ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 3,May 2021,pp. 519－528
DOI:10. 16337/j. 1004-9037. 2021. 03. 010

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于分类间隔增强的不平衡多标签学习算法

程玉胜[1,2]，曹天成[2]

(1. 安徽省高校智能感知与计算重点实验室(安庆师范大学),安庆 246133;2. 安庆师范大学创新团队,安庆 246133)

摘　要：传统的多标签学习算法一般没有考虑标签的不均衡性,从而忽略了标签不平衡给分类带来的影响。但统计发现,目前常用的多标签数据集均存在标签不均衡问题,且少数类标签往往更加重要。基于此,本文提出了一种基于分类间隔增强的不平衡多标签学习算法(Imbalanced multi-label learning algorithm based on classification interval enhanced, MLCIE),旨在利用各标签分类间隔的重构来增强分类器对少数类标签样本的学习效率,提升样本标签质量,从而减少多标签不平衡对分类器学习精度的影响。首先利用各标签密度与条件熵计算各标签的不确定性系数;然后构建分类间隔增强矩阵,将各标签独有的密度信息融入到原始标签矩阵中,获取平衡的标签空间;最后使用极限学习机作为线性分类器进行分类。本文在11个多标签标准数据集上与其他7种多标签学习算法进行对比实验,结果表明本文算法在解决标签不平衡问题上有一定效果。

关键词：多标签学习;标签不平衡;分类间隔;标签密度;极限学习机

中图分类号：TP391　　**文献标志码**:A

Imbalanced Multi-label Learning Algorithm Based on Classification Interval Enhanced

CHENG Yusheng[1,2], CAO Tiancheng[2]

(1. University Key Laboratory of Intelligent Perception and Computing of Anhui Province(Anqing Normal University), Anqing 246133,China;2. Innovation Team of Anqing Normal University, Anqing 246133,China)

Abstract: Traditional multi-label learning algorithms generally do not consider the label imbalance, so the impact of label imbalance on classification is not ignored. However, statistics show that the current multi-label datasets have the problem of label imbalance, and a few kinds of labels are often more important. Based on this, this paper proposes an imbalanced multi-label learning algorithm based on classification interval enhanced (MLCIE), which aims to enhance the learning efficiency and improve the quality of the sample label by using the reconstruction of each label classification interval, so as to reduce the impact of multi-label imbalance on the learning accuracy of the classifier. Firstly, the uncertainty coefficient of each label is calculated by using the density and conditional entropy of each label; Then the enhancement matrix of classification interval is constructed, so that the unique density information of each label is integrated into the original label matrix to obtain the balanced label space; Finally, the limit learning machine is used as the linear classifier for classification. In this paper, the proposed algorithm is compared with other seven multi-label learning algorithms on the 11 multi-label standard datasets. The results show that the proposed

收稿日期:2020-06-22;**修订日期**:2020-11-06

algorithm can solve the problem of label imbalance.

Key words: multi-label learning; label imbalance; classification interval; label density; extreme learning machine

引　言

多标签学习[1]一直是机器学习领域的研究热点之一。不同于普通的单标签分类,在多标签学习任务中,每个样本关联着一个或多个标签,数据维度也高于单标签数据。对此,许多学者都提出了针对多标签数据的学习算法,例如:二级分类(Binary relevance, BR)算法[2]和多标签分类器集成链(Ensembles of classifier chains, ECC)算法[3]通过增加分类器个数或标签的种类来解决多标签问题;反向传播多标签学习(Back-propagation for multi-label learning, BP-MLL)算法[4]引入排序损失因素,减少了分类器迭代次数,但大幅地增加了计算复杂度;多标签K近邻(K-nearest neighbor for multi-label learning, ML-KNN)算法[5]利用最大后验概率(Maximum a posterior, MAP)算法[6]预测待测样本的标签集,时间复杂度较低,但是分类精度与K值选择关系过大,缺乏客观性。

在数据分类领域中,当某些类别的实例远高于或远低于其他实例时,会发生数据不平衡的情况。与平衡数据相比,大多数算法在处理不平衡数据时表现不佳,分类器的性能偏向多数类,从而在少数类的判别上会发生更高的错误率。但是在实际应用中,少数类实例往往更加重要,所以研究重点需要更加关注少数类实例是否正确分类,例如,在肿瘤分类领域,非肿瘤患者是多数类,肿瘤患者是少数类,而显然对少数类的判别需要更加精确。这样的问题同时存在于故障检测、信用卡诈骗等领域。多标签不平衡数据的处理方法主要是基于数据层面和算法层面,但传统的不平衡数据处理方法不完全适用于多标签数据,近年来,越来越多针对多标签不平衡问题的方法被提出。Liu等[7]利用实例的局部标签分布,对数据进行合成过采样,在兼具全局与局部不平衡的同时提高了分类器的分类精度;Tsai等[8]在处理临床记录文本分析时,将类别标签进行分层,再加入卷积模型中,不仅提高了识别性能,同时解决了类别不平衡问题;Peng等[9]利用了多数标签和少数标签之间的相关性,使少数类的训练依托于多数类,从而减少训练不足的问题。但是上述这些方法均没有考虑利用标签密度信息来改善不平衡情况。

在多标签数据中,样本的标签密度分布可以直观地反映出数据的不平衡情况,多数类和少数类都会出现极端的密度分布。而多标签数据中标签数量相对较多,使得标签空间中蕴含着大量的隐藏信息,其中就包含标签密度信息。隐藏信息可以通过数据挖掘方法提取出来,为了提高多标签分类性能,许多学者对数据集中标签空间进行了深度挖掘。Hsu等[10]将标签向量投影到随机的低维空间中,构成新的标签空间,并在该空间中拟合回归模型,然后将这些预测投影回原始标签空间;Cheng等[11]针对不完备数据集,利用标签相关性挖掘了未标记与已标记数据间关系,并构建了近邻标签补全矩阵,解决了标签缺失的问题。同时,标签增强[12]是一种利用标签空间中潜在信息将原本的逻辑标签更新为连续型标签的方法,通过不同的潜在信息挖掘方法,使新的标签空间具有相应的潜在信息,从而提高分类精度。Li等[13]利用标签传播依赖思想得到了标签间的相关性,并将相关性信息用于标签增强,使标签空间富含更多的相关性信息,提高了分类精度;Hou等[14]将特征空间的局部拓扑结构“迁移”到标签空间,使得增强后的标签空间具有特征空间局部样本相关信息,有效地减少错分现象;Xu等[15]基于图拉普拉斯矩阵进行标签增强,有效使用了特征空间中样本的相似性,约束了标签空间中标签间的相关性,较大地提升了分类效率。

在数据集中挖掘有关标签密度的信息用于标签增强,可以有效解决类不平衡而带来的错分类问

题。基于此,本文提出了一种基于分类间隔增强的不平衡多标签学习算法(Imbalanced multi-label learning algorithm based on classification interval enhanced, MLCIE)。该算法挖掘了样本标签空间潜在的标签密度信息,并利用标签密度信息增大了标签正负类的分类间隔,重构了标签空间将原有逻辑标签值转变为数值型标签值,而不同标签的分类间隔也会引导分类器更加"关注"少数类样本。本文算法首先利用了训练集标签密度与条件熵计算出4种不确定性系数,然后使用分类间隔增强矩阵的构建方法,获得了包含更多密度信息的标签矩阵,用其代替原始标签空间,最后使用极限学习机作为线性分类器进行分类。本文算法在11个数据集上与7种多标签学习算法进行了对比实验,实验结果验证了该算法的可行性、有效性和稳定性。实验结果表明:本文算法在绝大多数情况下可以取得更好的预测结果且稳定性更高。

1 知识背景

1.1 多标签学习

多标签学习是多标签数据而提出的一种学习框架,在这个学习框架之下,样本都是由特征和标签构成的,学习的目标是将未知标签的样本对应上更多正确的标签。定义 $\boldsymbol{X}=\mathrm{R}^m$ 表示 m 维样本空间,样本集合为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_q\}$,类别标签集合 $\boldsymbol{L}=\{l_1,l_2,l_3,\cdots,l_n\}$,给定多标签训练集 $\boldsymbol{T}=\{(\boldsymbol{x}_i,\boldsymbol{Y}_i)|i=1,2,3,\cdots,q\}$。在特征空间中,样本 $\boldsymbol{x}_i$ 用 m 维特征向量 $\boldsymbol{x}_i=[\boldsymbol{x}_i^1,\boldsymbol{x}_i^2,\boldsymbol{x}_i^3,\cdots,\boldsymbol{x}_i^m]$ 来表示,样本 $\boldsymbol{x}_i$ 对应与标签空间中的标签集合记为 $\boldsymbol{Y}_i=[\boldsymbol{y}_i^1,\boldsymbol{y}_i^2,\boldsymbol{y}_i^3,\cdots,\boldsymbol{y}_i^n]$,当 $\boldsymbol{x}_i$ 含有标签 l_a 时,$\boldsymbol{y}_i^a=+1$,否则 $\boldsymbol{y}_i^a=-1$。

1.2 标签密度和多标签不平衡

在真实世界中,多标签不平衡普遍存在于各数据集中。在同一数据集中,含有某种标签的实例数可能远大于不含有该标签的实例数,这些标签被称为多数标签;而含有某种标签的样本数可能远小于不含有该标签的样本数,这些标签被称为少数标签。这种关系可以用标签密度展现出来,图1给出了Yeast数据集各标签真负类密度对比。由图1可直观地发现,Yeast数据中14个标签都存在不同程度的不平衡问题,其中除12号、13号以外的12个标签负类个数均大于正类个数,9号、10号、11号这3个标签的正负类比例甚至约达到1:9,14号标签更是几乎没有正类样本,这4个标签就被称为少数数标签;而在12号和13号标签上,正类数大于负类数,且样本个数悬殊也很大,则这两个标签为多数标签。

标签密度反映了数据的不平衡情况,也为不平衡情况影响分类精度阐述了原因,少数标签因正类样本训练不充分,会出现正类误分为负类的情况;多数标签因负类样本训练不充分,会出现负类误分为正类的情况。

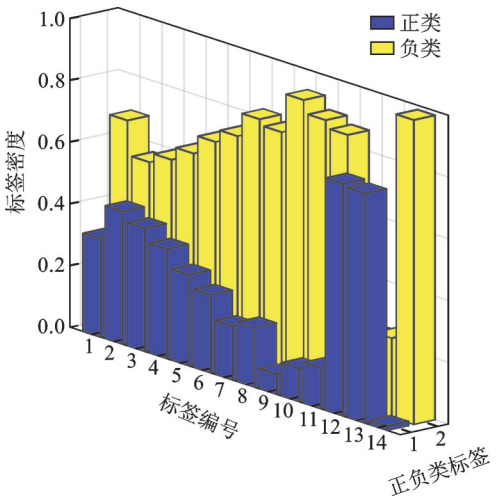

图1 Yeast数据集标签密度直方图

Fig.1 Density histogram of Yeast data set label

1.3 极限学习机

极限学习机(Extreme learning machine, ELM)[16]是一种单隐层前馈神经网络训练方法。此方法随机设置隐藏层权重和偏置,利用最小二乘的思想直接对输出层权重矩阵进行求解,只需要很少的训练时间,即可获得同等或更优的泛化性能。ELM求解单隐层前馈神经网络,可分为两个阶段:随机特征映射和线性参数求解。

设有 N 个随机样本的数据表示为 $\boldsymbol{D}=\left\{(\boldsymbol{X}_i,\boldsymbol{Y}_i)_{i=1}^N\right\}$,

$\boldsymbol{X}_i=\{\boldsymbol{x}_{i1},\boldsymbol{x}_{i2},\cdots,\boldsymbol{x}_{in}\}^{\mathrm{T}}$，$\boldsymbol{Y}_i=\{\boldsymbol{y}_{i1},\boldsymbol{y}_{i2},\cdots,\boldsymbol{y}_{im}\}^{\mathrm{T}}$，隐藏层神经元的个数为$L$，则单隐层前馈神经网络形式化定义为

$$f_L(\boldsymbol{X}_j)=\sum_{i=-1}^{L}\boldsymbol{\beta}_i g_i(\boldsymbol{X}_j) \tag{1}$$

式中：$\boldsymbol{\beta}_i=\{\beta_{i1},\beta_{i2},\cdots,\beta_{im}\}$表示第$i$层的输出权重，$g_i$表示第$i$个隐藏节点的输出，实质为激活函数，并可表现为

$$g_i(\boldsymbol{X}_j)=g(\boldsymbol{\omega}_i\cdot\boldsymbol{X}_j+b_i) \tag{2}$$

式中：$\boldsymbol{\omega}_i=\{\omega_{i1},\omega_{i2},\cdots,\omega_{im}\}$为输入权重，$b_i$表示第$i$个隐藏神经元的偏置。通常式(2)用来建模回归，对于分类问题可使用Sigmoid函数来限制输出值的范围，从而达到分类效果。

以上是ELM的随机特征映射阶段，对于第二阶段线性参数求解，通过最小化平方误差的近似误差来求解连接隐藏层和输出层的权值$\boldsymbol{\beta}$，可表示为

$$\min_{\beta}\ \left\|\boldsymbol{H}\boldsymbol{\beta}-\boldsymbol{Y}\right\|^2 \tag{3}$$

式中$\boldsymbol{H}$为隐藏层输出矩阵，即

$$\boldsymbol{H}=\begin{bmatrix}\boldsymbol{h}(\boldsymbol{x}_1)\\ \boldsymbol{h}(\boldsymbol{x}_2)\\ \vdots\\ \boldsymbol{h}(\boldsymbol{x}_N)\end{bmatrix}=\begin{bmatrix}h_1(x_1) & h_2(x_1) & \cdots & h_L(x_1)\\ h_1(x_2) & h_2(x_2) & \cdots & h_L(x_2)\\ \vdots & \vdots & \vdots & \vdots\\ h_1(x_N) & h_1(x_N) & \cdots & h_L(x_N)\end{bmatrix} \tag{4}$$

$\boldsymbol{Y}$为训练标签矩阵

$$\boldsymbol{Y}=\begin{bmatrix}\boldsymbol{y}_1^{\mathrm{T}}\\ \boldsymbol{y}_2^{\mathrm{T}}\\ \vdots\\ \boldsymbol{y}_N^{\mathrm{T}}\end{bmatrix}=\begin{bmatrix}y_{11} & y_{12} & \cdots & y_{1m}\\ y_{21} & y_{22} & \cdots & y_{2m}\\ \vdots & \vdots & \vdots & \vdots\\ y_{N1} & y_{N2} & \cdots & y_{Nm}\end{bmatrix} \tag{5}$$

通过式(3~4)，最小二乘解为

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{\dagger}\boldsymbol{Y} \tag{6}$$

式中$\boldsymbol{H}^{\dagger}$为$\boldsymbol{H}$的Moore-Penrose广义逆矩阵，表示为

$$\text{s.t.}\quad \boldsymbol{H}^{\dagger}=\begin{cases}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}} & \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\text{非奇异}\\ \boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})^{-1} & \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\text{非奇异}\end{cases} \tag{7}$$

最终求出的$\hat{\boldsymbol{\beta}}$即可预测位子标签，表示为

$$\hat{\boldsymbol{Y}}=\boldsymbol{H}\hat{\boldsymbol{\beta}} \tag{8}$$

2 MLCIE算法

2.1 基于标签密度构建分类间隔增强矩阵

对于标签分布不平衡的问题，本文通过标签密度这一先验知识，将标签密度信息添加到原标签空间，改造原标签空间的二元标签向量，使得原标签+1/−1变为连续型数值标签，构建出密度标签矩阵。这种转换可以使得分类器更加容易分类各标签，提高分类精度。对于多标签数据，N个示例的标记空间$\boldsymbol{Y}=\{\boldsymbol{Y}_1,\boldsymbol{Y}_2,\boldsymbol{Y}_3,\cdots,\boldsymbol{Y}_N\}$，$\boldsymbol{Y}_i=[\,y_i^1,y_i^2,y_i^3,\cdots,y_i^m]$，则标签密度可表示为

$$P^{+}(j)=\sum_{i=1}^{N}(y_i^j=1)/N \tag{9}$$

$$P^{-}(j)=\sum_{i=1}^{N}(y_i^j=-1)/N \tag{10}$$

式中:P_i^+为第i个标签的正类密度,P_i^-为第i个标签的负类密度,理论上P_i^+与P_i^-的和为1。本文通过计算条件熵来评价各标签分类正确或是错误带来的信息量大小。设条件熵共分为4种:已知标签为正时,通过分类器计算得到标签预测为正或负的条件熵;已知标签为负时,通过分类器计算得到标签预测为正或负的条件熵,可得出

$$\mathrm{HYY}=-p((P^+ +s),(P^+ +s))\log_2 p((P^+ +s)|(P^+ +s)) \tag{11}$$

$$\mathrm{HYN}=-p((P^+ +s),(P^- +s))\log_2 p((P^- +s)|(P^+ +s)) \tag{12}$$

$$\mathrm{HNY}=-p((P^- +s),(P^+ +s))\log_2 p((P^+ +s)|(P^- +s)) \tag{13}$$

$$\mathrm{HNN}=-p((P^- +s),(P^- +s))\log_2 p((P^- +s)|(P^- +s)) \tag{14}$$

式中:HYY代表已知第i个标签为正时,通过分类器计算预测出标签为正所带来的信息量大小;HYN代表已知第i个标签为正时,通过分类器计算预测出标签为负所带来的信息量大小;HNY代表已知第i个标签为负时,通过分类器计算预测出标签为正所带来的信息量大小;HNN代表已知第i个标签为负时,通过分类器计算预测出标签为负所带来的信息量大小。在实验数据集中,会出现某类标签密度为零的情况,这种情况会导致条件熵值无法计算,所以这里条件熵计算时引入了一个数值极小的平滑参数s,目的是在尽可能不改变结果大小的情况下,消除标签密度为0而带来的无法计算情况,一般设在$[1\times10^{-3},1\times10^{-5}]$内。已知条件熵越大,则随机变量不确定性也越大,所以将这4种条件熵称为4种不确定性系数。具体如表1所示。

表1 4种不确定性系数

Table 1 Four kinds of uncertainty coefficient

不确定性系数	已知原标签	预测标签
HYY	+1	+1
HYN	+1	−1
HNY	−1	+1
HNN	−1	−1

这4种条件熵的含义可以这样理解:现已知某样本的第i个标签为+1时,通过分类器预测出标签为+1的不确定性为HYY,通过分类器预测出标签为−1的不确定性为HNY。不确定性系数越大,说明预测的标签置信度越小,错误分类的代价就越大。这些错误是与每个标签的不平衡程度相关,若标签相对平衡,则不确定性系数就会小,相应地所求得的标签置信度也大,所以将这4种不确定性系数代入原训练集的标签矩阵中,就可以在训练分类器的过程中加入标签密度的信息,从而减小因不平衡所带来的错分类现象。基于此,本文构建了分类间隔增强矩阵$\boldsymbol{C}$,其元素为

$$\boldsymbol{C}_{ij}=\begin{cases}\alpha-(\mathrm{HYY}_j+\mathrm{HYN}_j) & y_{ij}=+1\\ \alpha-(\mathrm{HNY}_j+\mathrm{HNN}_j) & y_{ij}=-1\end{cases} \tag{15}$$

式中:$i\in\{1,2,3,\cdots,N\}$为样本编号,$j\in\{1,2,3,\cdots,M\}$为标签编号,α为平衡化参数,实验中取值范围为$[1,5]$。

2.2 标签空间平衡化重构

分类间隔增强矩阵$\boldsymbol{C}$包含了样本标签的密度信息,将其融入标签空间后可以使原标签空间含有标签密度信息,增大了正负标签的分类间隔。原始标签越不平衡,新的标签分类间隔将会越大,从而大幅降低分类时不平衡而带来的错分类现象。新的标签矩阵$\hat{\boldsymbol{Y}}$为

$$\hat{\boldsymbol{Y}}=\boldsymbol{Y}\times\boldsymbol{C} \tag{16}$$

为了更直观地体现实际效果,表2是本文在Yeast数据集中任取10个样本,列出了前3个标签改造前和改造后的标签值变化,平衡化参数α取值为2。

由表2可以看出，原始标签都是简单的+1或−1，以0为分类面，分类间隔为2，而改造后的标签空间中，分类面为0，分类间隔却不同，若标签密度越不平衡，分类间隔将会越大，并且在数值上会偏向于少数类。重构后的标签空间，标签密度在数值上趋于平衡，用于分类使用可以消除原有数据标签密度不平衡带来的影响，提高分类精度。

表2 改造前后标签变化

Table 2 Labels change before and after remoulding

序号	原始标签			改造后标签		
	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
1	−1	−1	+1	−1.384 2	−1.437 3	+1.613 1
2	−1	−1	−1	−1.384 2	−1.437 4	−1.435 9
3	−1	+1	+1	−1.384 2	+1.575 9	+1.613 1
4	−1	−1	+1	−1.384 2	−1.437 4	+1.613 1
5	+1	+1	−1	+1.713 0	+1.575 9	−1.435 9
6	−1	−1	+1	−1.384 2	−1.437 4	+1.613 1
7	−1	+1	+1	−1.384 2	+1.575 9	+1.613 1
8	−1	+1	+1	−1.384 2	+1.575 9	+1.613 1
9	−1	−1	−1	−1.384 2	−1.437 4	−1.435 9
10	−1	+1	+1	−1.384 2	+1.575 9	+1.613 1

2.3 基于分类间隔增强的不平衡多标签学习算法

根据多标签学习的目标，同时结合ELM学习模型，本文将ELM的随机映射函数$h(\boldsymbol{x})$将样本特征$\boldsymbol{x}_i$从输入空间映射到L维的特征空间，多标签极限学习机的输出函数$f_L(\boldsymbol{x})$为

$$f_L(\boldsymbol{x})=\boldsymbol{H}\boldsymbol{\beta}=\begin{bmatrix}h(\boldsymbol{x}_1)\\h(\boldsymbol{x}_2)\\\vdots\\h(\boldsymbol{x}_N)\end{bmatrix}_{N\times L}\begin{bmatrix}\boldsymbol{\beta}_1^{\mathrm{T}}\\\boldsymbol{\beta}_2^{\mathrm{T}}\\\vdots\\\boldsymbol{\beta}_3^{\mathrm{T}}\end{bmatrix} \tag{17}$$

根据式(8,17)求解目标函数式(8)，可得输出权重

$$\boldsymbol{\beta}=H^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{U}+\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{Y} \tag{18}$$

式中：U为正则系数，$\boldsymbol{I}$为L维单位矩阵。再结合由式(16)构造出的新标签空间，此时多标签输出函数就可表示为

$$f_L(\boldsymbol{x})=\boldsymbol{H}\boldsymbol{\beta}=\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{U}+\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\hat{\boldsymbol{Y}} \tag{19}$$

在传统的ELM中，权重和偏置是随机产生的，所以算法的输出很不稳定。本文采用RBF核函数来解决这个问题，即对$\boldsymbol{H}$使用核函数进行一次映射，操作如下

$$\Omega_{\mathrm{ELM}}=\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}:\Omega_{\mathrm{ELM}(i,j)}=h(\boldsymbol{x}_i,\boldsymbol{x}_j),K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\gamma\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2\right) \tag{20}$$

由式(18)可得：$\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}=\begin{bmatrix}K(\boldsymbol{x},\boldsymbol{x}_1)\\K(\boldsymbol{x},\boldsymbol{x}_2)\\\vdots\\K(\boldsymbol{x},\boldsymbol{x}_N)\end{bmatrix}^{\mathrm{T}}$，这样式(19)就可以表示为

$$f_L(\boldsymbol{x})=h(\boldsymbol{x})\boldsymbol{H}^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{C}+\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{Y}=\begin{bmatrix}K(\boldsymbol{x},\boldsymbol{x}_1)\\K(\boldsymbol{x},\boldsymbol{x}_2)\\\vdots\\K(\boldsymbol{x},\boldsymbol{x}_N)\end{bmatrix}^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{U}+\Omega_{\mathrm{ELM}}\right)^{-1}\hat{\boldsymbol{Y}} \tag{21}$$

本文通过拟合回归实验得到权重$\boldsymbol{\beta}$。在已知重构后标签矩阵$\hat{\boldsymbol{Y}}$和拟合得出权重$\boldsymbol{\beta}$的情况下，通过式(21)即可求得测试集预测标记集合$f_L(\boldsymbol{x})$，最小化目标函数为

$$f_E = \sum_{i=1}^{N} \left\| f_l(\boldsymbol{x}_i - \hat{\boldsymbol{Y}}_i) \right\|^2 \tag{22}$$

3 实验与结果分析

3.1 实验环境与评价指标

实验代码均在Matlab2016a中运行,硬件环境为Intel®Core$_{(TM)}$ i5-7500M 3.40 GHz CPU,8 GB内存,操作系统为Windows 10。本文实验的评价指标选用5个常用多标签学习评价指标来综合评价该算法性能,分别是:平均精度(Average precision, AP)、覆盖率(Coverage, CV)、汉明损失(Hamming loss, HL)、1-错误率(One-error, OE)和排序损失(Ranking loss, RL)[17]。为方便,简写为AP↑、CV↓、HL↓、OE↓和RL↓,其中↑表示数值越高越好,↓表示数值越低越好。

3.2 实验数据集

本文实验使用的11个多标签标准化数据集,均为雅虎网页数据集,其中涵盖了文本、音乐和图像等多个领域,详细信息如表3所示。

表3 多标签数据集详细描述

Table 3 Detailed description of multi-label datasets

数据集	训练集	测试集	标签数	特征数	数据类型
Art	2 000	3 000	26	462	艺术文本
Business	2 000	3 000	30	438	商业新闻
Computer	2 000	3 000	33	681	计算机文本
Education	2 000	3 000	33	550	教育资讯
Entertainment	2 000	3 000	21	640	娱乐新闻
Health	2 000	3 000	32	612	医药文本
Recreation	2 000	3 000	22	606	娱乐新闻
Reference	2 000	3 000	33	793	说明书文本
Science	2 000	3 000	40	743	科技文本
Society	2 000	3 000	27	636	社会科学
Social	2 000	3 000	39	1 047	社会新闻

3.3 对比算法与参数设置

本文选择了7个多标签学习算法作为对比算法,分别为:基于K近邻思想的多标签懒惰学习算法(Improred multi-label lazy learning approach, IMLLA)[18],基于二阶策略的多标签算法RankSVM(Rank support vector machine),基于核极限学习机(Kernl extreme learning machine, KELM)的多标签学习算法ML-KELM(Multi-label KEML)[19],基于径向基函数的多标签学习算法ML-RBF(Radial basis function neural network based multi-label learning algorithm)[20],基于近邻标记空间的非平衡化标签补全算法NeLC-NLS(Non-equilibrium labels completion of neighboring labels space)[21],基于类属属性思想的多标签学习算法LITF(Multi-label learing with label specific features)[22]。在本文算法MLCIE中,平滑参数$s=0.001$,平衡化参数$\alpha=2$,正则系数$U=1$,核函数选用RBF,核参数设为1。在ML-KNN中近邻个数$k=15$,平滑参数$s=1$。在IMLLA中近邻个数$k=15$。在RankSVM中代价参数设为1,核函数选用RBF。在ML-KELM中正则化系数设为1,核函数选用RBF,核参数设在[1,100]内。在ML-RBF中核参数设在[1,100]内。在LIFT中平滑参数设为0.1。考虑到对比算法的可行性和准确性,并减少误差的产生,各对比算法在每一个实验数据集都进行了10次实验。通过实验发现ML-RBF和LITF结果不稳定,所以选用10次结果得到的5种评价指标求平均数和标准差代表实验结果,其他算法结果稳定,10次实验结果相同。

3.4 实验结果及分析

本文在11个雅虎网页数据集上将本文算法与7个对比算法进行对比试验,这些数据集的特征数量从438~1 047不等,每个数据集包含2 000个训练集和3 000个测试集,属于高维度数据集。由于篇幅原因,本文只给出了AP指标实验结果,如表4所示,另4种指标结果类似,并给出了总体平均排序如表5所示。数据后有排位顺序,最优结果以粗体表示,并且每种算法在所有数据集上的平均排位列在最后

表4 各算法在11个数据集上的AP↑值

Table 4 AP↑ values of each algorithm on 11 data sets

数据集	MLCIE	ML-KNN	IMLLA	RankSVM	ML-KELM	NeLC-NLS	ML-RBF	LIFT
Art	0.624 9(1)	0.569 7(7)	0.488 5(8)	0.570 5(6)	0.614 3(3)	0.624 6(2)	0.606 6±0.001 8(4)	0.606 0±0.002 8(5)
Business	0.889 6(1)	0.882 2(4)	0.871 7(8)	0.871 8(7)	0.882 9(3)	0.886 8(2)	0.880 5±0.001 2(6)	0.882 0±0.001 4(5)
Computer	0.711 2(1)	0.662 0(6)	0.615 3(8)	0.615 7(7)	0.700 9(3)	0.707 2(2)	0.696 0±0.002 6(5)	0.697 9±0.002 7(4)
Education	0.647 4(2)	0.608 2(6)	0.528 5(8)	0.553 0(7)	0.637 9(3)	0.648 0(1)	0.621 1±0.002 2(5)	0.632 9±0.003 0(4)
Entertainment	0.692 1(1)	0.621 8(7)	0.549 7(8)	0.642 1(6)	0.684 7(4)	0.692 0(2)	0.679 4±0.003 0(5)	0.686 4±0.003 0(3)
Health	0.793 4(1)	0.756 2(6)	0.669 8(8)	0.674 5(7)	0.782 4(4)	0.791 4(2)	0.781 2±0.001 2(5)	0.782 6±0.001 7(3)
Recreation	0.634 1(1)	0.561 3(7)	0.412 7(8)	0.570 6(6)	0.628 7(3)	0.632 8(2)	0.620 1±0.002 6(5)	0.621 8±0.002 7(4)
Reference	0.717 3(1)	0.682 0(6)	0.585 9(8)	0.624 8(7)	0.708 7(4)	0.717 2(2)	0.709 6±0.002 1(3)	0.701 4±0.003 2(5)
Science	0.596 4(3)	0.548 9(6)	0.445 2(8)	0.503 5(7)	0.596 8(2)	0.603 2(1)	0.583 0±0.002 0(5)	0.590 0±0.002 1(4)
Social	0.774 3(1)	0.755 5(6)	0.691 4(7)	0.689 9(8)	0.771 9(3)	0.772 2(2)	0.764 4±0.001 5(5)	0.768 6±0.002 8(4)
Society	0.645 8(1)	0.618 1(6)	0.585 8(8)	0.593 0(7)	0.636 4(3)	0.643 8(2)	0.627 4±0.001 2(5)	0.634 2±0.002 9(4)
平均排序	1.27	6.09	7.91	6.82	3.18	1.82	4.36	4.09

表5 各算法的平均排序

Table 5 Average sorting of algorithms

评价指标	MLCIE	ML-KNN	IMLLA	RankSVM	ML-KELM	NeLC-NLS	ML-RBF	LIFT
AP	1.27	6.09	7.91	6.82	3.18	1.82	4.36	4.09
OE	2.27	6.00	7.36	7.55	3.18	1.55	3.45	4.64
RL	1.00	3.59	7.81	4.55	5.45	2.64	7.09	3.86
CV	1.00	3.27	7.27	4.27	5.73	3.82	7.73	2.91
HL	1.18	5.82	7.09	7.91	2.95	5.09	2.27	3.68
总体	1.344	4.954	7.488	6.220	4.098	2.984	4.980	3.836

1行,排位越小算法性能越优。

由表4可看到对于AP指标,MLCIE算法在Education数据集上略低于NeLC-NLS算法,排位第二,在Science数据集上排位第三,在其他数据集上均排位第一,平均排序为1.27,是8个算法中最优的。由表5可以看到,MLCIE算法AP、RL、CV、HL指标的平均排序都领先于其他算法,只是在OE指标上略逊与NeLC-NLS算法。综上实验结果表明,MLCIE算法整体性能比其他对比算法更加优异,尤其是在CV、RL两个指标上,这体现了分类间隔增强矩阵的有效性,也说明了MLCIE算法在具有更好分类精度的同时,兼具了良好的泛化能力。

3.5 算法性能分析

为了评价MLCIE算法在各数据集上的综合性能,本文选用显著性水平$\alpha=0.05$下的Nemenyi检验评估该算法与其他对比算法在雅虎网页数据集上的结果是否真实有效。当两个对比算法在所有数据集上平均排序的差值大于临界差值(Critical difference, CD),则认为这两个算法存在显著性差异,否则无显著性差异。

图2给出了在不同评价指标下每种算法之间的性能对比。图中对于没有显著性差异的算法用实线相连,各评价指标从左至右,算法性能依此降低。

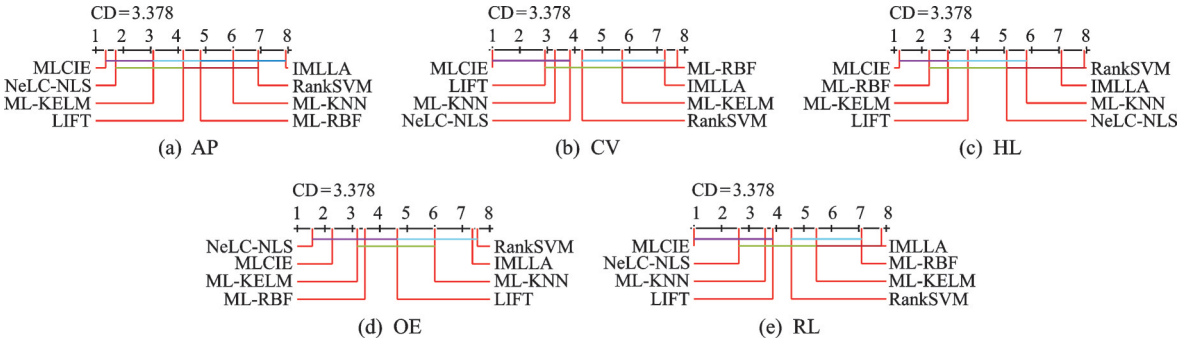

图2 各算法性能对比

Fig.2 Performance comparison of each algorithm

对于每种算法,都有35种实验对比结果(7种对比算法,5种评价指标),结合图2得到结论:在74%的情况下,MLCIE算法与其他算法有显著性差异,并且性能在97%的情况下占优。在AP指标上MLCIE算法与NeLC-NLS算法没有显著性差异;在CV指标上MLCIE算法与LIFT算法和ML-KNN算法没有显著性差异;在HL指标上MLCIE算法与ML-RBF算法没有显著性差异;在OE指标上MLCIE算法与NeLC-NLS算法、ML_KELM算法、ML-RBF算法没有显著性差异;在RL指标上MLCIE算法与NeLC-NLS算法和ML-KNN算法没有显著性差异。在5种评价指标的性能上,MLCIE算法仅在OE指标上低于NeLC-NLS算法,在其他指标上均为最优算法。

从上述统计假设检验分析可知,MLCIE算法性能最优,与其他对比算法显著性差异明显,进一步说明了MLCIE算法的有效性以及分类间隔增强矩阵的合理性。

4 结束语

在多标签学习中,为减少标签不平衡带来的影响,本文将标签密度信息融入到标签空间中,以改善分类器的分类性能。为此,本文引入了分类间隔增强矩阵,并提出了分类间隔增强的不平衡多标签学习算法。该算法有效地提升了标签空间的质量,让每个标签都包含了密度信息。实验结果表明,MLCIE算法优于一些常见的多标签学习算法。本文算法设计的重点在于验证标签不均衡对分类器精度的影响,因此在分类间隔增强矩阵的构建过程中将每个标签单独考虑计算,而未考虑标签之间的相关性。下一步的研究重点将兼顾标签的相关性信息,进一步丰富分类间隔增强矩阵的构建方法,提升分类器的性能。

参考文献:

[1] ZHOU Z H, ZHANG M L. Multi-label learning[C]//Proceedings of Encyclopedia of Machine Learning and Data Mining. US: Springer, 2017.

[2] SCHOLKOPF B, PLATT J, HOFMANN T. Multi-instance multi-label learning with application to scene classification[C]//Proceedings of International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2006:1609-1616.

[3] READ J, PFAHRINGER B, HOLMES G, et al. Classifier chains for multi-label classification[J]. Machine Learning, 2011, 85(3):333-359.

[4] ZHANG M L, ZHOU Z H. Multilabel neural networks with applications to functional genomics and text categorization[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10):1338-1351.

[5] ZHANG M L, ZHOU Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.

[6] ZHANG L, FU X, LI H, et al. An improved maximum a posterior-based estimation method coping with capture effect for RFID tags identification[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2020, 34(5):63-67.

[7] LIU B, TSOUMAKAS G. Dealing with class imbalance in classifier chains via random undersampling[J]. Knowledge-Based Systems, 2020, 192: 105292.1-105292.13.

[8] TSAI S C, CHENG T Y, CHEN Y N. Leveraging hierarchical category knowledge for data-imbalanced multi-label diagnostic text understanding[C]//Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019). Stroudsburg, PA: ACL, 2019: 39-43.

[9] PENG Y, HUANG E, CHEN G, et al. A general framework for multi-label learning towards class correlations and class imbalance[J]. Intelligent Data Analysis, Cambridge, MA: MIT Press, 2019, 23(2): 371-383.

[10] HSU D J, KAKADE S M, LANGFORD J, et al. Multi-label prediction via compressed sensing[C]//Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2009: 772-780.

[11] CHENG Y S, ZHAO D W, ZHAN W F, et al. Multi-label learning of non-equilibrium labels completion with mean shift[J]. Neurocomputing, 2018, 321:92-102.

[12] 耿新, 徐宁, 邵瑞枫. 面向标记分布学习的标记增强[J]. 计算机研究与发展, 2017, 54(6):1171-1184.
GENG Xin, XU Ning, SHAO Ruifeng. Label enhancement for label distribution learning[J]. Computer Research and Development, 2017, 54(6):1171-1184.

[13] LI Y K, ZHANG M L, GENG X. Leveraging implicit relative labeling-importance information for effective multi-label learning [C]//Proceedings of 2015 IEEE International Conference on Data Mining (ICDM). Los Alamitos, CA: IEEE Computer Society, 2015: 251-260.

[14] HOU P, GENG X, ZHANG M L. Multi-label manifold learning[C]//Proceedings of Thirtieth AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2016:1680-1686.

[15] XU N, LIU Y P, GENG X. Label enhancement for label distribution learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(4): 1632-1643.

[16] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: Theory and applications[J]. Neurocomputing, 2006, 70(1/2/3):489-501.

[17] 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.

[18] 张敏灵. 一种新型多标记懒惰学习算法[J]. 计算机研究与发展, 2012, 49(11):2271-2282.
ZHANG Minling. A new multi label lazy learning algorithm[J]. Computer Research and Development, 2012, 49(11): 2271-2282.

[19] LUO F, GUO W, YU Y, et al. A multi-label classification algorithm based on kernel extreme learning machine[J]. Neurocomputing, 2017, 260: 313-320.

[20] ZHANG M L. Ml-RBF: RBF neural networks for multi-label learning[J]. Neural Processing Letters, 2009, 29(2):61-74.

[21] 程玉胜, 赵大卫, 钱坤. 近邻标签空间非平衡化标签补全的多标签学习[J]. 模式识别与人工智能, 2018, 31(8):740-749.
CHENG Yusheng, ZHAO Dawei, QAIN Kun. Multi-label learning based on unbalanced label completion in nearest neighbor label space[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(8):740-749.

[22] ZHANG M L, WU L. Lift: Multi-label learning with label-specific features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(1): 107-120.

作者简介:

程玉胜(1969-),**通信作者**,男,教授,硕士生导师,研究方向:数据挖掘、粗糙集和机器学习等,E-mail:chengyshaq@163.com。

曹天成(1995-),男,硕士研究生,研究方向:数据挖掘、统计等,E-mail:641560829@qq.com。

(编辑:刘彦东)