第 56 卷 第 4 期
2020 年 7 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 56, No. 4
July, 2020

DOI:10.13232/j.cnki.jnju.2020.04.013

一种基于嵌入式的弱标记分类算法

李亚重[1],杨有龙[1*],仇海全[1,2]

(1.西安电子科技大学数学与统计学院,西安,710126;2.安徽科技学院信息与网络工程学院,蚌埠,233030)

摘 要:对于高维标签的分类问题,标签嵌入法已经受到广泛关注.现有的嵌入方法大都需要完整的标签信息,也没有将特征空间考虑在内;同时,由于数据进行人工标注的成本高以及噪声干扰等原因,仅能获得数据的部分标签信息,使得含有缺失标签的高维标签分类问题变得更加复杂.为解决这一问题,提出一种弱标记嵌入算法(Label Embedding for Weak Label Classification,LEWL).该算法利用矩阵的低秩分解模型,结合样本的流形结构恢复缺失标签;同时采用希尔伯特-施密特独立标准技术(Hilbert-Schmidt Independence Criterion,HSIC)使特征和标签相互作用,联合学习获得一个低维的嵌入空间,可以有效地减少模型的训练时间.通过在七个多标签数据集上与其他算法的对比实验,结果表明了所提算法的有效性.

关键词:弱标记学习,标签嵌入,低秩分解,希尔伯特-施密特独立标准,缺失标签

中图分类号:TP301.6 **文献标识码**:A

Label embedding for weak label classification

Li Yachong[1], Yang Youlong[1*], Qiu Haiquan[1,2]

(1. School of Mathematics and Statistics, Xidian University, Xi'an, 710126, China;
2. College of Information & Network Engineering, Anhui Science and Technology University, Bengbu, 233030, China)

Abstract: For the classification of high-dimensional labels, label embedding has attracted extensive attention of researchers in recent years. Current embedding methods require complete label information and do not take feature information into consideration. Meanwhile, due to the high cost of manual labeling and interference of noise, only part of the label information can be obtained. This makes the classification problem of high-dimensional labels with missing labels more complicated. To end this, a Label Embedding method for Weak Label Classification (LEWL) is proposed in this paper. The algorithm uses the low-rank factorization model on the label matrix and the flow pattern structure of the samples to recover the missing labels. In the meantime, the HSIC (Hilbert-Schmidt Independence Criterion) technique is adopted to obtain the low dimensional embedding space by making feature and labels interact with each other for joint learning, which can effectively reduce the training time of the model. Compared with other methods on seven data sets, comprehensive experimental results validate the effectiveness of proposed approach.

Key words: weak label classification, label embedding, the low-rank factorization on the matrix, HSIC(Hilbert-Schmidt Independence Criterion), missing labels

多标签分类中一个实例常常同时拥有多个标签,目前已经在文本分类[1]、图像注解[2]、基因功能分类[3]等方面得到广泛应用.现有的多标签分类算法可分两类:一类是问题转换法(Problem

基金项目:国家自然科学基金(61573266),安徽省高校自然科学研究重点项目(KJ2019A0816)

收稿日期:2020-03-03

*通讯联系人,E-mail:ylyang@mail.xidian.edu.cn

Transformation,PT),基本思想是把多标签分类任务转换成多个单标签分类任务,如:BR (Binary Relevance)[4]算法、LP (Label Powerset)[5]算法、CC (Classifier Chain)[6]算法等. 另一类是算法转换法(AA,Algorithm Adaptation),即对现有的单标签分类算法进行改进使其能够处理多标签数据. 代表算法有:ML-KNN[7],Adaboost. MH[8]等. 随着数据采集和存储技术的发展[9],标签数量呈指数型爆炸式增长,传统的多标签分类算法需要较大的时间成本,也面临着标签维数过高引起的"维数灾难"问题. 同时,多标签学习的输出是多个类别标签,且标签之间存在相关性,增加了模型学习的难度. 为了缓解这个问题,研究人员开始研究标记空间维度下降法(Label Space Dimension Reduction,LSDR),利用标签之间的关系来降低标签空间的维度,希望在提升分类精度的同时能有效减少整个模型的训练和预测时间.

标记空间维度下降法是针对高维标签向量提出的一种嵌入技术,把初始的标签空间转化成低维的嵌入空间,在低维嵌入空间中实现对向量更有效的表示[10]. 对于一个测试数据来说,学习器将其映射到低维嵌入空间,再通过解码器将其恢复到原始的二值空间,最终希望预测到的仍是原始标签空间下的标签向量. Hsu et al[11]认为标签空间具有输出稀疏性,标签向量存在小支撑,首次提出基于压缩感知的多标签预测方法(Multi-Label Prediction via Compressed Sensing,ML-CS),即利用压缩感知理论对标签空间进行压缩. 该方法采用随机生成的压缩函数,不能有效利用标签之间的关系来实现更好的压缩效果. Tai and Lin[12]提出PLST (Principal Label Space Transformation),通过对标签矩阵进行奇异值分解来降维. Chen and Lin[13]在PLST基础上提出CPLST (Conditional PLST)方法,在标签重构过程中引入相关特征信息,进一步提高模型对未知数据预测的准确率. 最近基于典型相关分析(Canonical Correlation Analysis,CCA)理论,Lin et al[14]提出E^2FE(End-to-End Feature-aware Label Space Encoding),该方法无需对编码方式进行任何假设,避免了不合理假设造成的风险. 由于对数据进行人工标注的成本太高、用户更新频率大及噪声干扰等其它原因[15],获取训练数据的全部标签显得非常困难. 在这种情况下产生了弱标记数据,即实例中含有未被标记或标记错误的标签. 本文主要讨论前一种情况,即数据只有部分标签信息可以获得. Sun et al[16]最早将弱标签问题引入多标签学习,并提出WELL (WEak Leak Learning),让每个标签的分类边界跨越低密度区域,并考虑了类不平衡问题. MLML (Multi-label Learning with Missing Labels)[17]算法首次明确区分负类标签和缺失标签,即正类、负类和缺失标签分别用+1,−1和0表示,该算法基于标签一致性和标签平滑性求出恢复后的完整标签. LRML (Low Rank Multi-label Classification with Missing Labels)[18]算法利用标签一致性(label consistency)和局部不变性(local invariance)假设得到完整的标签矩阵,同时又从特征空间到恢复后的标签空间学习线性函数矩阵,并假设其是低秩的. Han et al[19]提出ColEmbed (Collaborative Embedding),利用非线性嵌入将特征和标签嵌入到一个共享子空间,同时解决了特征不完整和标签缺失问题. 综上,解决高维且含有缺失标签的多标签学习问题十分必要.

为了解决上述问题,本文提出一种基于嵌入式的弱标记算法LEWL (Label Embedding for Weak Label Classification). 一方面通过对标签矩阵进行低秩分解来最小化嵌入空间返回原始标签空间的恢复误差. 为了提高对嵌入空间的可预测性,采用希尔伯特-施密特独立标准技术(Hilbert-Schmidt Independence Criterion,HSIC)使得特征空间和嵌入空间的依赖关系更加紧密,这样获得的实值低维嵌入空间把标签信息和特征信息同时考虑在内,标签和特征的嵌入过程是紧密相关同时进行的. 另一方面,由于矩阵的低秩分解对矩阵补全问题(标签恢复)起着重要作用,再利用样本流形结构对缺失标签进行填补. 最后将以上模型整合成一个优化问题,并提出了一个有效的求解方法. 实验结果表明,针对不同的数据集,本文提出的算法均具有较好的分类性能和泛化能力.

1　一种基于嵌入式的弱标记算法

1.1　定　义　$\boldsymbol{X}=\left[x_1,\cdots,x_n\right]^T\in R^{n\times m}$是由$n$个训练样本组成的数据矩阵，$Y=\left[y_1,\cdots,y_n\right]^T\in R^{n\times c}$是原始标签矩阵，$c$为类别个数，其中$\boldsymbol{y}_i\in\{0,0.5,1\}^c$是第$i$个样本$x_i$对应的标签．1，0，0.5分别表示正类、负类和缺失标签．若示例x_i含有第j个标签，则$Y_{i,j}=1$；若示例x_i没有第j个标签，则$Y_{i,j}=0$；$Y_{ij}=0.5$表示该示例缺失该标签．

1.2　标签嵌入模型　由于不同标签之间存在相关性，故可以把整个标签矩阵看成是低秩的[20−21]（即它的秩小于它的行数或列数）．例如，当标签“蓝天”和“白云”同时出现的时候，很有可能会出现标签“晴天”．对于一个矩阵来说，根据其部分元素来推断其所有元素也非常困难．Candès and Tao[22]证明在矩阵低秩的情况下，大多数矩阵可以通过求解核范数最小化问题来恢复元素．为降低求解的计算复杂度，Wen et al[23]提出一种基于矩阵分解的低秩拟合算法，目的是寻找两个或多个矩阵，其乘积对原始矩阵具有良好的逼近能力，即尽可能地减小分解矩阵乘积与原始矩阵之间的近似误差．利用这一思想，可以将原始标签矩阵$\boldsymbol{Y}\in R^{n\times c}$分解成两个规模更小的矩阵的乘积：

$$\boldsymbol{Y}=\boldsymbol{Z}\times\boldsymbol{D}\tag{1}$$

其中，$\boldsymbol{Z}\in R^{n\times d}$表示低维嵌入空间，集成了原始标签空间的整体信息，$\boldsymbol{D}\in R^{d\times c}$是基矩阵．

假设初始的标签矩阵表示为$Y=\left\{y_1,\cdots,y_n\right\}^T\in\{0,0.5,1\}^{n\times c}$，真实的标签矩阵为$\bar{\boldsymbol{Y}}$，定义$\Omega\subseteq\{1,\cdots,n\}\times\{1,\cdots,c\}$为观察到的$\boldsymbol{Y}$中的元素的位置集合（非0.5所对应的下标），$\Omega^c$为缺失元素的位置集合．通过最小化嵌入空间返回原始标签空间的恢复误差，提出下列低秩分解模型：

$$\begin{aligned}&\min_{Y,Z,D}\left\|\bar{Y}-\boldsymbol{ZD}\right\|_F^2\\&\quad s.t.\,\bar{Y}_{i,j}=Y_{i,j},\forall(i,j)\in\Omega\end{aligned}\tag{2}$$

从全局方面考虑了标签关系后，进一步从局部来考虑．样本平滑性假设：当两个样本x_i和x_j距离很近时，对应的标签$\bar{y}_i$和$\bar{y}_j$也是相似的．一般可以用正则化形式来呈现：

$$\sum_{i,j}^{n}\frac{1}{2}\omega_{i,j}\left\|\bar{y}_i-\bar{y}_j\right\|^2=tr(\bar{\boldsymbol{Y}}^T\boldsymbol{L}\bar{\boldsymbol{Y}})\tag{3}$$

其中$\boldsymbol{L}=\boldsymbol{P}-\boldsymbol{W}$是拉普拉斯矩阵，$\boldsymbol{P}\left(\boldsymbol{P}_{ii}=\sum_{j=1}^{n}\omega_{ij}\right)$为对角矩阵，$\boldsymbol{W}$表示样本相似性矩阵，如式(4)所示：

$$\omega_{i,j}=\begin{cases}\exp\left(-\dfrac{\left\|x_i-x_j\right\|^2}{t}\right)\\\qquad\qquad x_i\in N_k(x_j)\,\text{or}\,x_j\in N_k(x_i)\\0\qquad\qquad\text{otherwise}\end{cases}\tag{4}$$

于是得到下列缺失标签恢复模型：

$$\begin{aligned}&\min_{Y,Z,D}\left\|\bar{\boldsymbol{Y}}-\boldsymbol{ZD}\right\|_F^2+tr(\bar{\boldsymbol{Y}}^T\boldsymbol{L}\bar{\boldsymbol{Y}})\\&\quad s.t.\,\bar{\boldsymbol{Y}}_{i,j}=\boldsymbol{Y}_{i,j},\forall(i,j)\in\Omega\end{aligned}\tag{5}$$

这个模型能够通过恢复原始标签矩阵$\boldsymbol{Y}$中的缺失元素来得到真实的标签矩阵$\bar{\boldsymbol{Y}}$．而且求得的真实标签除了缺失标签之外，剩下的标签保持不变，可以避免把正确的标签恢复错误．

1.3　特征嵌入模型　理论研究表明[24]，强相关性往往意味着强可预测性．为了在标签空间压缩过程中更有效地使用特征信息，同时提高嵌入空间的可预测性（学习器在学习阶段学习相应的映射），本文认为应该尽量使嵌入空间与特征空间的依赖相关性最大化．

衡量依赖性的指标有很多，Gretton et al[25]于2005年提出希尔伯特-施密特独立标准（HSIC），该标准衡量的是两组变量在再生核希尔伯特空间（Reproducing Kernel Hilbert Spaces）中映射之间协方差算子的平方希尔伯特-施密特范数．由于其具有简洁的数学形式和优雅的理论性质，目前已经在很多领域被用来衡量变量之间的依赖性．

假定$\mathscr{F}$和$\mathscr{G}$分别表示两组变量$\boldsymbol{X}$和$\boldsymbol{Y}$的再生核希尔伯特空间，给定数据集$\boldsymbol{Z}=\left\{(x_1,y_1),\cdots,(x_n,y_n)\right\}\subseteq\boldsymbol{X}\times\boldsymbol{Y}$，则$HSIC$的一个经验型估计是：

$$HSIC(Z,\mathscr{F},\mathscr{G})=(n-1)^{-2}tr(\boldsymbol{KHLH})\tag{6}$$

其中，$\boldsymbol{H}_{n\times n}=\boldsymbol{I}-\dfrac{1}{n}\boldsymbol{ee}^t$是中心矩阵，$\boldsymbol{e}_{n\times1}$是一个全1向量，$\boldsymbol{I}_{n\times n}$是一个单位矩阵．$\boldsymbol{K}\in R^{n\times n}$和$\boldsymbol{L}\in R^{n\times n}$分别是特征核函数$\boldsymbol{K}_{i,j}=k(x_i,x_j)$和标签核函数$\boldsymbol{L}_{i,j}=l(x_i,x_j)$，显然值越大说明两个变量的依赖性越强．

为了方便起见,对于L本文考虑线性核$L=ZZ^T$. 由于$(n-1)^{-2}$是常数项,式(6)等同于$\max_Z tr(KHZZ^TH)=\max_Z tr(Z^THKHZ)$. 获得特征核矩阵$K$最有效的方法是利用高斯核函数进行计算[26],具体定义为:

$$K(x,x')=\exp\left(-\alpha\|x-x'\|^2\right)$$

其中,$\alpha>0$是核参数.

结合标签嵌入和特征嵌入模型得到最终的弱标记优化模型:

$$\begin{aligned}&\min_{\bar{Y},Z,D}\left\|\bar{Y}-ZD\right\|_F^2+\beta tr(\bar{Y}^TL\bar{Y})-\lambda tr(KHZZ^TH)\\&\quad s.t.\,\bar{Y}_{i,j}=Y_{i,j},\forall(i,j)\in\Omega\end{aligned}\tag{7}$$

为了便于求解,定义$P_\Omega(X)$为矩阵X在集合Ω上的正交投影算子:

$$\left(P_\Omega(X)\right)_{i,j}=\begin{cases}X_{i,j} & (i,j)\in\Omega\\0 & (i,j)\in\Omega^c\end{cases}\tag{8}$$

优化模型可被写成:

$$\begin{aligned}&\min_{\bar{Y},Z,D}\left\|\bar{Y}-ZD\right\|_F^2+\beta tr(\bar{Y}^TL\bar{Y})-\lambda tr(KHZZ^TH)\\&\quad s.t.P_\Omega(\bar{Y})=P_\Omega(Y)\end{aligned}\tag{9}$$

目标函数的前两项通过低秩分解和样本的流形结构来恢复缺失标签,第一项和第三项采用标签和特征相互作用联合学习获得了低维嵌入空间和解码矩阵. 因此,本文提出的优化模型不仅解决了弱标记问题,还有效地缓解了高维标签带来的影响,同时充分利用了标签之间的关系,提高了模型的分类精度.

2 模型的求解

为了更准确有效地求解目标函数,本文提出一种交替迭代的优化求解算法. 即将优化问题划分为几个子问题进行交替迭代更新,直至收敛到问题的稳定点或局部极值点.

(1)更新矩阵D:当固定矩阵Z和$\bar{Y}$时,D为唯一的变量,求解问题(9)可以转化为式(10):

$$L(D)=\min_D\left\|\bar{Y}-ZD\right\|_F^2\tag{10}$$

式(10)对D求导,并令其等于0:

$$\frac{\partial L(D)}{\partial D}=Z^T(ZD-\bar{Y})=0$$

可得到关于D的闭式解:

$$D=(Z^TZ)^{-1}Z^T\bar{Y}\tag{11}$$

为了消除嵌入空间中噪声和冗余信息的影响,同时使得解码过程更加方便,本文假设Z是列向量正交矩阵,即$Z^TZ=I$. 因此式(11)被重写为式(12):

$$D=Z^T\bar{Y}\tag{12}$$

(2)更新矩阵Z:类似的,当固定D和$\bar{Y}$时,目标函数(式(9))可写成:

$$\begin{aligned}L(Z)=&\min_Z\left\|\bar{Y}-ZD\right\|_F^2-\lambda tr(KHZZ^TH)=\\&\max_Z tr(Z^T\bar{Y}\bar{Y}^TZ)+tr(Z^THKHZ)=\\&\max_Z tr\left(Z^T(\bar{Y}\bar{Y}^T+HKH)Z\right)\\&s.t.Z^TZ=I\end{aligned}\tag{13}$$

通过特征值分解[13,27],优化问题(式(13))能容易地被求解,最后得出Z是由$A=YY^T+\lambda(HKH)$的前k个最大的特征值对应的特征向量组成.

(3)更新矩阵$\bar{Y}$:固定Z和D,从式(9)中可以得到关于$\bar{Y}$的优化模型:

$$\begin{aligned}L(\bar{Y})=&\min_{\bar{Y}}\left\|\bar{Y}-ZD\right\|_F^2+\beta tr(\bar{Y}^TL\bar{Y})\\&s.t.P_\Omega(\bar{Y})=P_\Omega(Y)\end{aligned}\tag{14}$$

为了求解式(14),需要引入拉格朗日乘子$\Lambda\in R^{n\times c}$,其拉格朗日函数为:

$$\begin{aligned}L(\bar{Y},\Lambda)=&\left\|\bar{Y}-ZD\right\|_F^2+\beta tr(\bar{Y}^TL\bar{Y})+\\&\Lambda\cdot P_\Omega(\bar{Y}-Y)\end{aligned}\tag{15}$$

要得到$\bar{Y}$的最优解,需先确定$P_\Omega(\bar{Y})$和$P_{\Omega^c}(\bar{Y})$分别对应的两个子问题. 其中,$P_\Omega(\bar{Y})$对应的子问题为:

$$\begin{aligned}L(\bar{Y},\Lambda)=&\left\|P_\Omega(\bar{Y}-ZD)\right\|_F^2+\beta tr\left(P_\Omega(\bar{Y}^TL\bar{Y})\right)+\\&\Lambda\cdot P_\Omega(\bar{Y}-Y)\end{aligned}\tag{16}$$

Karush-Kuhn-Tucker (KKT)条件需要如下式子成立:

$$\frac{\partial L}{\partial\Lambda}=P_\Omega(\bar{Y}-Y)=0\tag{17}$$

$$\frac{\partial L}{\partial\bar{Y}}=P_\Omega\left((\bar{Y}-ZD)+\beta L\bar{Y}\right)=\Lambda\tag{18}$$

另外,$P_{\Omega^c}(\bar{Y})$对应的子问题为:

$$L(\bar{Y})=\left\|P_{\Omega^c}(\bar{Y}-ZD)\right\|_F^2+\beta tr\left(P_{\Omega^c}(\bar{Y}^TL\bar{Y})\right)\tag{19}$$

类似的,它的KKT条件为:

$$\frac{\partial L}{\partial\bar{Y}}=P_{\Omega^c}\left((\bar{Y}-ZD)+\beta L\bar{Y}\right)=0\tag{20}$$

根据式(17)和式(20),得到 $\bar{Y}$ 的解:

$$P_{\Omega}(\bar{Y})=P_{\Omega}(Y)$$
$$P_{\Omega^c}(\bar{Y})=P_{\Omega^c}\left((I+\beta L)^{-1}ZD\right) \tag{21}$$

综上所述可以求得关于 $D, Z, \bar{Y}$ 的闭式解,其算法伪代码如下.

算法:LEWL算法

输入:数据矩阵 X,原始标签矩阵 Y,嵌入维数 K,参数 λ 和 β,精度 $\varepsilon=10^{-6}$,最大迭代次数 $N_{\mathrm{MAX}}=2000$

输出:补全的标签矩阵 $\bar{Y}$,测试数据 x_i 的可预测标签集.

1. 初始化:令 $\bar{Y}=Y$;
2. While不收敛do
3. 固定 $\bar{Y}$ 并根据式(13)更新 Z;
4. 固定 $\bar{Y}, Z$ 并根据式(12)更新 D;
5. 固定 Z, D 并根据式(21)更新 $\bar{Y}$;
6. 判断收敛条件:直至目标函数收敛(函数值不再变化或变化小于一定精度)或达到最大迭代次数 N_{MAX};
7. end while;
8. 基于 $\left\{(x_i,z_i)\right\}_{i=1}^{n}$ 训练得到回归模型 $f(x)$;
9. 给定一个新的测试样例 x_i,预测其标签集合:

$h(x_i)=round\left(f(x_i)D\right)$.

首先,从特征空间 X 到嵌入空间 Z 学习一个回归模型.由于 Z 维度较低,在训练过程中极大地降低了计算复杂度,提高了学习效率.其次,通过解码函数 D 将其解码到初始标签空间中.在整个过程中无须考虑编码形式.在步骤9中对测试数据预测时,结果可能包含非二值情况,这时需要选取一个阈值来进行决定属于哪一个类.Tai and Lin[12]证明固定值0.5是一个简单有效的方法.为了提升分类性能,本文采用类似文献[28－29]中提出的一种自适应阈值法.具体的,用步骤9对训练数据进行预测得到预测矩阵,将其预测值按降序(或升序)方式排列成一个一维向量,通过最大化(最小化)训练数据的评分标准来找到最好的分割点,如果超过这个阈值则为1,否则为0.

3　实　验

3.1　数据集　为了验证所提算法LEWL的有效性,从不同领域选取了七个公开的多标签数据集,具体信息如表1所示,其中标签基数(Label cardinality)表示每个样本的平均标签数.

表1　数据集基本信息

Table 1　The characteristics of the datasets

数据集	样本	特征	标签	领域	标签基数
Emotions	593	72	6	music	1.869
Yeast	2417	103	14	biology	4.237
CAL500	502	68	174	music	26.044
Medical	978	1449	45	text	1.245
Langlog	1460	1004	75	images	1.18
Enron	1702	1001	53	text	3.378
Corel5k	5000	499	374	images	3.522

3.2　评价指标　在多标签分类中由于每个样本同时归属多个标签,因此传统的单标签分类评价指标已不再适用.本文采用以下三个常见的多标签分类指标[30]评价算法的性能.对于一个给定的数据集 $D=\left\{(x_i,y_i)\middle|1\leqslant i\leqslant n\right\}$,其中 $y_i\in\{0,1\}^c$ 表示第 i 个样本的真实标记向量.

(1)排序损失(Rank Loss):用于衡量样本类别标签排序中出现排序错误的程度,如式(22)所示:

$$Rloss=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|Q_i\right|}{\left|y_i^{+}\right|\left|y_i^{-}\right|} \tag{22}$$

其中,

$$Q_i=\left\{(y',y'')\middle|f(x_i,y')\leqslant f(x_i,y''),\ (y',y'')\in y_i^{+}\times y_i^{-}\right\}$$

y_i^{+} 和 y_i^{-} 分别表示 y_i 中正类和负类标签的数目.Rank Loss值越小表明预测函数 f 的性能越好.

(2)平均精度(Average Precision):用于衡量整体样本标签预测的平均精确度,如式(23)所示:

$$Ap=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left|y_i^{+}\right|}\sum_{y\in y_i^{+}}\frac{\left|y'\middle|rank_f(x_i,y')\leqslant rank_f(x_i,y''),y'\in y_i^{+}\right|}{rank_f(x_i,y)} \tag{23}$$

Average Precision值越大表明预测函数 f 的性能越好.

(3)Macro F_1:是查准率(Precision)和召回率(Recall)的调和平均值,如式(24)所示:

$$\text{Macro}\,F_1=\frac{1}{c}\sum_{i=1}^{c}\frac{2p_ir_i}{p_i+r_i}$$
$$s.t.\,p_i=\frac{TP}{TP+FP},r_i=\frac{TP}{TP+FN} \tag{24}$$

其中,TP表示真正例的个数,即正类被预测为正类的数量;FP表示假正例的个数,即负类被预测为正类的数量;FN表示假负例的个数,即正类被预测为负类的数量. Macro F_1越大,表明算法的性能越好.

以上评价指标的值都在$[0,1]$区间中,一个好的多标签分类算法应该具有较低的Rank Loss及较高的Average Precision和Macro F_1.

3.3 实验设置 从标签数据中随机选取30%和70%的标签作为缺失标签,缺失标签用0.5表示. 为验证所提算法的有效性,与BR (Binary Relevance)[4], CPLST (Conditional Principal Label Space Transformation)[12], MLML (Multi-label Learning with Missing Labels)[17]和LRML (Low Rank multi-label classification with Missing Labels)[18]四种算法进行了比较. 由于缺失标签的存在,BR和CPLST不能直接进行分类和预测,为了方便起见它们采用最简单的填补方法,即把缺失标签看成负类标签来对待.

对于所有比对算法来说,利用随机森林从特征空间到嵌入空间学习训练模型. 模型参数如树的最大深度和个数通过灰度搜索分别从5,10,…,35和2,4,…,40中选择. 在本文的实验中,对每个数据集进行5折交叉验证,其中四份被用来训练,一份用来测试. 算法MLML中的权衡参数λ_x和λ_c、LRML中的α,β和γ以及所提算法LEWL中的λ和β通过交叉验证来选择.

3.4 结果及分析 通过前文提到的评价指标对多标签分类算法进行多角度比较,表2、表3和表4分别是在排序损失、平均精度和Macro F_1三个评价指标下对缺失标签的恢复情况,表5对它们的恢复结果进行了t检验,表6、表7和表8分别给出了在三个指标下对测试数据的预测结果. 其中ρ表示标签缺失率,表中的粗体字表明其在配对检验中优于其他结果,或与最优结果之间无显著性差异.

表2 缺失标签在平均精度上的恢复结果

Table 2 Recovery results for missing labels on Average Precision

Data set	ρ	LEWL	MLML	LRML	CPLST	BR
Emotions	0.3	0.918±0.008	**0.920±0.0044**	0.843±0.011	0.886±0.005	0.886±0.005
	0.7	0.839±0.011	**0.846±0.012**	0.772±0.013	0.781±0.010	0.781±0.010
Yeast	0.3	**0.894±0.003**	0.891±0.001	0.792±0.010	0.869±0.002	0.869±0.002
	0.7	**0.811±0.004**	0.805±0.003	0.716±0.011	0.738±0.005	0.738±0.005
CAL500	0.3	**0.832±0.006**	0.791±0.005	0.784±0.014	0.816±0.004	0.816±0.004
	0.7	**0.663±0.006**	0.633±0.007	0.652±0.010	0.623±0.004	0.623±0.004
Medical	0.3	**0.790±0.019**	0.776±0.029	0.772±0.042	0.766±0.021	0.766±0.021
	0.7	**0.602±0.024**	0.588±0.029	0.531±0.044	0.542±0.009	0.542±0.009
Langlog	0.3	**0.860±0.005**	0.811±0.002	0.852±0.031	0.827±0.002	0.827±0.002
	0.7	**0.709±0.006**	0.667±0.003	0.704±0.022	0.643±0.006	0.643±0.006
Enron	0.3	0.819±0.007	**0.823±0.006**	0.764±0.051	0.795±0.006	0.795±0.006
	0.7	0.632±0.020	**0.667±0.028**	0.543±0.058	0.585±0.013	0.585±0.013
Corel5k	0.3	**0.749±0.005**	0.735±0.006	0.742±0.004	0.733±0.002	0.733±0.002
	0.7	**0.515±0.008**	0.510±0.011	0.513±0.009	0.511±0.004	0.511±0.004

从实验结果可以很明显地发现:缺失的标签越少,即观察到的标签越多,算法的恢复性能和预测性能越好,这也符合事实. 无论是对缺失数据的恢复还是对未知数据的预测,LEWL在绝大多数情况下都能获得比其他对比算法更好的结果. 这也说明LEWL使用策略的有效性,即在恢复缺失标签的同时又利用标签和标签以及特征和标签之间的关系学习到一个低维、实值的嵌入空间.

表3　缺失标签在Macro F_1上的恢复结果

Table 3　Recovery results for missing labels on Macro F_1

Data set	ρ	LEWL	MLML	LRML	CPLST	BR
Emotions	0.3	0.882±0.009	**0.905±0.006**	0.857±0.048	0.847±0.013	0.847±0.013
	0.7	0.775±0.010	0.801±0.009	**0.814±0.035**	0.666±0.011	0.666±0.011
Yeast	0.3	**0.883±0.007**	0.874±0.006	0.805±0.015	0.850±0.003	0.85±0.003
	0.7	**0.749±0.008**	0.738±0.007	0.717±0.023	0.660±0.002	0.66±0.002
CAL500	0.3	0.845±0.004	0.776±0.005	0.755±0.021	**0.849±0.008**	0.849±0.008
	0.7	**0.667±0.007**	0.633±0.007	0.643±0.020	0.657±0.012	0.657±0.012
Medical	0.3	**0.804±0.053**	0.783±0.029	0.755±0.005	0.796±0.049	0.796±0.049
	0.7	0.566±0.042	**0.593±0.029**	0.565±0.006	0.557±0.038	0.557±0.038
Langlog	0.3	**0.859±0.002**	0.856±0.002	0.855±0.041	0.85±0.001	0.85±0.001
	0.7	**0.685±0.003**	0.683±0.003	0.682±0.066	0.663±0.005	0.663±0.005
Enron	0.3	**0.845±0.013**	0.822±0.006	0.754±0.016	0.822±0.015	0.822±0.015
	0.7	0.633±0.033	0.648±0.028	0.63±0.013	**0.651±0.014**	0.651±0.014
Corel5k	0.3	**0.817±0.012**	0.785±0.007	0.801±0.025	0.813±0.017	0.813±0.017
	0.7	**0.619±0.011**	0.605±0.013	0.614±0.021	0.612±0.016	0.612±0.016

表4　缺失标签在排序损失上的恢复结果

Table 4　Recovery results for missing labels on Rank Loss

Data set	ρ	LEWL	MLML	LRML	CPLST	BR
Emotions	0.3	**0.132±0.010**	0.133±0.013	0.229±0.017	0.262±0.0121	0.262±0.012
	0.7	0.274±0.005	**0.271±0.011**	0.324±0.016	0.498±0.018	0.498±0.018
Yeast	0.3	**0.116±0.003**	0.138±0.002	0.209±0.008	0.261±0.003	0.261±0.003
	0.7	**0.227±0.004**	0.258±0.007	0.288±0.007	0.502±0.007	0.502±0.007
CAL500	0.3	**0.130±0.001**	0.239±0.006	0.249±0.011	0.259±0.0055	0.259±0.005
	0.7	**0.248±0.001**	0.394±0.011	0.339±0.009	0.503±0.006	0.503±0.006
Medical	0.3	**0.107±0.007**	0.254±0.011	0.155±0.008	0.257±0.022	0.257±0.022
	0.7	**0.148±0.022**	0.444±0.012	0.205±0.012	0.505±0.011	0.505±0.011
Langlog	0.3	**0.170±0.004**	0.233±0.004	0.204±0.051	0.251±0.002	0.251±0.002
	0.7	**0.332±0.012**	0.339±0.007	0.342±0.018	0.493±0.007	0.493±0.007
Enron	0.3	**0.129±0.008**	0.191±0.010	0.213±0.043	0.257±0.008	0.257±0.008
	0.7	**0.272±0.029**	0.363±0.021	0.291±0.043	0.513±0.011	0.513±0.011
Corel5k	0.3	**0.108±0.002**	0.271±0.008	0.178±0.008	0.259±0.003	0.259±0.003
	0.7	0.403±0.004	0.496±0.014	**0.356±0.007**	0.503±0.004	0.503±0.004

从表2至表4可以看到,这三个表只关注对缺失标签的恢复效果.LEWL,MLML和LRML分别用不同的方式来处理缺失标签,故在多个数据集上性能都优于BR和CPLST.这也从另一个侧面反映了处理缺失标签的必要性.本文建立的LEWL模型能够确保除了缺失标签外,其余可观察到的标签保持不变,这样就避免了在恢复的过程中误把正确的标签填补错误.

表5给出LEWL和和其余四种算法对缺失标签的恢复结果在不同标准上的t检验结果.取置信度为95%,$\alpha=0.05$.可以看出四种算法在平均精度和排序损失上置信水平值均小于0.05,说明

表5　不同算法关于恢复结果的t检验

Table 5　t-test of different algorithms on recovery results

	MLML	LRML	CPLST	BR
Average Precision	0.0455	0.0004	0.0001	0.0001
Macro F_1	0.2273	0.0179	0.0230	0.0230
Rank Loss	0.0015	0.0008	0.0003	0.0003

LEWL 和其余四种算法在平均精度和排序损失上均存在显著性差异,也进一步证明LEWL算法对缺失标签具有一定的恢复作用. 在Macro F_1标准上,MLML 的 p 值为 0.2273,大于 0.05,表明LEWL的恢复结果和MLML没有显著性差异.

从表6至表8可以看出,LEWL 在常用分类性能评价指标上均有提升. 首先,对于一个分类器来说,从一个密集的、低维实值的嵌入空间学习比从一个稀疏的、高维二值的空间学习更加准确高效. 其次,对于不平衡数据集,例如Corel5k数据集有300多个标签,而每个样本仅仅只有不超过10个正类标签,所以传统的BR分类器难以获得好的分类结果. MLML算法聚焦于恢复缺失标签,不关注分类过程,因此它的预测结果比BR好,但比新提出的LEWL算法略差. LRML和CPLST算法都属于标签嵌入法,两者考虑的是标签和特征之间的线性关系,对比实验结果也表明了本文所提算法采用HSIC的有效性和合理性.

表6 测试数据在平均精度上的预测结果

Table 6 Prediction results for test data on Average Precision

Data set	ρ	LEWL	MLML	LRML	CPLST	BR
Emotions	0.3	**0.732±0.026**	0.716±0.01999	0.729±0.0222	0.662±0.038	0.645±0.020
	0.7	**0.708±0.012**	0.687±0.039	0.682±0.016	0.558±0.023	0.568±0.021
Yeast	0.3	**0.708±0.006**	0.654±0.005	0.441±0.024	0.587±0.004	0.573±0.005
	0.7	**0.698±0.005**	0.648±0.008	0.44±0.026	0.448±0.008	0.450±0.003
CAL500	0.3	**0.365±0.017**	0.31±0.013	0.202±0.011	0.248±0.012	0.248±0.004
	0.7	**0.369±0.013**	0.307±0.013	0.197±0.009	0.170±0.008	0.175±0.009
Medical	0.3	**0.805±0.025**	0.734±0.009	0.775±0.008	0.675±0.049	0.289±0.029
	0.7	0.718±0.039	0.685±0.018	**0.721±0.038**	0.332±0.030	0.112±0.011
Langlog	0.3	**0.543±0.019**	0.419±0.016	0.410±0.017	0.461±0.048	0.469±0.019
	0.7	0.530±0.016	**0.531±0.019**	0.399±0.017	0.308±0.029	0.309±0.011
Enron	0.3	**0.536±0.041**	0.395±0.051	0.348±0.007	0.331±0.045	0.296±0.043
	0.7	**0.493±0.054**	0.346±0.055	0.252±0.011	0.213±0.020	0.197±0.025
Corel5k	0.3	**0.038±0.003**	0.027±0.008	0.027±0.005	0.030±0.007	0.027±0.007
	0.7	0.022±0.002	0.025±0.007	0.026±0.004	**0.027±0.007**	0.023±0.007

表7 测试数据在Macro F_1上的预测结果

Table 7 Prediction results for test data on Macro F_1

Data set	ρ	LEWL	MLML	LRML	CPLST	BR
Emotions	0.3	0.606±0.057	0.576±0.051	**0.616±0.036**	0.429±0.069	0.389±0.021
	0.7	0.596±0.038	0.523±0.061	**0.610±0.036**	0.152±0.033	0.160±0.041
Yeast	0.3	**0.461±0.004**	0.334±0.011	0.367±0.008	0.251±0.009	0.234±0.009
	0.7	**0.452±0.006**	0.313±0.008	0.366±0.009	0.058±0.009	0.062±0.004
CAL500	0.3	**0.234±0.012**	0.073±0.007	0.212±0.011	0.035±0.005	0.046±0.003
	0.7	**0.232±0.008**	0.066±0.008	0.205±0.010	0.008±0.002	0.015±0.002
Medical	0.3	**0.358±0.032**	0.311±0.012	0.325±0.002	0.310±0.023	0.098±0.018
	0.7	**0.306±0.020**	0.258±0.008	0.287±0.019	0.171±0.029	0.01±0.003
Langlog	0.3	**0.448±0.018**	0.164±0.011	0.389±0.028	0.171±0.025	0.185±0.014
	0.7	**0.434±0.017**	0.232±0.009	0.374±0.029	0.048±0.009	0.059±0.0120.
Enron	0.3	**0.166±0.011**	0.065±0.013	0.120±0.015	0.063±0.009	0.042±0.008
	0.7	**0.149±0.010**	0.051±0.014	0.101±0.016	0.022±0.004	0.013±0.011
Corel5k	0.3	**0.025±0.001**	0.018±0.005	0.020±0.003	0.010±0.003	0.009±0.003
	0.7	**0.019±0.001**	0.015±0.006	0.016±0.003	0.007±0.004	0.006±0.002

表8　测试数据在排序损失上的预测结果

Table 8　Prediction results for test data on Rank Loss

Data set	ρ	LEWL	MLML	LRML	CPLST	BR
Emotions	0.3	0.446±0.024	0.509±0.042	**0.441±0.019**	0.692±0.052	0.731±0.018
	0.7	0.457±0.030	0.561±0.071	**0.453±0.016**	0.924±0.023	0.915±0.013
Yeast	0.3	**0.385±0.008**	0.484±0.009	0.709±0.007	0.652±0.008	0.679±0.011
	0.7	**0.401±0.011**	0.496±0.013	0.722±0.005	0.940±0.011	0.933±0.011
CAL500	0.3	**0.461±0.011**	0.577±0.006	0.518±0.012	0.890±0.014	0.885±0.003
	0.7	**0.457±0.011**	0.584±0.018	0.528±0.008	0.972±0.008	0.969±0.005
Medical	0.3	**0.139±0.037**	0.476±0.014	0.302±0.019	0.345±0.050	0.785±0.027
	0.7	**0.168±0.005**	0.529±0.016	0.347±0.017	0.462±0.040	0.977±0.006
Langlog	0.3	**0.468±0.014**	0.716±0.017	0.629±0.057	0.716±0.053	0.689±0.025
	0.7	**0.487±0.015**	0.761±0.021	0.625±0.050	0.811±0.031	0.887±0.019
Enron	0.3	**0.315±0.038**	0.500±0.056	0.548±0.008	0.792±0.041	0.835±0.041
	0.7	**0.349±0.038**	0.569±0.055	0.771±0.019	0.948±0.015	0.967±0.023
Corel5k	0.3	**0.538±0.001**	0.745±0.002	0.697±0.002	0.825±0.001	0.889±0.001
	0.7	**0.549±0.001**	0.783±0.004	0.721±0.002	0.889±0.004	0.903±0.002

但LEWL在Emotions数据集上的效果不是很好,一方面是由于这个数据集只有六个标签,在压缩的过程中可能会损失有用信息;另一方面可能是由于优化过程陷入了局部最优.

表9给出了不同算法在预测结果上的t检验结果.表中所有算法的p值均小于0.05,说明LEWL算法性能较好.

表9　不同算法关于预测结果的t检验

Table 9　t-test of different algorithms on prediction results

	MLML	LRML	CPLST	BR
Average Precision	0.0046	0.0054	0.0000	0.0007
Macro F_1	0.0008	0.0007	0.0001	0.0000
Rank Loss	0.0007	0.0016	0.0001	0.0000

为了分析LEWL的收敛性,本文刻画了Emotions和Yeast数据集的目标函数值随着迭代次数的变化曲线,如图1所示.结果表明,目标函数值在经过较少次迭代后就趋于稳定.总体来说,该算法具有收敛速度快、迭代次数少的特点.对于其它数据集也有类似结果.

3.5　缺失标签的影响　为了探索缺失标签和嵌入维数的影响,在Emotions和Medical两个数据集上进行实验.如图2所示,ρ表示缺失标签率,在0.3到0.7之间变化;d表示嵌入空间维数.以

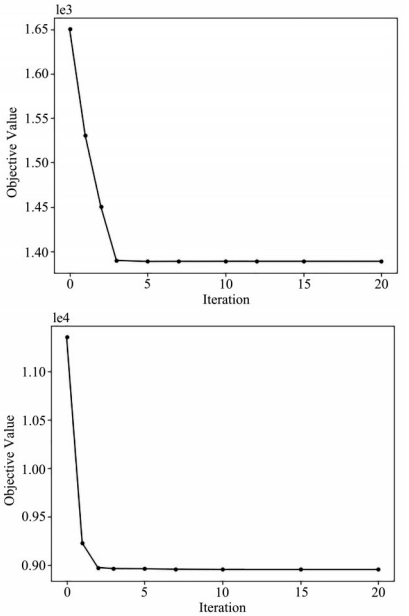

图1　LEWL在Emotions(上图)和Yeast(下图)数据集上的收敛性

Fig. 1　Convergence of LEWL on Emotions(up) and Yeast(down) datasets

Emotions数据集为例,当ρ从0.3变到0.7时,AP从0.732下降到0.708,降低了2.4%;而当嵌入维数从6变成2时,AP从0.734下降到0.603,降低了13.1%.可以发现,随着ρ不断变化,LEWL的性能变化不大,相对还是比较稳定的,这也暗示其具有鲁棒性.然而,当嵌入维数d特别小的时

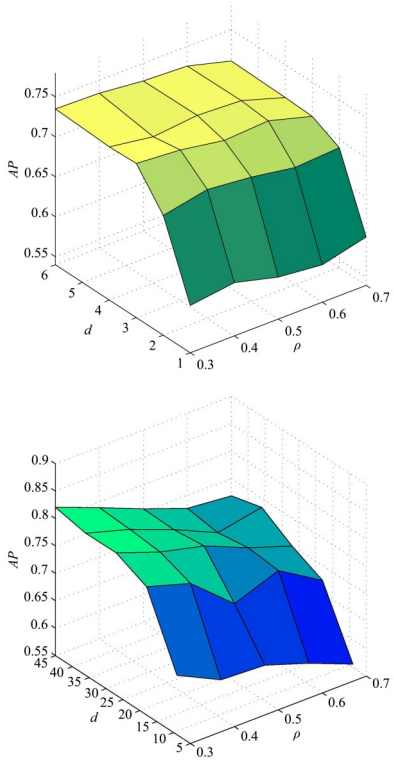

图 2　在 Emotions(上图)和 Medical(下图)数据集上的平均精度

Fig.2　Average Precision on Emotions(up) and Medical (down) datasets

候,这两个数据集的性能变差,这可能是由于嵌入空间维数太低,不能很好地提取全部标签的信息. 因此,对于每个数据集来说,选择合适的嵌入维数是至关重要的.

3.6　参数敏感度分析　在提出的LEWL算法中有两个参数:λ和β. 权衡参数λ控制特征嵌入和标签嵌入的权重比例,β是正则化参数. 为了研究λ和β的影响,分别选取$\lambda=0.001, 0.01, 0.1, 1, 10, 100$和$\beta=0, 0.001, 0.01, 0.1, 1, 10$六种情况,在 Enron数据集上进行实验,如图3所示.

从图3可以看出,当$\beta=0$时,该算法完全忽略样本的流形结构;当$0<\beta<10$时,曲线逐渐上升,说明样本平滑性假设对于缺失标签恢复具有一定的作用;$\beta\geqslant 10$时,则效果逐渐下降,这说明需要选取合适的β来使效果达到最佳. 对于λ来说,当$\lambda=0$时,效果较差,随着λ的不断增加,性能不断提升,证明了算法采用HSIC标准的有效性和合理性. 在其它数据集上也有类似的结果.

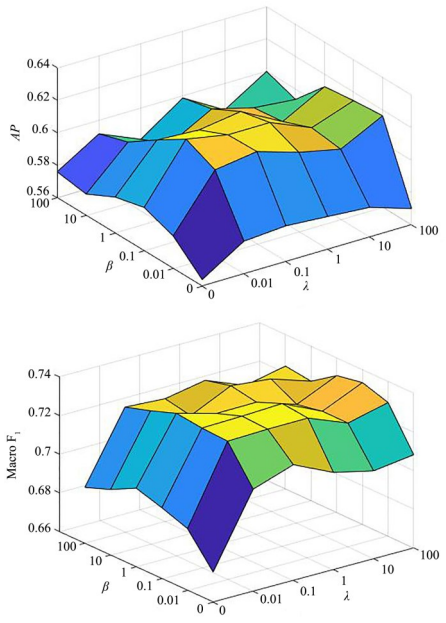

图 3　λ和β对 Emotions 数据集的影响

Fig.3　Varying λ and β on Emotions

4　结　论

针对高维弱标记问题,本文提出一种基于嵌入式的弱标记分类算法LEWL. 首先,将矩阵分解融入多标签算法中:一方面矩阵分解确实是一种有效的降维方法,通过利用标签之间的关系获得一个低维的嵌入空间;另一方面再结合样本的流形结构来恢复缺失的标签. 其次,为了增强嵌入空间的可预测性,它的学习过程应当与特征空间有较强的相关性. 本文采用HSIC技术来增强特征空间和嵌入空间的依赖性. 实验结果表明,本文算法对缺失标签的恢复具有良好的效果,而且获得的嵌入空间具有可预测性和可恢复性,同时有效地缓解了面对高维标签空间时的低效性. 但多标签学习领域还存在诸多挑战,例如准确地选取嵌入空间维数,高度类不平衡问题等,今后将会进一步考虑标签之间的相关性,使多标签的分类性能有更大的提升.

参考文献

[1] Katakis I, Tsoumakas G, Vlahavas I. Multilabel text classification for automated tag suggestion//

Proceedings of the ECML-PKDD/2008 Workshop on Discovery Challenge. Antwerp, Belgium: Springer, 2008,18:5.

[2] Jia X, Sun F M, Li H J, et al. Image multi-label annotation based on supervised nonnegative matrix factorization with new matching measurement. Neurocomputing,2017,219:518－525.

[3] Elisseeff A, Weston J. A kernel method for multi-labelled classification//Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver,Canada:MIT Press,2001:681－687.

[4] Boutell M R, Luo J B, Shen X P, et al. Learning multi-label scene classification. Pattern Recognition, 2004,37(9):1757－1771.

[5] Tsoumakas G, Vlahavas I. Random k-labelsets: An ensemble method for multilabel classification// European Conference on Machine Learning. Springer Berlin Heidelberg,2007:406－417.

[6] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification. Machine Learning,2011,85(3):333－359.

[7] Zhang M L, Zhou Z H. ML-KNN: a lazy learning approach to multi-label learning. Pattern Recognition,2007,40(7):2038－2048.

[8] Freund Y, Schapire R. A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence,1999,14(5):771－780.

[9] 马宏亮,万建武,王洪元.一种嵌入样本流形结构与标记相关性的多标记降维算法.南京大学学报(自然科学),2019,55(1):92－101. (Ma M L, Wan J W, Wang H Y. A multi-label dimensionality reduction algorithm embedded sample manifold structure and label correlation. Journal of Nanjing University (Natural Science),2019,55(1):92－101.)

[10] 彭成伦.多义性机器学习中的标记嵌入方法研究.硕士学位论文.南京:东南大学,2018. (Peng C L. Research on label embedding in ambiguous machine learning. Master Dissertation. Nanjing: Southeast University,2018.)

[11] Hsu D J, Kakade S M, Langford J, et al. Multi-label prediction via compressed sensing. 2009, arXiv: 0902.1284.

[12] Tai F, Lin H T. Multilabel classification with principal label space transformation. Neural Computation,2012,24(9):2508－2542.

[13] Chen Y N, Lin H T. Feature-aware label space dimension reduction for multi-label classification// Advances in Neural Information Processing Systems. Lake Tahoe, NV, USA: Neural Information Processing Systems Foundation, Inc., 2012, 2: 1529－1537.

[14] Lin Z J, Ding G G, Han J G, et al. End-to-end feature-aware label space encoding for multilabel classification with many classes. IEEE Transactions on Neural Networks and Learning Systems,2018,29 (6):2472－2487.

[15] 刘阳.多标签数据分类技术研究.博士学位论文.西安:西安电子科技大学,2018. (Liu Y. Research on Multi-label data classification technology. Ph. D. Dissertation. Xi′an:Xidian University,2018.)

[16] Sun Y Y, Zhang Y, Zhou Z H. Multi-label learning with weak label//Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta, GE, USA:AAAI Press,2010:593－598.

[17] Wu B Y, Liu Z L, Wang S F, et al. Multi-label learning with missing labels//2014 22nd International Conference on Pattern Recognition. Stockholm, Sweden:IEEE,2014:1964－1968.

[18] Guo B, Hou C, Shan J, et al. Low rank multi-label classification with missing labels//2018 24th International Conference on Pattern Recognition (ICPR2018). Beijing,China:IEEE,2018:417－422.

[19] Han Y F, Sun G L, Shen Y, et al. Multi-label Learning with Highly Incomplete Data via Collaborative Embedding//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, United Kingdom:ACM Press,2018:1494－1503.

[20] Xu M, Jin R, Zhou Z H. Speedup matrix completion with side information: application to multi-label learning//Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Montreal,Canada:MIT Press,2013:2301－2309.

[21] Xu L L, Wang Z, Shen Z F, et al. Learning low-rank label correlations for multi-label classification with missing labels//2014 IEEE International Conference on Data Mining. Shenzhen, China: IEEE, 2014: 1067－1072.

[22] Candès E J, Tao T. The power of convex relaxation: near-optimal matrix completion. IEEE Transactions on Information Theory, 2010, 56(5): 2053－2080.

[23] Wen Z W, Yin W T, Zhang Y. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. Mathematical Programming Computation, 2012, 4(4): 333－361.

[24] Zhang Y, Schneider J. Multi-label output codes using canonical correlation analysis// Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, FL, USA: JMLR, 2011: 873－882.

[25] Gretton A, Bousquet O, Smola A, et al. Measuring statistical dependence with Hilbert-Schmidt norms// International Conference on Algorithmic Learning Theory. Springer Berlin Heidelberg, 2005: 63－77.

[26] Han S J, Qubo C, Meng H. Parameter selection in SVM with RBF kernel function// World Automation Congress 2012. Puerto Vallarta, Mexico: IEEE, 2012: 1－4.

[27] Lin Z J, Ding G G, Hu M Q, et al. Multi-label classification via feature-aware implicit label space encoding// Proceedings of the 31st International Conference on International Conference on Machine Learning. Beijing, China: JMLR. org, 2014: 325－333.

[28] Han Y H, Wu F, Jia J Z, et al. Multi-task sparse discriminant analysis (MtSDA) with overlapping categories// Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlana, GA, USA: AAAI Press, 2010: 469－474.

[29] Pacharawongsakda E, Theeramunkong T. Towards more efficient multi-label classification using dependent and independent dual space reduction// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2012: 383－394.

[30] Zhang M L, Zhou Z H. A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(8): 1819－1837.

（责任编辑 杨可盛）