研究与开发

一种适应于非完备标签数据和标签关联性的多标签分类方法

张丽娜[1,2],戴灵鹏[2],匡泰[1]

(1. 浙江安防职业技术学院信息工程系,浙江 温州 325016;
2. 温州大学生命与环境科学学院,浙江 温州 325035)

摘　要:多标签分类已在很多领域得到了实际应用,所用标签大多具有很强的关联性,甚至存在非完备标签或部分标签遗失。然而,现有的多标签分类算法难以同时处理这两种情况。基于此,提出一种新的概率模型处理方法,实现同时对具有标签关联性和遗失标签情况进行多标签分类。该方法可以自动获知和掌握多标签的关联性。此外,通过整合遗失的标签信息,该方法能够提供一个自适应策略来处理遗失的标签。在完备标签和非完备标签的数据上进行实验,结果表明,与现有的多标签分类算法相比,提出的方法得到了较好的分类预测评价值。

关键词:非完备标签;标签关联性;多标签分类;概率模型

中图分类号:TP311　　**文献标识码**:A

doi: 10.11959/j.issn.1000-0801.2016197

A multi-label classification method for disposing incomplete labeled data and label relevance

ZHANG Lina[1,2], DAI Lingpeng[2], KUANG Tai[1]

1. Department of Information Engineering, Zhejiang College of Security Technology, Wenzhou 325016, China
2. College of Life and Environmental Science, Wenzhou University, Wenzhou 325035, China

Abstract: Multi-label classification methods have been applied in many real-world fields, in which the labels may have strong relevance and some of them even are incomplete or missing. However, existing multi-label classification algorithms are unable to handle both issues simultaneously. A new probabilistic model that can automatically learn and exploit multi-label relevance was proposed on label relevance and missing label classification simultaneously. By integrating out the missing information, it also provides a disciplined approach to handle missing labels. Experiments on a number of real world data sets with both complete and incomplete labels demonstrated that the proposed method can achieve higher classification and prediction evaluation scores than the existing multi-label classification algorithms.

Key words: incomplete label, label relevance, multi-label classification, probabilistic model

收稿日期:2016-05-04;**修回日期**:2016-07-10

通信作者:张丽娜,zln_zcst@163.com

基金项目:浙江省教育科学规划基金资助项目(No.2016SCG188);浙江省自然科学基金资助项目(No.LY14C03007)

Foundation Items: Education Science Department Foundation of Zhejiang Province(No.2016SCG188), The Natural Science Foundation of Zhejiang Province of China (No.LY14C03007)

1 引言

在传统的标签数据分类问题中，一个样点通常只分配一个标签。但在实际应用中，大部分都涉及多标签分类，即一个样点可分配多个标签，如文本分类、图像标注以及基因方程的分析[1]等。正是由于该类问题的不断出现，也吸引了越来越多的学者进行多标签分类问题的研究。

经过多年的研究和发展，目前已出现了大量的多标签分类方法，最为经典且应用最广泛的是问题转换类方法[2,3]。在此类方法中，较常用的一种方法是二值相关（binary relevance，BR）[4]算法。该方法建立在标签之间相互独立假设的基础上，当标签存在很大的关联性时，该方法的效果不佳。针对 BR 方法的局限性，一种解决方案是假定标签的关联性是一种先验信息或很容易被估计，如 Read 等人[5]提出了一种链式分类器（classifier chain，CC）算法，该方法以一定的次序将一系列 BR 分类器串联起来，后一个分类器的结果总是依赖于前一个分类器，这样便考虑了标签的关联性。然而，其弊端是如果前一个分类器存在误差，这种关联也会将误差进行传递积累。孙霞等人[6]提出了一种基于 Haddoop 框架的传播算法，该方法能适应大规模数据，但误差同样会在迭代算法中传播。在 Hariharan 等人[7]提出的最大边界多标签分类器（max-margin multi-label classifier，M3L）中，标签的关联性被从训练集中计算出的成对标签关联所代替，取得了不错的效果，但该方法在样本数量较少时会出现一些误差。此外，一些方法通过标签转换进行去标签关联，即对去关联的标签进行分离学习，如 Hsu 等人[8]提出的利用随机矩阵进行标签转换、在标签矩阵中采用奇异值分解法的主标签空间转换（principal label space transformation，PLST）[9]、在标签矩阵和输入样本矩阵中均采用奇异值分解法的条件主标签空间转换（conditional principal label space transformation，CPLST）[10]以及在最小二乘回归基础上利用 SVD 技术的多标签分类算法[11]，经过这些方法转换后的标签都可以单独进行处理。再有，一些学者在方法的精度等方面也做了一些研究，如李远航等人[12]、许美香等人[13]通过主动学习的方法提高分类器的精度和效率；徐晓丹等人[14]通过对数据进行预处理来提高分类器的精度。

在多标签分类算法中，另一个重要的问题是标签遗失。针对此问题，一些学者也提出了相应的解决方案。其中一种解决方法是丢弃所有没有标签的样本，但会损失大量的标签信息。另一种方法是估算遗失的标签，尤其是正标签，如群索引多标签排序（multi-label ranking with group lasso，MLRGL）[15]算法，它将分类问题转化为双边排序问题进行解决；快速图像标记[16]（fast image tagging，FastTag）法，它假设标签一致性被破坏，该方法能对可能遗失的正标签进行恢复，但不能有效地处理标签关联性问题，且只能对正标签进行恢复。Yu 等人[17]根据经验误差最小化，提出了一种对遗失标签的分析方法，该方法也没有准确考虑标签的关联性，限制了它的应用。

针对目前大部分方法分别在处理标签关联和标签遗失问题中存在的缺陷以及难以对这两个问题同时进行处理的情况，提出一个新的概率模型，该模型能同时处理标签关联和标签遗失问题，它不需准确地找出标签转换，只需将概率模型重新表述成原始的标签空间，在原始标签空间中自动获知和掌握标签的关联性。此外，通过对遗失的标签信息进行整合，以提供一种自适应的方法来处理遗失标签。最后，在具有完备标签和遗失标签数据上进行试验，结果表明，该方法所获得的效果优于现有经典方法。

2 提出的方法

第 2 节主要介绍同时处理非完备标签数据和标签关联性的多标签分类方法。首先进行最基本的步骤，即标签转换，在此基础上，提出一个新的概率模型，对标签关联和遗失标签同时进行处理，并给出优化子问题的推导过程。

2.1 标签转换

在多标签分类问题中，对于一个给定的训练集$\{(x,y)\}$，其中，$x\in R^d$是输入，$y\in\{0,1\}^m$是对应的输出，其中，m是样点数，d是样点维度。标签转换[9,10]是指将每一个y转换成$\tilde{y}=Py$，其中，$P\in\mathbb{R}^{\tilde{m}\times m}$是转换矩阵，经转换之后的标签$\{\tilde{y}_i\}_{i=1}^{\tilde{m}}$是不相关的，即可以单独对待。首先，采用一个线性模型作为初始模型（加权为$\tilde{w}_i$，偏差为$\tilde{b}_i$），即$\tilde{y}_i$被假定为：

$$\tilde{y}_i \mid x,\tilde{w}_i,\tilde{b}_i,\tilde{\sigma}_i \sim \mathcal{N}(\tilde{w}_i^{\mathrm{T}}x+\tilde{b}_i,\tilde{\sigma}_i^2) \tag{1}$$

其中，$i=1,\cdots,\tilde{m}$，$\tilde{\sigma}_i^2$是噪声方差。注意：用于标签转换的高斯噪声一般是实值（尽管原始值是二值）。式（1）可改写为如下形式：

$$\tilde{y} \mid x,\tilde{W},\tilde{b},\tilde{\Omega} \sim \mathcal{N}(\tilde{W}^{\mathrm{T}}x+\tilde{b},\tilde{\Omega}) \tag{2}$$

其中，$\tilde{W}=[\tilde{w}_1,\cdots,\tilde{w}_{\tilde{m}}]$，$\tilde{b}=[\tilde{b}_1,\cdots,\tilde{b}_{\tilde{m}}]^{\mathrm{T}}$，$\tilde{\Omega}=\mathrm{diag}(\tilde{\sigma}_1^2,\cdots,\tilde{\sigma}_{\tilde{m}}^2)$。在$\tilde{w}_i$中加入一个高斯先验信息，即$l_2$为正则化矩阵：

$$\tilde{w}_i \mid \tilde{\Sigma}_i \sim \mathcal{N}(0,\tilde{\Sigma}_i) \tag{3}$$

其中，$\tilde{\Sigma}_i = \mathrm{diag}\left\{\frac{1}{\alpha_{i,1}},\cdots,\frac{1}{\alpha_{i,d}}\right\}$。由于此$\tilde{m}$个标签是独立的，因此，$\{\alpha_{i,d}\}_{i=1}^{\tilde{m}}$值不同。根据学习值$\tilde{y}$可以反推到原始的标签空间，得到：

$$z = P^{\dagger}\tilde{y} \tag{4}$$

其中，$P^{\dagger}$为转换矩阵P的逆矩阵。目前，标签转换类方法主要的区别在于转换矩阵P如何获得。例如，在CPLST方法[10]中，矩阵P是正交的，主要是通过同时最小化转换空间中的训练误差$\left\|PY-\tilde{W}^{\mathrm{T}}X\right\|_{\mathrm{F}}^2$和转换标签时编码原始标签的误差$\left\|Y-P^{\dagger}PY\right\|_{\mathrm{F}}^2$得到，其中，$\|\cdot\|_{\mathrm{F}}$指F范数。

2.2 标签关联处理

第2.1节主要介绍了标签转换的基本方法，接下来将建立一个模型，利用该模型对标签关联性进行处理。

2.2.1 模型建立及分析

首先，建立一个概率模型。通过使用式(2)和式(4)以及$W=\tilde{W}P^{\dagger\mathrm{T}}$、$b=P^{\dagger}\tilde{b}$和$\Omega=P^{\dagger}\tilde{\Omega}P^{\dagger\mathrm{T}}$，可以得到，在对式(4)中$z$进行凑整之前的多标签预测遵从下列正态分布：

$$z\mid x,W,b,\Omega\sim\mathcal{N}(W^{\mathrm{T}}x+b,\Omega) \tag{5}$$

其中，Ω为m个噪声方差为$\tilde{\sigma}_i^2$的样点组成的对角矩阵$\tilde{\Omega}$与转换矩阵P的逆矩阵的作用关系，即代表关联性。尽管假设$\tilde{y}$是独立的，但由于式(4)中存在共享矩阵$P^{\dagger}$，因此，$\tilde{z}_i$是高度关联的，且Ω不是对角阵。此外，由于Ω直接对标签相互关系进行编码，所以，很容易获得先验信息。例如，可以采用：

$$p(\Omega)\propto\exp\left\{-\frac{1}{\lambda_1}\left\|\Omega^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2-\frac{1}{\lambda_2}\left\|\Omega^{-1}\right\|_{1}\right\} \tag{6}$$

其中，$\lambda_1,\lambda_2>0$是两个自由参数，用来增加Ω^{-1}的稀疏度和收缩性，Ω^{-1}指精度矩阵，Ω_{ij}^{-1}表示标签i和标签j之间部分相关[18]。增加Ω^{-1}的稀疏度意味着大部分的标签是条件不相关的。这里，$p(\Omega)$是根据Ω^{-1}表述，而不是Ω。为了模拟从z到二值预测y的凑整误差，通过下面的正态分布来近似：

$$y\mid z\sim\mathcal{N}\left\{z,\frac{1}{\lambda_0}I\right\} \tag{7}$$

其中，$\lambda_0>0$。

最终，W的分布可以从$\tilde{W}$中按如下步骤得出。首先，对W^{T}进行矢量化：

$$\begin{aligned}\mathrm{vec}(W^{\mathrm{T}})&=\left[W_{(1,:)},\cdots,W_{(d,:)}\right]^{\mathrm{T}}\\&=\left[\tilde{W}_{(1,:)}P^{\dagger\mathrm{T}},\cdots,\tilde{W}_{(d,:)}P^{\dagger\mathrm{T}}\right]^{\mathrm{T}}\\&=Q\left[\tilde{W}_{(1,:)},\cdots,\tilde{W}_{(d,:)}\right]^{\mathrm{T}}\end{aligned} \tag{8}$$

其中，$Q=\begin{bmatrix}P^{\dagger}&&0\\&\ddots&\\0&&P^{\dagger}\end{bmatrix}$。利用式(3)，可以得到：

$$\mathrm{vec}(W^{\mathrm{T}})\sim\mathcal{N}\left(0,Q\begin{bmatrix}\mathrm{diag}(\alpha_1)&&0\\&\ddots&\\0&&\mathrm{diag}(\alpha_d)\end{bmatrix}Q^{\mathrm{T}}\right) \tag{9}$$

其中，$\alpha_j=\left[\alpha_{1,j},\cdots,\alpha_{m,j}\right]^{\mathrm{T}}$；$W_{(j,:)}$相互独立，每一个$W_{(j,:)}$的分布为：

$$W_{(j,:)}\mid\Sigma_j\sim\mathcal{N}(0,\Sigma_j),\ \ j=1,2,\cdots,d \tag{10}$$

其中，$P^{\dagger}\mathrm{diag}(\alpha_j)P^{\dagger\mathrm{T}}$不是对角阵，指对特征$j$的任务关联。与式(6)类似，对$\Sigma_j$加入先验信息得到：

$$p(\Sigma_j)\propto\exp\left\{-\frac{1}{\beta_1}\left\|\Sigma_j^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2-\frac{1}{\beta_2}\left\|\Sigma_j^{-1}\right\|_{1}\right\} \tag{11}$$

其中，$\beta_1,\beta_2>0$。图1是代表整个模型的一个图解表示，可清楚呈现其建立过程。具体地，首先通过自由参数λ_1和λ_2根据式(6)来获得Ω的信息；同时，通过参数β_1和β_2根据式(11)来获得Σ_j的信息，进而得到$W_{(j,:)}$的分布。根据式(5)，通过已知的训练集x、偏差b以及得到的W和Ω来得到z，进而根据式(7)通过给定的参数λ_0来获得y。提出的模型相较传统方法模型的优势在于：现有的部分方法未考虑标签间的关联性，或考虑了相关性，但由于串联分类器存在误差或样本数量较少等原因，其关联会将误差传递积累，进而导致更多的误差。而提出的模型不需要准确地找出标签转换，只需将概率模型重新表述成原始的标签空间，在原始标签空间中自动获知和掌握标签的关联性，目标是获得能够实现同时对具有标签关联性和遗失标签情况进行多标签分类求解模型（具体求解处理见第2.2.2节和第2.3节）。

接下来对该模型进行一些分析。如果$\Sigma_1=\Sigma_2=\cdots=\Sigma_d=\Sigma$，则式(8)可重写成：

$$W\mid\Sigma\sim\mathcal{Q}_{d,m}(0,I\otimes\Sigma) \tag{12}$$

其中，$\otimes$是克罗内克积，$\mathcal{Q}_{d,m}(\cdot,\cdot)$是$d\times m$矩阵变量的正态分布[19]。可以看到，通过降低式(7)中的凑整误差，该模型可退化至具有输出和任务结构的多输出回归模

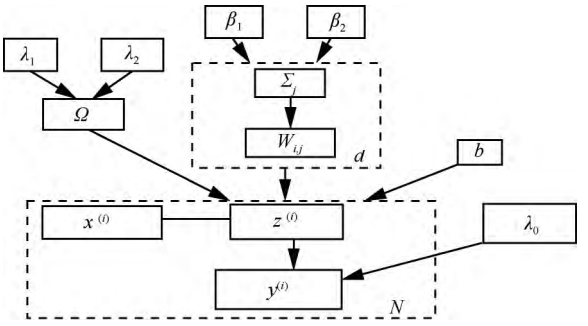

图 1　提出模型的一个图解表示

型(multi-output regression task structure, MROTS)[20]。对式(7)中凑整问题进行准确建模在分类问题中很重要。研究表明,凑整会导致编码误差$\left\|Y-P^{\dagger}\hat{Y}\right\|^2$,这种误差和转换标签空间中的训练误差可以一同用于指导寻找一个好的标签转换。正如 Rai 等人[20]所讨论的,对于多输出回归,MROTS 可以包含最近的很多方法,如基于协方差估计的多元回归(multivariate regression with covariance estimation, MRCE)[21]和多任务关联学习(multi-task relationship learning, MTRL)[22]。

2.2.2　模型推导求解

对建立的模型进行推导求解。对于给定的 N 个样点,令 $X=[x^{(1)},\cdots,x^{(N)}]$ 为输入矩阵,$Y=[y^{(1)},\cdots,y^{(N)}]$ 为对应的标签矩阵。其中,上标$^{(i)}$表示第 i 个样点。当 $Z=[z_1,\cdots,z_n]$ 时,则 $\{Z,W,\Omega,\{\Sigma_j\}_{j=1}^d\}$ 的后验分布可以表示为:

$$\begin{aligned}&p(Z,W,\Omega,\{\Sigma_j\}_{j=1}^d\mid X,Y,b)\\&\propto p(Y\mid Z)p(Z,W,\Omega,\{\Sigma_j\}_{j=1}^d\mid X,b)\\&=p(Y\mid Z)p(Z\mid W,X,\Omega,b)p(W\mid\{\Sigma_j\}_{j=1}^d)\cdot p(\Omega)p(\{\Sigma_j\}_{j=1}^d)\\&=p(\Omega)\prod_{j=1}^d p(W_{(j,\cdot)}\mid\Sigma_j)p(\Sigma_j)\cdot\prod_{i=1}^N p(y^{(i)}\mid z^{(i)})p(z^{(i)}\mid x^{(i)},W,\Omega,b)\end{aligned}\tag{13}$$

下面,使用交替最大化方法[23]对式(11)进行求解,交替获得后验信息最大值,即在式(13)中每次固定其他变量而求一个变量的最大值。这里,每一个子问题是凸性的,且容易求解,具体如下。

(1)固定其他变量,求解 Z

最优化子问题是:

$$\begin{aligned}\min_z&\sum_{i=1}^N\lambda_0\left\|z^{(i)}-y^{(i)}\right\|^2+(z^{(i)}-W^{\mathrm{T}}x^{(i)}-b)^{\mathrm{T}}\cdot\\&\Omega^{-1}(z^{(i)}-W^{\mathrm{T}}x^{(i)}-b)\end{aligned}\tag{14}$$

显然,每一个 $z^{(i)}$ 都可以单独求解。令 $z^{(i)}$ 的导数为 0,可以得到:

$$z^{(i)}=(\lambda_0 I+\Omega^{-1})^{-1}(\Omega^{-1}(W^{\mathrm{T}}x^{(i)}+b)+\lambda_0 y^{(i)})\tag{15}$$

(2)固定其他变量,求解 W 和 b

最优化子问题为:

$$\begin{aligned}\min_{W,b}&\operatorname{tr}(Z-W^{\mathrm{T}}x-bl^{\mathrm{T}})^{\mathrm{T}}\Omega^{-1}\\&(Z-W^{\mathrm{T}}x-bl^{\mathrm{T}})+\sum_{j=1}^d W_{(j,\cdot)}\Sigma_j^{-1}W_{(j,\cdot)}^{\mathrm{T}}\end{aligned}\tag{16}$$

令 W 的导数为 0,可以得到一个封闭形式的解:

$$\operatorname{vec}(W)=\left(\Omega^{-1}\otimes(XX^{\mathrm{T}})+\sum_{j=1}^d\Sigma_j^{-1}\otimes E_j\right)^{-1}\operatorname{vec}(C)\tag{17}$$

其中,$C=X(Z-bl)^{\mathrm{T}}\Omega^{-1}$,对于 b,令其导数为 0,可以得到:

$$b=\frac{1}{N}(Z-W^{\mathrm{T}}X)l\tag{18}$$

(3)固定其他变量,求解 Ω^{-1}

根据式(6)的先验信息,最优化子问题为:

$$\begin{aligned}\min_{\Omega^{-1}}&\operatorname{tr}(Z-W^{\mathrm{T}}X-bl^{\mathrm{T}})^{\mathrm{T}}\Omega^{-1}(Z-W^{\mathrm{T}}X-bl^{\mathrm{T}})-\\&N\lg\left|\Omega^{-1}\right|+\lambda_1\operatorname{tr}(\Omega^{-1})+\lambda_2\left\|\Omega^{-1}\right\|_1\end{aligned}\tag{19}$$

Ω^{-1} 可以通过采用标准的稀疏协方差估计算法求解[25]。

(4)固定其他变量,求解 Σ_j^{-1}

采用式(11)中的先验信息,对每一个 Σ_j 可以得到最优化子问题:

$$\min_{\Sigma_j^{-1}}W_{(j,\cdot)}\Sigma_j^{-1}W_{(j,\cdot)}^{\mathrm{T}}-\lg\left|\Sigma_j^{-1}\right|+\beta_1\operatorname{tr}(\Sigma_j^{-1})+\beta_2\left\|\Sigma_j^{-1}\right\|_1\tag{20}$$

同样,Σ_j^{-1} 也可以通过稀疏的逆协方差估计算法进行求解。

2.3　遗失标签处理

基于提出的模型,对遗失标签情况进行处理。如前所说,标签向量也许会有一些遗失记录。假设样点 $x^{(i)}$ 具有 l_i 个观测标签以及 $u_i=m-l_i$ 个遗失标签,可将 $y^{(i)}$ 和 $z^{(i)}$ 分别记为 $[(y_l^{(i)})^{\mathrm{T}},(y_u^{(i)})^{\mathrm{T}}]^{\mathrm{T}}$ 和 $[(z_l^{(i)})^{\mathrm{T}},(z_u^{(i)})^{\mathrm{T}}]^{\mathrm{T}}$,其中,$y_l^{(i)}\in\mathbb{R}^{l_i}$,$y_u^{(i)}\in\mathbb{R}^{u_i}$。同样,对每一个 i,通过将第 l_i 行/列与首次观测的标签进行对应,可将 Ω^{-1} 记为 $\begin{bmatrix}U_i & V_i\\ V_i^{\mathrm{T}} & Q_i\end{bmatrix}$,其中,$U_i\in\mathbb{R}^{l_i\times l_i}$,$V_i\in\mathbb{R}^{l_i\times u_i}$,$Q_i\in\mathbb{R}^{u_i\times u_i}$。

不同于直接估算遗失标签值的方法[15],这里所获得的信息直接来源于观测标签,类似于式(13),可以得到:

$$\begin{aligned}&p(\{z_{l_i}^{(i)}\}_{i=1}^N,W,\Omega,\{\Sigma_j\}_{j=1}^d\mid X,\{y_l^{(i)}\}_{i=1}^N,b)\propto p(\Omega)\\&\prod_{j=1}^d p(W_{(j,\cdot)}\mid\Sigma_j)p(\Sigma_j)\cdot\prod_{j=1}^d p(y_l^{(i)}\mid z_l^{(i)})p(z_l^{(i)}\mid x^{(i)},W,\Omega,b)\end{aligned}\tag{21}$$

注意：$p(y_l^{(i)} \mid z_l^{(i)}) = \Pi_{j \in l_i} p(y_j^{(i)} \mid z_j^{(i)})$，因此，$p(y^{(i)} \mid z^{(i)})$也容易计算得到。重要地，$p(z_l^{(i)} \mid x^{(i)}, W, \Omega, b)$呈现了遗失标签元素的计算，即：

$$p(z_l^{(i)} \mid W, x^{(i)}, \Omega, b) = \int p([(z_l^{(i)})^{\mathrm{T}}, (z_u^{(i)})^{\mathrm{T}}]^{\mathrm{T}} \mid W, b, x^{(i)}, \Omega) \mathrm{d}z_u^{(i)} \quad (22)$$

则仍然满足正态分布[24]：

$$z_l^{(i)} \mid W, x^{(i)}, \Omega \sim \mathcal{N}(W_{(:,l_i)}^{\mathrm{T}} x^{(i)} + b_{l_i}, \tilde{U}_i) \quad (23)$$

其中，$\tilde{U}_i = U_i - V_i Q_l^{-1} V_i^{\mathrm{T}}$，$W_{(:,l_i)}$是 W 的子矩阵，W 的列对应于 l_i 观测标签。注意：每一个$z_l^{(i)}$在整个Ω矩阵中相对于$\tilde{U}_i$是独立的。因此，尽管有遗失标签存在，推理过程仍然可以利用标签关联信息。

如第 2.2.2 节所示，本节也将使用交替最大化法来求解式(21)中的后验信息最大值。对于Σ_j^{-1}的最大化子问题与之前相同，因此其校正不变。

(1)固定其他变量，求解$\{z_{l_i}^{(i)}\}_{i=1}^N$

最优化子问题为：

$$\min_{z_l^{(i)}} \sum_{i=1}^N \left\| z_{l_i}^{(i)} - y_{l_i}^{(i)} \right\|^2 + (z_l^{(i)} - W_{(:,l_i)}^{\mathrm{T}} x - b_{l_i})^{\mathrm{T}} \tilde{U}_i (z_l^{(i)} - W_{(:,l_i)}^{\mathrm{T}} x - b_{l_i}) \quad (24)$$

令每一个$z_l^{(i)}$的导数为 0，可以得到：

$$z_l^{(i)} = W_{(:,l_i)}^{\mathrm{T}} x^{(i)} + b_{l_i} - \tilde{U}_i^{-1} y_l^{(i)} \quad (25)$$

(2)固定其他变量，求解 W 和 b

最优化子问题为：

$$\min_{W,b} \sum_{i=1}^N (z_l^{(i)} - W_{(:,l_i)}^{\mathrm{T}} x - b_{l_i})^{\mathrm{T}} \tilde{U}_i (z_l^{(i)} - W_{(:,l_i)}^{\mathrm{T}} x - b_{l_i}) + \sum_{j=1}^d W_{(j,:)} \Sigma_j^{-1} W_{(j,:)}^{\mathrm{T}} \quad (26)$$

不同于式(16)，这里无法对该凸问题得到封闭解，而需要通过降低梯度对 W 进行最优化，即对 b 可得到一个封闭解：

$$b = \left(\sum_{i=1}^N \Xi_i(\tilde{U}_i) \right)^{-1} \sum_{i=1}^N \Xi_i(\tilde{U}_i (z_l^{(i)} - W_{(:,l_i)}^{\mathrm{T}} x^{(i)})) \quad (27)$$

其中，Ξ_i是一个将矩阵$A \in \mathbb{R}^{l_i \times l_i}$扩展为$\begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m \times m}$的操作。

(3)固定其他变量，求解Ω^{-1}

最优化子问题为：

$$\min_{\Omega^{-1}} \sum_{i=1}^N ((y_l^{(i)} - W_{(:,l_i)}^{\mathrm{T}} x^{(i)} - b_{l_i})^{\mathrm{T}} \tilde{U}_i (y_l^{(i)} - W_{(:,l_i)}^{\mathrm{T}} x^{(i)} - b_{l_i}) - \ln|\tilde{U}_i|) + \lambda_1 tr(\Omega^{-1}) + \lambda_2 \left\| \Omega^{-1} \right\|_1 \quad (28)$$

通过软阈值迭代算法[25]进行求解，将式(28)分解为两部分：

$$f(\Omega) = ((y_l^{(i)} - W_{(:,l_i)}^{\mathrm{T}} x^{(i)} - b_{l_i})^{\mathrm{T}} \cdot \tilde{U}_i (y_l^{(i)} - W_{(:,l_i)}^{\mathrm{T}} x^{(i)} - b_{l_i}) - \ln|\tilde{U}_i|) + \lambda_1 tr(\Omega^{-1}) \quad (29)$$

$$g(\Omega) = \lambda_2 \left\| \Omega^{-1} \right\|_1 \quad (30)$$

Ω是半正定的，而不是投影梯度下降的，即每一次迭代，需要投影到半正定的圆锥体上，这在计算方面比较耗时。因此，需要根据因式分解对Ω^{-1}进行校正。在每一次迭代中有：

- 将Ω^{-1}因式分解为 GG^{T}，对 $f(\Omega)$计算关于 G 的一阶梯度降；
- 根据修正 G 的重新计算Ω^{-1}；
- 通过对Ω^{-1}的每一个因子进行收缩稀疏化：

$$\max(|(\Omega^{-1})_{ij}| - \tau, 0) \cdot \mathrm{sign}((\Omega^{-1})_{ij}) \quad (31)$$

其中，$\tau = \lambda_2 \eta$，η 是梯度降的补偿。

3 实验效果与分析

在第 3 节中，对 Guillaumin 等人[26]采用的 4 个图像标注的数据集(表 1)进行试验。对每一个图像都提取 1 000个 SIFT 特征。

表 1 数据集

数据集	标签数/个	样点数/个	每个样点的平均正标签数/个	每个样点的最大负标签数/个
MIRFLICKR	38	25 000	4.7	17
COREL5K	260	4 999	3.4	5
ESPGAME	268	23 641	4.7	15
IAPRTCL12	291	19 627	5.7	23

将提出的处理非完备数据和标签关联的多标签分类方法 (multi-label classification with incomplete labeled data and label relevance，ILDLR)与下列经典方法进行对比分析：

- 二值相关(BR)算法[4]；
- 条件的主标签空间转换(CPLST)法[10]；
- 具有群索引的多标签排序(MLRGL)法[15]；
- 快速图像标记(FastTag)法[16]。

由于在 CPLST 方法中的转换标签是实际值，因此，使用脊回归作为其初始模型。所有方法的参数调整都基于一

个已验证的集合，该集合通过对实验数据采样 25%获得。下面两个参数被广泛的用于多标签分类的效果评价[26]：

$$\text{macro}\_F1=\frac{1}{m}\sum_{j=1}^{m}\frac{2\sum_{i=1}^{N}\hat{y}_j^{(i)}y_j^{(i)}}{\sum_{i=1}^{N}\hat{y}_j^{(i)}+\sum_{i=1}^{N}y_j^{(i)}} \tag{32}$$

$$\text{micro}\_F1=\frac{1}{N}\sum_{i=1}^{N}\frac{2\sum_{j=1}^{m}\hat{y}_j^{(i)}y_j^{(i)}}{\sum_{j=1}^{m}\hat{y}_j^{(i)}+\sum_{j=1}^{m}y_j^{(i)}} \tag{33}$$

其中，macro_$F1$ 和 micro_$F1$ 值越大，效果越好。

3.1　在完备标签数据上的实验

在 4 组完备数据集上进行实验，并与其他几种方法进行对比。基于 10 折交叉验证方法，对数据进行分类及 $F1$ 值评价，结果见表 2，其中，“±”表示 10 次验证计算的 $F1$ 评价值的浮动范围。可以看到，在完备标签的情况下，提出的 ILDLR 方法对所有的数据处理效果都比其他方法要好。对于 MLRGL 方法，在 COREL5K、ESPGAME 和 IAPRTCL12 3 个数据集上的处理效果与其他几种方法的效果差距较大。

3.2　在非完备标签数据上的实验

通过以下方式生成遗失标签。之前提及有 m 个标签，对每一个训练样本集，选择其中的一半作为观测值，其余的为遗失值。由于每个样点都普遍具有极少的正标签，且标签进行随机拆分有可能导致仅仅观测到负标签，因此，对每一个样点，保证有 k=1,2,3 个观测到的正标签（如果一个样本的正标签数少于 k 个，那么它所有的正标签将被挑选出来）。将 ILDLR、MLRGL、FastTag 和 BR 4 种方法进行比较，这 4 种方法都能处理遗失标签。

表 3 呈现的是 10 折交叉验证的结果，可以看到，ILDLR 的效果要优于其他方法（除了在 MIRFLICKR 数据集中当 k=2 时）。当遗失标签的数量远大于 k 的最大值时，效果并不会随着 k 的增大而变好。并且当 k=1 时，通过 ILDLR 获得的 $F1$ 与表 2 中具有完备标签所获得的值很接近。对于数据 COREL5K，ILDLR 在标签遗失的情况下反而效果更好，其原因是：对于完备标签，需要学习 $m\times n$ 的标签矩阵 Z。而当标签遗失时，只需要学习观测标签矩阵 Z 对应的子矩阵。这样，尽管式（22）依赖于分布 $p([(z_l^{(i)})^{\mathrm{T}},(z_u^{(i)})^{\mathrm{T}}]^{\mathrm{T}}\mid W,x^{(i)},\Omega)$ 的结果，且可能会引进一些误差，但减少了自由参数的个数。因此，提出的 ILDLR 方法在标签遗失的情况下也会得到很好的处理效果。

4　结束语

对多标签分类提出了一个概率模型，然后受标签转换方法的启发，此模型在原始标签空间而不是在转换标签空间进行表述，这可以灵活地处理标签相关和遗失标签情况，并且推导过程简单。在完备的数据和具有非完备的遗失标签数据上的实验都表明，提出的方法比现有的其他经典方法效果更好。

参考文献：

[1] ZHANG M L, ZHOU Z H. A review on multi-label learning algorithms [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837.

[2] MADJAROV G, KOCEV D, GJORGJEVIKJ D, et al. An extensive experimental comparison of methods for multi-label learning[J]. Pattern Recognition, 2012, 45(9): 3084-3104.

[3] 王霄，周李威，陈耿，等. 一种基于标签相关性的多标签分类算法[J]. 计算机应用研究，2014, 31(9): 2609-2614.
WANG X, ZHOU L W, CHEN G, et al. Correlation label-based multi-label classification algorithm.[J]. Application Research of Computers, 2014, 31(9): 2609-2614.

[4] TSOUMAKAS G, KATAKIS I, VLAHAVAS I. Mining multi-label data[M]. New York: Springer, 2010: 667-685.

表 2　具有完备标签数据的结果

参数	数据集	ILDLR	BR	CPLST	MLRGL	FastTag
macro_$F1$	MIRFLICKR	0.499 6±0.005 1	0.494 1±0.004 2	0.493 2±0.004 2	0.495 6±0.007 3	0.467 9±0.005 3
	COREL5K	0.208 1±0.056 3	0.201 0±0.046 6	0.200 4±0.049 3	0.131 5±0.044 4	0.203 6±0.038 0
	ESPGAME	0.237 8±0.013 1	0.231 4±0.016 2	0.232 7±0.016 7	0.125 1±0.007 2	0.228 4±0.013
	IAPRTCL12	0.246 7±0.040 5	0.235 9±0.042 2	0.237 5±0.042 4	0.126 7±0.041 1	0.228 5±0.041 1
micro_$F1$	MIRFLICKR	0.467 1±0.003 4	0.461 1±0.005 5	0.461 1±0.004 7	0.466 1±0.009 2	0.436 2±0.006 1
	COREL5K	0.208 3±0.055 5	0.196 7±0.048 1	0.198 2±0.048 1	0.130 5±0.043 6	0.202 2±0.036 2
	ESPGAME	0.227 6±0.014 0	0.221±0.017 3	0.221 7±0.018 0	0.118 9±0.001 0	0.218 9±0.017 4
	IAPRTCL12	0.240 8±0.045 1	0.231 5±0.043 1	0.231 0±0.042 7	0.123 9±0.041 2	0.221 1±0.043 6

表 3　10 折交叉验证结果

参数	数据集	ILDLR	MLRGL	FastTag	BR
macro_$F1(k=1)$	MIRFLICKR	0.467 5±0.001 4	0.458 1±0.008 9	0.461 1±0.010 2	0.455 6±0.002 1
	COREL5K	0.240 7±0.054 1	0.110 8±0.015 1	0.199 7±0.010 1	0.142 2±0.027 9
	ESPGAME	0.233 1±0.019 1	0.147 1±0.011 1	0.179 1±0.007 8	0.193 9±0.014 1
	IAPRTCL12	0.237 1±0.045 3	0.058 9±0.038 9	0.192 4±0.004 9	0.209 6±0.033 4
micro_$F1(k=1)$	MIRFLICKR	0.467 8±0.002 1	0.457 8±0.006 7	0.462 3±0.010 3	0.453 4±0.001 1
	COREL5K	0.240 0±0.053 9	0.111 1±0.013 4	0.198 7±0.011 1	0.141 9±0.027 8
	ESPGAME	0.233 1±0.019 2	0.144 8±0.010 1	0.187 9±0.007 8	0.193 2±0.014 3
	IAPRTCL12	0.236 1±0.045 3	0.056 7±0.040 1	0.193 1±0.004 2	0.209 6±0.034 4
macro_$F1(k=2)$	MIRFLICKR	0.453 2±0.006 5	0.456 7±0.007 8	0.434 4±0.051 0	0.443 4±0.002 6
	COREL5K	0.240 5±0.050 9	0.110 5±0.015 1	0.181 2±0.011 9	0.016 7±0.011 1
	ESPGAME	0.231 1±0.020 2	0.145 5±0.011 4	0.221 6±0.004 9	0.183 4±0.012 6
	IAPRTCL12	0.234 4±0.043 7	0.058 8±0.038 7	0.181 1±0.006 5	0.197 1±0.033 1
micro_$F1(k=2)$	MIRFLICKR	0.435 4±0.005 6	0.439 8±0.006 5	0.415 4±0.044 5	0.422 1±0.004 6
	COREL5K	0.239 7±0.053 5	0.111 1±0.014 3	0.181 1±0.005 7	0.015 6±0.009 1
	ESPGAME	0.223 1±0.024 5	0.144 5±0.012 1	0.216 7±0.008 9	0.174 4±0.013 9
	IAPRTCL12	0.231 1±0.035 4	0.058 7±0.037 7	0.185 7±0.008 4	0.193 1±0.032 1
macro_$F1(k=3)$	MIRFLICKR	0.460 1±0.002 3	0.456 7±0.008 7	0.454 3±0.017 8	0.412 1±0.001 1
	COREL5K	0.235 6±0.053 1	0.112 1±0.014 3	0.180 1±0.004 5	0.025 6±0.015 4
	ESPGAME	0.223 4±0.023 2	0.145 7±0.011 1	0.214 5±0.006 7	0.160 1±0.010 1
	IAPRTCL12	0.231 1±0.033 4	0.058 8±0.037 8	0.183 2±0.004 3	0.164 5±0.023 4
micro_$F1(k=3)$	MIRFLICKR	0.444 4±0.003 1	0.435 6±0.004 3	0.415 1±0.016 1	0.386 7±0.002 4
	COREL5K	0.236 7±0.052 1	0.107 6±0.015 4	0.173 2±0.005 6	0.024 5±0.016 1
	ESPGAME	0.212 3±0.025 5	0.142 1±0.011 2	0.203 4±0.005 6	0.156 7±0.012 3
	IAPRTCL12	0.224 5±0.037 8	0.056 7±0.038 1	0.174 5±0.004 8	0.164 3±0.022 1

[5] READ J, PFAHRINGER B, HOLMES G, et al. Classifier chains for multi-label classification [C]//European Conference on Machine Learning, June 14-18, 2009, Montreal, Canada. New Jersey: IEEE Press, 2009: 254-269.

[6] **孙霞，张敏超，冯筠，等**. Hadoop **框架下的多标签传播算法**[J]. **西安交通大学学报**, 2015, 49(5): 134-139.
SUN X, ZHANG M C, FENG J, et al. A label propagation algorithm for multi-label classification [J]. Journal of Xi'an Jiaotong University, 2015, 49(5): 134-139.

[7] HARIHARAN B, ZELNIK M L, VISHWANATHAN S, et al. Large scale max-margin multi-label classification with priors[C]// 27th International Conference on Machine Learning, June 21-24, 2010, Haifa, Isreal. New Jersey: IEEE Press, 2010: 423-430.

[8] HSU D, KAKADE S, LANGFORD J, et al. Multi-label prediction via compressed sensing[J]. Computer Science, 2009: 772-780.

[9] TAI F, LIN H. Multi-label classification with principal label space transformation [J]. Neural Computation, 2012, 24 (9): 2508-2542.

[10] CHEN Y N, LIN H T. Feature-aware label space dimension reduction for multi-label classification [J]. Advances in Neural Information Processing Systems , 2012(2): 1538-1546.

[11] **马宗杰，刘华文**. **基于奇异值分解-偏最小二乘回归的多标签分类算法**[J]. **计算机应用**, 2014, 34(7): 2058-2061.
MA Z J, LIU H W. Multi-label classification based on singular value decomposition-partial least squares regression[J]. Journal of Computer Applications, 2014, 34(7): 2058-2061.

[12] **李远航，刘波，唐侨**. **面对多标签图数据的主动学习**[J]. **计算机科学**, 2014, 41(11): 260-264.
LI Y H, LIU B, TANG Q. Active learning for multi-label classification on graphs[J]. Computer Science, 2014, 41(11): 260-264.

[13] **许美香，孙福明，李豪杰**. **主动学习的多标签图像分类在线分类**[J]. **中国图像图形学报**, 2015, 20(2): 237-244.
XU M X, SUN F M, LI H J. Online multi-label image classification with active learning [J]. Journal of Image and Graphics, 2015, 20(2): 237-244.

[14] **徐晓丹，姚明海，刘华文，等**. **基于** kNN **的多标签分类预处理方法**[J]. **计算机科学**, 2015, 42(5): 106-108.
XU X D, YAO M H, LIU H W, et al. Pre-processing method of multi-label classification based on kNN[J]. Computer Science, 2015, 42(5): 106-108.

[15] BUCK S, JIN R, JAIN A. Multi-label learning with incomplete class assignments [C]//IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2011, Providence, RI, USA. New Jersey: IEEE Press, 2012: 2801-2808.

[16] CHEN M, ZHENG A, WEINBERGER K Q. Fast image tagging[C]//30th International Conference on Machine Learning, June 16-21, 2013, Atlanta, GA, USA. New Jersey: IEEE Press, 2013: 1274-1282.

[17] YU H F, JAIN P, DHILLON I S. Large-scale multi-label learning with missing labels[C]//31st International Conference on Machine Learning, June 21-26, 2014, Beijing, China. New Jersey: IEEE Press, 2014: 593-601.

[18] PETTERSON J, CAETANO T. Submodular multi-label learning[J]. Advances in Neural Information Processing Systems, 2011: 1512-1520.

[19] GUPTA A, NAGAR D. Matrix variate distributions [M]. Boca Raton: Chapman & Hall/CRC Press, 2000.

[20] RAI P, KUMAR A, III H D. Simultaneously leveraging output and task structures for multiple-output regression[C]//Advances in Neural Information Processing Systems, December 3-8, 2012, South Lake Tahoe, USA. New Jersey: IEEE Press, 2012: 3194-3202.

[21] ROTHMAN A J, LEVINA E, ZHU J. Sparse multivariate regression with covariance estimation[J]. Journal of Computational and Graphical Statistics, 2010, 19(4): 947-962.

[22] ZHANG Y, YEUNG D Y. A convex formulation for learning task relationships in multi-task learning [C]//26th Conference on Uncertainty in Artificial Intelligence, July 8-11, 2010, Los Angeles, USA. New Jersey: IEEE Press, 2010: 733-742.

[23] BERTSEKAS D P. Nonlinear programming [M]. Nashua: Athena Scientific, 1999.

[24] BISHOP C M. Pattern recognition and machine learning[M]. New York: Springer-Verlag, 2006: 125-153.

[25] BECK A, TEBOULLE M. A fast iterative shrinkage-thresholding algorithm for linear inverse problem[J]. SIAM Journal on Imaging Sciences, 2009, 2(1): 183-202.

[26] GUILLAUMIN M, MENSINK T, VERBEEK J, et al. Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation [C]//International Conference on Computer Vision, September 29-October 2, 2009, Kyoto, Japan. New Jersey: IEEE Press, 2009: 309-316.

[作者简介]

张丽娜（1980–），女，浙江安防职业技术学院信息工程系讲师，主要研究方向为数据挖掘、图像处理、模式识别。

戴灵鹏（1975–），男，博士，温州大学生命与环境科学学院副教授，主要研究方向为模式识别。

匡泰（1964–），男，浙江安防职业技术学院信息工程系副教授，主要研究方向为大数据、人工智能。