, \infty, m = 1, \cdots, M$.
    for each label $c$ do
        Draw $\theta_{ct}$ independently from Beta distribution $Beta(1, \beta), t = 1, \cdots, \infty, c = 1, \cdots, C$.
        $\pi^*_{ct} = \theta_{ct} \prod_{i=1}^{t-1} (1 - \theta_{ci}), t = 1, \cdots, \infty$.
        Draw $w_{ct}$ independently from Gaussian distribution $N(\mu_0, \Sigma_0), t = 1, \cdots, \infty, c = 1, \cdots, C$.
        $s_{mc} \sim Mult(1; \pi^*_{c1}, \cdots \pi^*_{c\infty}), m = 1, \cdots, M$.
        $w_m = \{w_{c,s_{mc}}\}|_{c=1}^{C}$.
    end for
    for each instance $n$ in task $m$ do
        $z_{mn} \sim Mult(1; \pi_{m1}, \cdots \pi_{m\infty}), m = 1, \cdots, M, n = 1, \cdots, N_m$.
        $x_{mn} \sim Mult(\phi_{m,z_{mn}})$.
        $y_{mn} \sim Mult(\psi_{m,z_{mn}} .* \sigma(x_{mn}^T .* w_m))$.
    end for
end for
```

Generally, variational inference method is to approximate the posterior distribution p of interest using a variational distribution q [Jordan *et al.*, 1999; Ghahramani and Beal, 2001]. By minimizing the Kullback-Leibler (KL) divergence between p and q, the calculation of the posterior distribution can be transformed to an optimization problem. Particularlly, by assuming that the variational distribution could be factorized with different parts in the exponential family, the analytic form of the variational distribution q could be obtained. Under this assumption, the factorized variational distribututions of the LABTAG model are as follows:

$$
\begin{aligned}
q(v, \psi, \phi, z, \theta, w, s) = & \prod_{m=1}^{M} \prod_{k=1}^{K-1} q(v_{mk}) \prod_{m=1}^{M} \prod_{k=1}^{K} q(\psi_{mk}) \\
& \prod_{m=1}^{M} \prod_{k=1}^{K} q(\phi_{mk}) \prod_{m=1}^{M} \prod_{n=1}^{N_m} q(z_{mn}) \prod_{c=1}^{C} \prod_{t=1}^{T^*} q(\theta_{ct}) \\
& \prod_{t=1}^{T^*} \prod_{c=1}^{C} q(w_{ct}) \prod_{m=1}^{M} \prod_{c=1}^{C} q(s_{mc})
\end{aligned} \tag{1}
$$

where $q(v_{mk})$ are Beta distributions with parameter (r^m_{1k}, r^m_{2k}); $q(\psi_{mk})$ are dirichlet distributions with parameter $\tilde{\psi}_{mk}$; $q(\phi_{mk})$ are dirichlet distributions with parameter $\tilde{\phi}_{mk}$;

$q(z_{mn})$ are multinomial distributions with parameter $\tilde{z}_{mn}$; $q(\theta_{ct})$ are Beta distributions with parameter $(\tau^c_{1t}, \tau^c_{2t})$; $q(w_{ct})$ are normal distributions with parameters $(\tilde{\mu}_{ct}, \tilde{\Sigma}_{ct})$; $q(s_{mc})$ are multinomial distributions with parameter $\tilde{s}_{mc}$. It is worth noting that we utilize truncated stick-breaking representations here [Blei *et al.*, 2006]. In the proposed model, K and T^* are all truncation levels which could be set freely. Next, the estimations of the parameters of the variational distribution will be introduced in detail.

Estimating the parameters of $q(v_{mk})$:
From the Bayesian nonparametric setting, $q(v_{mk})$ is the same distribution type as $P(v_{mk}|z, \alpha)$, in which the difference is the parameter. According to Bayes' theorem [Rosen, 2007], the posterior distribution can be written in exponential family as follows:

$$\begin{aligned} P(v_{mk}|z,\alpha) &\propto \prod_{n=1}^{N_m} P(z_{mn}|v_{mk})P(v_{mk}|\alpha) \\ &\propto \exp\{\sum_{n=1}^{N_m} (\mathrm{I}[z_{mn} > k] + \alpha - 1)\log(1 - v_{mk}) \\ &\qquad + \mathrm{I}(z_{mn} = k)\log v_{mk}\} \end{aligned} \tag{2}$$

According to the principle of the variational inference [Blei *et al.*, 2006], the parameters of the variational distribution $q(v_{mk})$ equals to the expectation of the natrual parameter of the posterior distribution under the variational distribution and can be calculated as following equations:

$$\begin{aligned} r^m_{1k} &= 1 + \sum_{n=1}^{N_m} E_q I(z_{mn} = k) = 1 + \sum_{n=1}^{N_m} \tilde{z}^k_{mn}, \\ r^m_{2k} &= \alpha + \sum_{n=1}^{N_m} \sum_{j=k+1}^{K} E_q I(z_{mn} > k) \\ &= \alpha + \sum_{n=1}^{N_m} \sum_{j=k+1}^{K} \tilde{z}^j_{mn} \end{aligned} \tag{3}$$

Estimating the parameters of $q(\phi_{mk})$:
Again, the posterior distribution $P(\phi_{mk}|z, X, \varpi)$ is a Dirichlet distribution which is also in the exponential family. Hence, the parameters $\tilde{\phi}_{mk}$ of variational distribution $q(\phi_{mk})$ can be obtained as follows:

$$\begin{aligned} P(\phi_{mk}|z,X,\varpi) &\propto \exp\{(\varpi - 1 + \\ &\qquad \sum_{n=1}^{N_m} x_{mn}\mathrm{I}[z_{mn} = k])\log(\phi_{mk})\} \\ \tilde{\phi}_{mk} = \varpi + \sum_{n=1}^{N_m} x_{mn} E_q I(z_{mn} = k) &= \varpi + \sum_{n=1}^{N_m} x_{mn}\, \tilde{z}^k_{mn} \end{aligned} \tag{4}$$

Estimating the parameters of $q(\psi_{mk})$:
As disscussed before, in different tasks, different feature-label patterns can be captured by (ϕ_{mk}, ψ_{mk}). Similar with the estimation of $\tilde{\phi}_{mk}$, ψ_{mk} also has a Dirichlet process prior, therefore, $\tilde{\psi}_k$ can be estimated similarly:

$$\tilde{\psi}_{mk} = \delta + \sum_{n=1}^{N_m} y_{mn}\, \tilde{z}^k_{mn} \tag{5}$$

Estimating the parameters of $q(z_{mn})$:
The posterior of z_{mn} is a multinomial distribution as follows: $P(z_{mn} = i|\cdot) \propto \exp\{\log v_{mi} + \sum_{j=1}^{i-1} \log(1 - v_{mj}) + y_{mn}\log\psi_{mi} + x_{mn}\log\phi_{mi}\}$.

Hence $\tilde{z}^i_{mn}$ equals to the expectation of the natural parameter of the posterior distribution under the variational distribution q:

$$\begin{aligned} \tilde{z}^i_{mn} \propto \exp\{E_q \log v_{mi} &+ \sum_{j=1}^{i-1} E_q \log(1 - v_{mj}) \\ &+ y_{mn}E_q\log\psi_{mi} + x_{mn}E_q\log\phi_{mi}\} \end{aligned} \tag{6}$$

where,

$$\begin{aligned} E_q \log v_{mi} &= \Psi(r^m_{1i}) - \Psi(r^m_{1i} + r^m_{2i}), \\ E_q \log(1 - v_{mj}) &= \Psi(r^m_{2j}) - \Psi(r^m_{1j} + r^m_{2j}), \\ y_{mn}E_q \log\psi_{mi} &= \sum_{c=1}^{C} y^c_{mn}\Psi(\tilde{\psi}^c_{mi}) - \Psi(\sum_v \tilde{\psi}^c_{vi}), \\ x_{mn}E_q \log\phi_{mi} &= \sum_{d=1}^{D} x^d_{mn}\Psi(\tilde{\phi}^d_{mi}) - \Psi(\sum_v \tilde{\phi}^d_{vi}), \end{aligned} \tag{7}$$

in which $\Psi(\cdot)$ is the digmma function, which is the first derivative of the log Gamma function. C is the number of the labels and D is the number of the features.

Estimating the parameters of $q(\theta_{ct})$:
Similar with estimating r, τ is calculated as follows:

$$\tau^c_{1k} = 1 + \sum_{m=1}^{M} \tilde{s}^t_{mc}, \quad \tau^c_{2k} = \beta + \sum_{m=1}^{M}\sum_{j>t}^{T} \tilde{s}^j_{mc} \tag{8}$$

Estimating the parameters of $q(w_{ct})$:
Since the sigmoid function is not in the exponential family, it is difficult to transform the the posterior distribution of $P(w_{tc}|\cdot)$ to a exponential expression. Hence, we utilize a variational approximation to calculate the posterior distribution [Jaakkola and Jordan, 1997; Xue *et al.*, 2007], which is shown as follows.

$$\begin{aligned} \tilde{\mu}_{ct} &= \tilde{\Sigma}_{ct}[{\Sigma_0}^{-1}\mu_0 + \sum_{m=1}^{M} \tilde{s}^t_{mc} \sum_{n=1}^{N_m} (y^c_{mn} - \frac{1}{2})x_{mn}] \\ \tilde{\Sigma}_{ct} &= [{\Sigma_0}^{-1} + 2\sum_{m=1}^{M} \tilde{s}^t_{mc} \sum_{n=1}^{N_m} |\rho(\xi^c_{mn})| x_{mn}x^T_{mn}]^{-1} \end{aligned} \tag{9}$$

where $\xi^c_{mn} = \sqrt{\sum_{t=1}^{T} \tilde{s}^t_{mc}\, x^T_{mn}(\tilde{\mu}_{ct}\, \mu^T_{ct} + \tilde{\Sigma}_{ct})x_{mn}}$

Estimating the parameters of $q(s_{mc})$:
s_{mc} is another indicator variable which reveals which cluster each task belongs under different labels. Similar with estimating $\tilde{z}^i_{mn}$, $\tilde{s}^i_{mc}$ can be estimated as follows:

$$\begin{aligned} \tilde{s}^i_{mc} \propto \exp\{E_q \log\theta_{ci} &+ \sum_{j=1}^{i-1} E_q(1 - \log\theta_{cj}) \\ &+ \sum_{n=1}^{N_m} [\rho(\xi^c_{mn})x^T_{mn}(\tilde{\mu}_{ci}\, \tilde{\mu}_{ci}{}^T + \tilde{\Sigma}_{ci})x_{mn} \\ &\quad + (y^c_{mn} - \tfrac{1}{2})\tilde{\mu}_{ci}{}^T x_{mn} + \log(\sigma(\xi^c_{mn})) \\ &\quad - \tfrac{1}{2}\xi^c_{mn} - \rho(\xi^c_{mn}){\xi^c_{mn}}^2]\} \end{aligned} \tag{10}$$

3.5 Prediction

Given the learned parameters $\Theta = \{z, s, v, \phi, \psi, \theta, w\}$ and a new test sample x_{m,n^*}, we need to predict the corresponding label vector $y_{m,n^*} \in (0,1)^C$. We assume y_{m,n^*}

Data set	Metrics	LABTAG	BR	CDN	BNMC	ML-kNN
LDOS-CoMoDa	Accuracy (%)	**42.19**	39.59	33.71	37.00	41.47
	F1 score (%)	**57.47**	52.83	47.19	50.88	55.23
	Hamming loss	0.2457	0.2698	0.2431	**0.1798**	0.2034
	One-error	**0.0063**	0.4416	0.4789	0.0406	**0.0063**
	Rank loss	**0.4513**	0.5119	0.5745	0.6076	0.5210
Enron Email Corpus	Accuracy (%)	**69.06**	66.20	54.65	41.46	59.45
	F1 score (%)	**72.22**	70.63	56.86	44.40	64.55
	Hamming loss	**0.0135**	0.0226	0.0192	0.0294	0.0148
	One-error	**0.1798**	0.2350	0.3702	0.4859	0.2208
	Rank loss	**0.2608**	0.2850	0.4059	0.5459	0.3782
TripAdvisor	Accuracy (%)	**67.45**	64.21	61.17	64.42	64.52
	F1 score (%)	**68.61**	65.12	62.21	64.42	65.34
	Hamming loss	**0.1313**	0.1453	0.1335	0.1423	0.1405
	One-error	**0.2906**	0.3361	0.3216	0.3558	0.3339
	Rank loss	**0.3081**	0.3497	0.3434	0.3558	0.3480

Table 1: Mean value of evaluation metrics on all tasks

is drawn from a multinomial distribution with parameters $[\varsigma^1_{m,n^*},\dots,\varsigma^C_{m,n^*}]$, in which ς^c_{m,n^*} can be calculated as: $\varsigma^c_{m,n^*} \propto P(y^c_{m,n^*} = 1|x_{m,n^*},\widetilde{\mu},\widetilde{\Sigma},\widetilde{s}) \times P(y^c_{m,n^*} = 1|\widetilde{z},\widetilde{\phi},\widetilde{\psi})$. Particularly, since $P(y^c_{m,n^*} = 1|x_{m,n^*},\widetilde{\mu},\widetilde{\Sigma},\widetilde{s})$ does not have a accurate analytic form, we utilize the approximation form as follows: $\sum_{t=1}^{T^*} \widetilde{s}^t_{mc}\,\sigma(\frac{\widetilde{\mu}_{ct}\,x_{m,n^*}}{\sqrt{1+\frac{PI}{8}x^T_{m,n^*}\,\widetilde{\Sigma}_{ct}\,x_{m,n^*}}})$ according to literature [Xue *et al.*, 2007].

4 Experimental Evaluation

4.1 Data Preparation

We use three public data sets for the performance evaluation, including TripAdvisor, LDOS-CoMoDa and Enron Corpus data sets, to evaluate the performance of all compared algorithms. TripAdvisor and LDOS-CoMoDa are two context-aware data sets, which are used for context recommendation systems [Zheng *et al.*, 2014]. TripAdvisor is a hotel rating data set, in which the content information includes user country, rating, etc. and the context is the trip type. LDOS-CoMoDa [Adomavicius and Tuzhilin, 2015] is a movie rating data set, which utilizes the movie contents including movie year, genre, actor, etc. to recommend the contexts consisting of time, location, day and companion. Finally, the Enron Email Corpus contains email information (email content and recipients) from Enron [Klimt and Yang, 2004; Carvalho and Cohen, 2007]. The statistics of three data sets are shown in Table 2. These three data sets contain different number of tasks ranging from 10 to 46, different numbers of labels and features, which are able to prove the robustness of the proposed model.

4.2 Baselines and Evaluation Metrics

Baselines. We compare our model LABTAG with the following baselines,

- Binary Relevance (BR) transforms the multi-label data set to k single label data sets, then adopts some basic algorithms on each data set, such as Decision Tree, SVM, Logistic Regression, etc., finally combines the results to form the prediction. In our experiment setting, the Decision Tree is used as the basic algorithm.

Statistics	LDOS	Enron Corpus	TripAdvisor
# of tasks	10	18	46
# of labels	17	58	5
# of features	146	93	460
Ave. # of ins.	22	10	28

Table 2: Description of datasets.

- ML-kNN [Zhang and Zhou, 2007] is from the traditional KNN algorithm, which utilizes k nearest neighbors in the training data to predict the instances in the test data.
- Conditional Dependency Networks (CDN) [Guo and Gu, 2011] is a cyclic directed graphical model by considering the label dependency, in which Gibbs sampling is utilized for inference.
- Bayesian Nonparametric Multi-label Classification (BNMC) [Nguyen *et al.*, 2016] is a Bayesian nonparametric framework which can learn the unknown number of label correlations automatically and handle the missing label samples naturally.

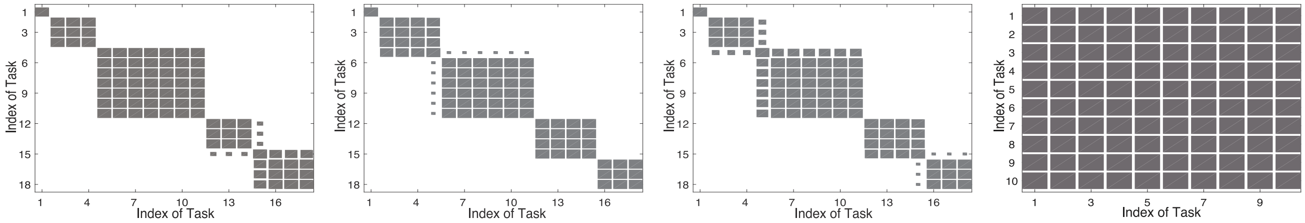
We utilize MEKA toolbox [Read *et al.*, 2016] for the implementation of BR and CDN, and download the Matlab code from the author's website for BNMC and ML-KNN.

Evaluation Metrics. We adopt five widely used metrics for performance evaluation [Zhang and Zhou, 2014], including Accuracy, F1-score, Hamming loss, One-error and Rank loss. For Accuracy and F1-score, larger value indicates the better performance, while for Hamming loss, One-error and Rank loss, smaller value indicates the better performance.

The hyper-parameter settings of LABTAG model are as follows: $\alpha = 1$, $\delta = 0.01$, $\varpi = 0.07$, $\beta = 1$, $\mu_0 = \mathbf{0}$, $\Sigma_0 = 10\text{I}$. The truncation threshold is set as $0.001 \times \#Train$ and the learning rate is set as 0.01. In each task, 50% data are used for training and the remaining 50% for test.

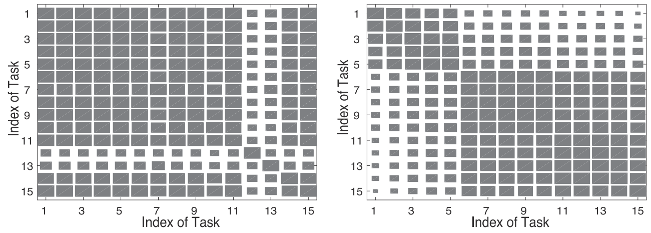
4.3 Numerical Results

The comparison of the classification performance between the proposed model and the baseline methods is shown in Table 1, and the best results w.r.t each evaluation metric are marked in bold. Besides, Fig. 5 shows the mean Accuracy and



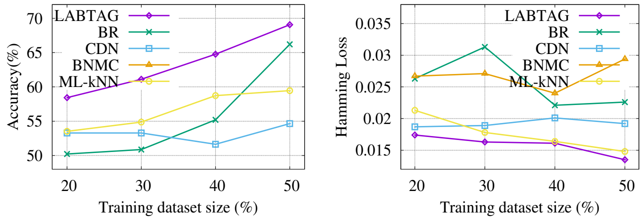
(a) Label 6 in Enron dataset.
(b) Label 46 in Enron dataset.
(c) Label 47 in Enron dataset.
(d) Label 1 in LDOS dataset.

Figure 3: Hinton diagram for the between-task similarity under different labels in Enron and LDOS dataset.



(a) Label 2.
(b) Label 3.

Figure 4: Hinton diagram for the between-task similarity in TripAdvisor dataset.



(a) Mean Accuracy.
(b) Mean Hamming Loss.

Figure 5: Mean Accuracy and Hamming Loss under different training dataset sizes in Enron dataset.

Hamming Loss under different training dataset sizes. From the results in Table 1 and Fig. 5, we have the following insightful observations,

- Our model LABTAG can achieve the best performance compared with all baselines on three data sets in term of all five metrics, except that on LDOS-CoMoDa data set, BNMC obtains a better result in term of Hamming loss and ML-kNN obtains a better result in term of One-error.
- LABTAG outperforms BR, CDN, BNMC, and ML-kNN, which shows the importance of applying multi-task learning to handling multi-label classification problem. The reason is that the small number of instances in each task can lead to under-fitting if each task is trained separately, while LABTAG can learn multiple tasks jointly to enhance performance. Similar tasks can share the same model parameters under different labels, in which one task could benefit the knowledge from other similar tasks.
- The success of LABTAG also attribute to the adoption of Bayesian nonparametric approach, based on which we do not need to specify the number of clusters.
- The proposed LABTAG model can also perform better than all baselines under different sizes of training data, even when the size of training data is small.

4.4 Model Analysis

To explain the cluster structures of multiple tasks under different labels in each data set, we calculate the between-task similarity as follows: (1) We obtain the results of 10 random trials; (2) In each random run, we output $\{\overset{\sim}{s^t_{mc}}\}_{t=1}^{T^*}$ which indicates the probability that task m belongs to cluster t under label c. Hence we utilize $\arg\max_t \overset{\sim}{s^t_{mc}}$ as the cluster index of task m under label c; (3) We construct a task similarity matrix, in which the element (i, j) records the number of occurrences that task i and task j are grouped into the same cluster among the total 10 random runs. The Hinton diagram for the between-task similarity matrices of different data sets is shown in Fig. 3 and Fig. 4, in which the larger size of blocks brings to higher between-task similarity. Specifically, we introduce three different cluster patterns of tasks under different labels (label 6, 46, 47) from the Enron data set as displayed in Figs. 3(a), 3(b), 3(c). Task 5 shares the same parameters with tasks 6∼11 under label 6 and label 47, while task 5 has high similarity with tasks 2∼4 under label 46. The same with task 15, task 15 is clustered differently under label 6 and label 46. We can observe the similar results from the TripAdvisor data set as shown in Figs. 4. Tasks 1∼11 and tasks 14∼15 form a large cluster together under label 2 while tasks 1∼5 and tasks 6∼15 form two clusters respectively under label 3. This reveals the motivation of the proposed LABTAG model that one task could belong to different clusters with other tasks under different labels. However, on LDOS-CoMoDa data set shown in Fig. 3(d), all tasks are grouped together and share the same parameters. This is the reason why LABTAG performs relatively worse performance on this data set.

5 Conclusion

In this paper, we addressed the problem of multi-task multi-label learning, which is particular suitable to deal with the applications of multiple highly correlated tasks with sparse training instances. Specifically, we proposed a label-sensitive task grouping framework (LABTAG) by Bayesian nonparametric approach for multi-task multi-label classification. In this framework, LABTAG took advantages of both the label correlations and similar task grouping under different labels to enhance the performance of classification. We evaluated the performance of the proposed model on three public multi-label data sets, which validated the superiority of the proposed model over the state-of-the-arts.

References

[Adomavicius and Tuzhilin, 2015] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 191–226. Springer, 2015.

[Bakker and Heskes, 2003] Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *JMLR*, 4(May):83–99, 2003.

[Blei *et al.*, 2006] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

[Boutell *et al.*, 2004] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[Carvalho and Cohen, 2007] Vitor R Carvalho and William Cohen. Recommending recipients in the enron email corpus. *Machine Learning*, 2007.

[Chen *et al.*, 2012] Xi Chen, Qihang Lin, Seyoung Kim, Jaime G Carbonell, and Eric P Xing. Smoothing proximal gradient method for general structured sparse regression. *AOAS*, pages 719–752, 2012.

[Dembczynski *et al.*, 2010] Krzysztof Dembczynski, Weiwei Cheng, and Eyke Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *ICML*, volume 10, pages 279–286, 2010.

[Evgeniou and Pontil, 2004] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *SIGKDD*, pages 109–117. ACM, 2004.

[Ghahramani and Beal, 2001] Zoubin Ghahramani and Matthew J Beal. Propagation algorithms for variational bayesian learning. pages 507–513, 2001.

[Ghamrawi and McCallum, 2005] Nadia Ghamrawi and Andrew McCallum. Collective multi-label classification. In *CIKM*, pages 195–200. ACM, 2005.

[Gong *et al.*, 2013] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv*, 2013.

[Guo and Gu, 2011] Yuhong Guo and Suicheng Gu. Multi-label classification using conditional dependency networks. In *IJCAI*, volume 22, page 1300, 2011.

[Ishwaran and James, 2001] Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *JASA*, 96(453):161–173, 2001.

[Jaakkola and Jordan, 1997] T Jaakkola and M Jordan. A variational approach to bayesian logistic regression models and their extensions. In *AISTATS Workshop*, volume 82, page 4, 1997.

[Jordan *et al.*, 1999] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):105–161, 1999.

[Kim and Xing, 2010] Seyoung Kim and Eric P Xing. Tree-guided group lasso for multi-task regression with structured sparsity. 2010.

[Klimt and Yang, 2004] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *ECML*, pages 217–226. Springer, 2004.

[Lawrence and Platt, 2004] Neil D Lawrence and John C Platt. Learning to learn with the informative vector machine. In *ICML*, page 65. ACM, 2004.

[Li *et al.*, 2012] Ying-Xin Li, Shuiwang Ji, Sudhir Kumar, Jieping Ye, and Zhi-Hua Zhou. Drosophila gene expression pattern annotation through multi-instance multi-label learning. *TCBB*, 9(1):98–112, 2012.

[Liu and Chen, 2015] Shuhua Monica Liu and Jiun-Hung Chen. A multi-label classification based approach for sentiment classification. *ESWA*, 42(3):1083–1093, 2015.

[Nguyen *et al.*, 2016] Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. A bayesian nonparametric approach for multi-label classification. In *ACML*, pages 254–269, 2016.

[Read *et al.*, 2011] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359, 2011.

[Read *et al.*, 2016] Jesse Read, Peter Reutemann, Bernhard Pfahringer, and Geoff Holmes. MEKA: A multi-label/multi-target extension to Weka. *JMLR*, 17(21):1–5, 2016.

[Rosen, 2007] Kenneth H Rosen. Discrete mathematics and its applications. *AMC*, 10:12, 2007.

[Tsoumakas *et al.*, 2011] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *TKDE*, 23(7):1079–1089, 2011.

[Xue *et al.*, 2007] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-task learning for classification with dirichlet process priors. *JMLR*, 8(Jan):35–63, 2007.

[Zhang and Zhang, 2010] Min-Ling Zhang and Kun Zhang. Multi-label learning by exploiting label dependency. In *SIGKDD*, pages 999–1008. ACM, 2010.

[Zhang and Zhou, 2007] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[Zhang and Zhou, 2014] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *TKDE*, 26(8):1819–1837, 2014.

[Zheng *et al.*, 2014] Yong Zheng, Bamshad Mobasher, and Robin Burke. Context recommendation using multi-label classification. In *WI and IAT*, volume 2, pages 288–295. IEEE, 2014.

[Zhou *et al.*, 2012] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Multi-task learning: Theory, algorithms, and applications. In *https://www. siam. org/meetings/sdm12/zhou_chen_ye. pdf*, 2012.