

类不平衡的公共和标签特定特征多标签分类

张海翔¹, 李培培², 胡学钢²

(1. 蚌埠医学院附属合肥市第二人民医院 讯息处, 安徽 合肥 230012;
2. 合肥工业大学 大数据知识工程教育部重点实验室, 安徽 合肥 230601)

摘要:多标签分类主要解决实例数据对应多个标签问题, 现有多标签方法大多利用所有特征组成的相同数据表示来区分所有标签, 由于每个标签自身特点不同, 统一的特征不能完全区分标签, 给模型训练带来负面作用和时间成本增加, 如何利用对每个标签而言最具有辨别力的特征来提高模型分类性能成为一种难题, 此外现实中类不平衡问题同样会导致多标签学习模型的性能下降。基于此, 提出一种类不平衡的公共和标签特定特征多标签分类方法。首先, 找到种子实例的最近邻居, 然后通过插值技术得到合成实例的特征来解决类不平衡问题; 其次, 为了找出对每个标签最具代表性的特征, 引入 l_1 , $l_{2,1}$ 正则化约束系数矩阵提取标签的特定特征和公共特征; 最后, 使用标签相关性实现关联标签的模型输出相似, 实例相关性保证关联特征共享对应标签分布信息提高分类性能。实验表明所提方法与其他多标签分类方法相比获得了更好的分类精度。

关键词:多标签分类; 类不平衡; 公共特征; 标签特定特征; 标签相关性

中图分类号: TP183

文献标识码: A

文章编号: 1673-629X(2024)02-0046-07

doi: 10.3969/j.issn.1673-629X.2024.02.007

Class Imbalance Multi-label Classification with Common and Label Specific Features

ZHANG Hai-xiang¹, LI Pei-pei², HU Xue-gang²

(1. Information Division, The Second People's Hospital of Hefei Affiliated to Bengbu Medical College,
Hefei 230012, China;

2. Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology),
Ministry of Education, Hefei 230601, China)

Abstract: Multi-label classification mainly deals with the problem that instances data is associated with multiple class labels. Most of the existing multi-label methods use the same data representation consisting of all features to distinguish all labels. However, due to the different characteristics of each label, unified features cannot fully differentiate them, which brings negative effects and increases time cost to model training. Therefore, it becomes a challenge to improve the model classification performance by utilizing the most discriminative features for each label. In addition, the problem of class imbalance in reality can also result in a decline in the performance of multi-label learning models. Motivated by this, we propose a new approach of class imbalance multi-label classification with common and label specific features. Firstly, we find the nearest neighbors of seed instances, and then use interpolation techniques to obtain the features of synthetic instances to solve the problem of class imbalance. Secondly, in order to find the most representative features for each label, we introduce l_1 -norm and $l_{2,1}$ -norm regularizers constraint coefficient matrix to extract label-specific features and common features. Finally, we use label correlation to achieve similar model output of associated labels, and instance correlation to ensure that associated features share corresponding label distribution information to improve classification performance. Extensive experiments show a competitive performance of proposed method against other multi-label learning approaches.

Key words: multi-label classification; class-imbalance; common features; label-specific features; label correlation

收稿日期: 2023-03-22

修回日期: 2023-07-26

基金项目: 国家自然科学基金资助项目(61976077, 62076085, 62120106008); 蚌埠医学院科技计划项目(2022byzd225sk)

作者简介: 张海翔(1996-), 男, 工程师, 硕士, 通信作者, 研究方向为数据挖掘与人工智能; 李培培(1982-), 女, 副研究员, 硕导, 博士, CCF 会员(18097M), 研究方向为数据挖掘与人工智能; 胡学钢(1961-), 男, 教授, 博导, 博士, CCF 会员(13977S), 研究方向为数据挖掘与人工智能。

0 引言

多标签分类^[1-2]指利用一组已标记数据训练出模型对未标记的样本进行分类。现实中对事物的描述通常用多个标签进行描述,如视频注解、文本分类和生物信息学。常见处理方法分为:问题转化和算法适应。前者将多标签任务转换为一个或多个单标签分类任务,后者则将传统机器学习算法直接处理多标签数据。同样深度学习近年来在多标签医疗领域应用广泛,如:深度 CNN^[3]对 26 种心脏异常进行多标签分类,X 光图像分类有 ConvNeXt 网络^[4]与 BioBert 编码的语义向量相结合,EfficientNetB4 架构^[5]与转移学习方法进行结合用以提高胸部 X 光图像分类准确性,以及判别核卷积网络(DKNet)^[6]用于眼科疾病智能识别。

大多数多标签分类模型都面临类不平衡问题^[7],尤其当负类实例的数量远大于正类实例的数量,导致分类器偏向于负类实例,分类性能降低。不平衡问题可分为:标签内部的不平衡、标签间的不平衡和标签集之间的不平衡。在标签内部不平衡中每个标签通常包含极高数量的负样本和极少量的正样本^[8]。标签间不平衡考虑数据集中单个标签的频率,其中一个标签(正类)的数量可能高于另一个标签的正类数量^[9]。标签集^[10]的稀疏频率,如果考虑到完整标签集每个类别的正样本与负样本比例可能与常见标签集相关联,由于标签稀疏性,通常存在较多的频繁标签集和唯一标签集。这也意味着一些标签集被认为是大多数,而其余标签集同时被认为是少数情况。

现实数据同样面临数据维度爆炸问题,对模型训练将消耗过多资源。研究人员使用特征降维的技术从原始数据集筛选出对全部标签具有代表意义的特征子集,称为公共特征,如方法 SCMFS^[11]使用耦合矩阵分解技术找出特征与标签矩阵间公共部分。而实际中每个标签在特征空间都有对其最相关的特征称为标签特定特征,如方法 IMLSF^[12]。该方法分为两种:特征转换和 l_1 范数,如 LIFT^[13]通过特征转换将标签正负实例转换为单标签特定特征,但这种方法无法识别出哪些特征是某标签的特定特征。以上方法只考虑公共特征或标签特定特征的优势,未将两种优势同时考虑。CLML^[14]综合两种方法的优势,提出基于公共特征与标签特定特征方法,通过 $l_1, l_{2,1}$ 范数限定系数矩阵选出每个标签的特定特征与公共特征,但该方法未能适应类不平衡数据环境。

因而,该文提出类不平衡的公共和标签特定特征多标签分类方法,采用启发式重采样技术解决类不平衡问题,然后综合标签公共特征和标签特定特征的优势进行数据筛选,不仅找出对所有标签都有意义的公共特征集合,还为每一个标签找出最具代表意义的特

定特征。最后采用标签相关性实现关联标签的相似模型输出,实例相关性保证关联特征共享对应标签分布信息,提高了多标签分类精度。

主要贡献如下:

(1)所提方法考虑少数标签列表,使用这些标签作为种子出现的实例来生成新实例,解决类不平衡问题。

(2)为降低训练过程带来的资源消耗,利用 $l_1, l_{2,1}$ 范数限定模型系数矩阵,结合每一个标签自身特点找出其对应的特定特征和多个标签的公共特征。

(3)为提高分类精度,假设相似标签之间具有相似输出,相关实例可共享对应标签分布,来约束模型的系数。

1 相关工作

类不平衡是多标签分类过程面临的难题之一,样本与对应标签并非分布在同一数据空间中。多标签分类只采用问题转换或算法适应策略不能很好地解决该问题。针对不平衡问题,已有方法可分为四类:重采样、分类器自适应、集成方法和代价敏感方法。重采样方法^[15]对数据集的预处理产生新的平衡多标签数据,独立于分类器组。基于 LP 变换^[16]的重采样方法(LP-RUS)将多标签数据集转换为一个多类数据集,每个不同的标签组合(标签集)作为一个类处理。但基于 LP 的重采样在解决不平衡问题时受到数据集中标签稀疏性的限制。分类器自适应通过改进现有机器学习方法适应。Luo 等^[17]提出基于非对称分阶段损失函数,动态调整正样本和负样本的损失代价方法解决不平衡问题。集成方法将几个基本模型结合起来产生最佳预测模型,如 BR-IRUS^[18]。代价敏感方法使用不同成本度量来描述任何特定错误分类样本的成本,旨在使总成本最小化。如 SOSHF^[19]通过代价敏感聚类将多标签学习任务转换为不平衡的单标签分类类型。

在多标签分类过程中学习公共特征和标签的特定特征能有效提高计算效率和分类性能。公共特征方法指通过某种方法从原始特征空间中提取对分类过程有意义的特征子集。Zhu 等^[20]面对缺失标签空间引入流形正则化将特征相似样本在补全标签空间中也接近一致,构建模型时补全标签矩阵,结合实例相关性约束模型系数。MIFS^[21]为降低缺失标签在标签相关性中的不利因素,将原标签空间分解至低维,在低维空间进行公共特征筛选。以上方法在模型构建过程中只选择了被所有标签共享的公共特征,而现实中每一个标签都应该在特征空间中有与之对应的特定特征。

其中特征转化的标签特定特征方法有 LIFT, LIFTAce^[22]和 LSDM^[23],LIFTAce 利用聚类技术结合

标签相关性假设满足相关关系的标签共享聚类结果。LSDM 通过调整比例参数,针对单个标签的正负实例聚类重建特定特征空间。IMLSF 使用加速近端梯度方法,以迭代的方式快速有效地求解目标函数,找出每个标签对应的具体特征。基于 l_1 范数的特定特征方法如 LLSF^[24] 假设标签与特征子集关联,运用线性回归模型区分出对标签具有代表意义的特征。LSFCI^[25] 通过概率邻域图模型计算实例相关性,在学习标签特定特征时同时考虑实例、标签相关性。

2 类不平衡的公共和标签特定特征多标签分类

本节给出多标签分类问题定义: U 是实例集, L 是标签集, X 为 d 维的输入空间集, Y 是有 l 个标签的决策属性集。输入数据矩阵 $X = [x_1, x_2, \dots, x_n]^T \in R^{n \times d}$, $x_i = [x_{i1}, x_{i2}, \dots, x_{id}]$, $Y = [y_1, y_2, \dots, y_n]^T \in R^{n \times l}$, $y_i = [y_{i1}, y_{i2}, \dots, y_{il}]$ 。多标签分类是从训练集中学习一种映射 $f(\cdot): X \rightarrow Y$, 然后对测试数据预测标签。

2.1 类不平衡多标签分类方法

为解决类不平衡问题, 该文利用重采样策略 MLSMOTE 单独处理每个少数标签出现的实例集, 每个少样本都将是新的合成样本, 新实例由特征子集与合成标签集构成, 合成标签集指在参考样本及其邻居出现的标签都在合成实例中。该过程主要分为三步: 第一步选择一个少数实例作为参考点。在标签空间中计算类不平衡比率得到当前标签的不平衡程度。IRLb1 表示单个标签不平衡的程度, MeanIR 表示所有标签的 IRLb1 的均值, 见式 1 和式 2:

$$\text{IRLb1}(i) = \frac{\max\left\{\sum_{j=1}^U h(l_i, y_j)\right\}}{\sum_{j=1}^U h(l_i, y_j)} \quad (1)$$

$$h(l_i, y_j) = \begin{cases} 1, & l_i \in y_j \\ 0, & l_i \notin y_j \end{cases}$$

$$\text{MeanIR} = \frac{1}{|L|} \sum_{i=1}^U \text{IRLb1}(i) \quad (2)$$

其中, l_i 为 L 的第 i 个标签, $1 \leq i \leq |L|$, y_j 为 x_i 对应的标签集。若 $\text{IRLb1}(i) \geq \text{MeanIR}$ 表示该标签比其他标签在标签空间中更加稀疏, 将其放入少数类中得到少数类实例包, 反之放入多数类。第二步少数实例筛选完成后选择一个与其最近邻居的集合。集合大小由参数 K 确定, 并从邻居集合中随机选择一个实例 refNeigh 作为参考。第三步特征集和标签集的产生, 对每一个样本和其参考实例 refNeigh 合成实例特征值将沿着连接这两个样本的线进行插值。新实例的标签

集, 从邻域相关性计算参考样本及其邻居中每个标签的出现次数, 包括在合成标签集中出现一半或更多实例中的标签。合成的新实例样本最终被添加到数据集中, 对每一个少数类实例包样本和剩余标签做以上步骤处理。为达到更好的平衡效果, 每个标签的 IRLb1 在新标签开始时需重新评估其平衡率, 如果一个少数标签在处理过程中达到了 MeanIR 值, 将被排除在合成样本生成过程中。

2.2 公共特征和标签特定特征学习

该文选择线性回归模型分类器, 通过投影矩阵 W 关联特征与标签空间, 为提取标签特定特征, 采用 l_1 范数方法把投影矩阵中元素稀疏性和参数缩小, 同时引入 $l_{2,1}$ 范数提取公共特征。以上过程可表示为式 3:

$$\min_W \frac{1}{2} \|XW - Y\|_F^2 + \beta \|W\|_1 + \lambda_2 \|W\|_{2,1} \quad (3)$$

其中, $W = [w_1, w_2, \dots, w_l] \in R^{d \times l}$ 为回归模型系数矩阵, 且 W 中第 j 个标签系数向量为 $w_j = [w_{1j}, w_{2j}, \dots, w_{dj}]^T$, w_{ij} 表示第 i 个特征对第 j 个标签辨别程度, $w_{ij} \neq 0$ 表示特征与标签存在辨别程度, 该特征是第 j 个标签的特定特征。 β 和 λ_2 分别控制系统矩阵稀疏性和公共特征、标签特定特征数量。此外学习过程中经常引入标签相关性提高分类性能。但如果简单认为标签间存在相关, 某一标签的特定特征对另一相关标签而言也是特定特征, 相应的系数向量也相似, 此假设不成立^[14]。因而, 该文假设标签相关对应输出模型 XW 也相似, 且相关性越高, 相似度越接近, 使用正则项表示为式 4:

$$\sum_{i,j} S_{ij} \|Xw_i - Xw_j\| = \text{tr}(XWL_1(XW)^T) \quad (4)$$

其中, S_{ij} 表示标签 i 与标签 j 的相关性, L_1 是标签相关性矩阵的拉普拉斯矩阵, 因而式 3 可转化为式 5:

$$\min_W \frac{1}{2} \|XW - Y\|_F^2 + \frac{\alpha}{2} \text{tr}(XWL_1(XW)^T) + \beta \|W\|_1 + \lambda_2 \|W\|_{2,1} \quad (5)$$

借鉴方法 LSFCI 引入实例相关性增强标签特定特征选择的结果, 两实例之间存在强相关性, 对应标签空间也存在相关性。该文也引入此技术并在 K 个实例邻居之间计算相似度 C , 使用正则项表示为式 6:

$$\min_W \sum_{i,j} C_{ij} \|x_i W - x_j W\| = \text{tr}((XW)^T L_2 (XW)) \quad (6)$$

其中, C_{ij} 表示实例 i 与实例 j 的相关性矩阵, L_2 是实例相关性矩阵的拉普拉斯矩阵。最终目标函数表示为:

$$\min_W \frac{1}{2} \|XW - Y\|_F^2 + \frac{\alpha}{2} \text{tr}(XWL_1(XW)^T) +$$

$$\frac{\lambda_1}{2} \text{tr}((XW)^T L_2(XW)) + \beta \|W\|_1 + \lambda_2 \|W\|_{2,1} \quad (7)$$

其中, $\alpha, \beta, \lambda_1, \lambda_2$ 为常数参数。

2.3 分类模型处理与优化

式7存在范数 $l_1, l_{2,1}$ 需进一步优化该凸函数。首先放缩 $\|W\|_{2,1}$ 为 $\text{tr}(W^T A W)$, A 为对角矩阵, 且 $A_{ii} = \frac{1}{2 \|w_i\|_2}$, 对于 l_1 范数正则化使用加速近端梯度法求解。 $f(W)$ 表示为:

$$f(W) = \frac{1}{2} \|XW - Y\|_F^2 + \frac{\alpha}{2} \text{tr}(XW L_1(XW)^T) + \frac{\lambda_1}{2} \text{tr}((XW)^T L_2(XW)) + \lambda_2 \|W\|_{2,1} \quad (8)$$

且 $f(W)$ 记为 $\|\nabla f(W_1) - \nabla f(W_2)\|_F^2 \leq L_f \|\Delta W\|$, L_f 为利普希茨常数, $\Delta W = W_1 - W_2$, $g(W) = \beta \|W\|_1$ 。

由近端梯度算法 $F(W)$ 近似优化表示为:

$$Q(W, W^{(t)}) = f(W^{(t)}) + \langle \nabla f(W^{(t)}), W - W^{(t)} \rangle + \frac{L_f}{2} \|W - W^{(t)}\|_F^2 + g(W) \quad (9)$$

通过定义 $G^{(t)} = W^{(t)} - \frac{1}{L_f} \nabla f(W^{(t)})$, 优化过程进一步表示:

$$W = \arg \min_W Q(W, W^{(t)}) = \arg \min_W \frac{1}{2} \|W - G^{(t)}\|_F^2 + \frac{\beta}{L_f} \|W\|_1 \quad (10)$$

收敛过程迭代至 t 次, 当满足 $\psi_{i+1}^2 - \psi_i^2 \leq \psi_i^2$ 时, 收敛速度加快, 每次迭代优化系数矩阵为:

$$W_{i+1} = S_{L_f}^{L_f}[G^{(t)}] = \arg \min \frac{\beta}{L_f} \|W\|_1 + \frac{1}{2} \|W - G^{(t)}\|_F^2 \quad (11)$$

当给定系数矩阵 $W_1, W_2, \Delta W = W_1 - W_2$, 有:

$$\begin{aligned} \|\nabla f(W_1) - \nabla f(W_2)\|_F^2 &= \|X^T X \Delta W + \alpha X^T X \Delta W L_1 + \lambda_1 X^T L_2 X \Delta W + \lambda_2 A \Delta W\|_F^2 \\ &\leq 4 \|X^T X \Delta W\|_F^2 + 4 \|\alpha X^T X \Delta W L_1\|_F^2 + 4 \|\lambda_1 X^T L_2 X \Delta W\|_F^2 + 4 \|\lambda_2 A \Delta W\|_F^2 \\ &\leq 4 (\|X^T X\|_2^2 + \|\lambda_1 X^T L_2 X\|_2^2 + \|\lambda_2 A\|_2^2 + \|\alpha X^T X\|_2^2 \cdot \|L_1\|_2^2) \|\Delta W\|_F^2 \end{aligned} \quad (12)$$

利普希茨常数 L_f 为:

$$L_f = [4(\sigma_{\max}^2(X^T X) + \sigma_{\max}^2(\lambda_2 A) +$$

$$\sigma_{\max}^2(\lambda_1 X^T L_2 X) + \sigma_{\max}^2(\alpha X^T X) \cdot \sigma_{\max}^2(L_1)]^{1/2} \quad (13)$$

其中, $\sigma_{\max}^2(\cdot)$ 为计算最大奇异值函数。

3 实验及其结果分析

3.1 数据集与对比算法

为验证对比所提方法是否取得明显优势, 在多个数据集上进行实验。表1给出了实验数据集信息, 包含数据量、特征数、标签数、MaxIR 和 MeanIR, 其中 MaxIR 和 MeanIR 分别代表最大和平均类别不平衡比。

表1 数据集

| Datasets | Instances | Features | Labels | MaxIR | MeanIR |
|----------------|-----------|----------|--------|-------|--------|
| Cal500 | 502 | 68 | 174 | 88.8 | 20.58 |
| Medical | 978 | 1 449 | 45 | 226 | 89.5 |
| Enron | 1 702 | 1 001 | 53 | 913 | 73.95 |
| Science | 5 000 | 743 | 40 | 293.5 | 48 |
| Rcv1 (subset1) | 6 000 | 944 | 101 | 691 | 54.49 |
| Corel5K | 5 000 | 499 | 374 | 1 120 | 189.57 |
| Arts | 5 000 | 462 | 26 | 1 233 | 69.45 |
| Education | 5 000 | 550 | 33 | 1 576 | 179.59 |
| Genbase | 645 | 1 186 | 27 | 171 | 37.31 |

将所提方法与常见多标签分类方法进行比较, 包括: LLSF, LLSF-DL, LIFT, MLCIB, LSFCI, LSFMLL, JLCLS, CLML。其中 LLSF 假设任意强相关的两个类标签可以对应特征, 利用线性回归方法学习每个类标签的标签特定特征。LLSF-DL^[26] 在 LLSF 基础上引入高阶的标签相关性。LIFT 通过聚类分析手段在正负实例中学习标签特定特征, 参数 γ 设置为 0.1。MLCIB^[27] 通过学习标签正则化, 把原始标签空间映射到新空间中处理缺失标签和类不平衡问题, 参数 α, β, γ 在 $[0, 1]$ 调整, 步长为 0.1。LSFCI 借助标签、实例的相关性学习标签特定特征, 参数 α, γ 值在 $[2^{-10}, 2^{10}]$ 范围变化, 步长为 2, 参数 η 值在 $[2^{-12}, 2^{12}]$ 范围变化且步长为 2, 阈值 τ 设置为 0.5。LSFMLL^[28] 在模型训练过程中结合标签特定特征和相关性内容。JLCLS^[29] 利用标签关系型补全缺失标签矩阵, 扩展原始标签矩阵内容。所有对比算法相应的参数均按照文献中的建议进行设置, 所有的参数值均为其在各个数据集上的最优解。

3.2 评价指标与实验结果分析

本节主要分析所提方法与对比算法间的实验对比, 实验数据评价指标使用 Average Precision, Ranking Loss, Micro F1, Macro F1, F1, Hamming Loss。各指标含义如下: 给定多标签测试数据集 $T = \{(x_i, Y_i)\}_{i=1}^n$,

$Y_i \in \{0,1\}^l$ 表示真实数据集, 测试样本 x_i , $h(x_i)$ 为第 i 个样本的预测结果, $f(x_i, y)$ 为 x_i 属于 y 的置信度。Average Precision 评估相关标签排名高于特定标签 $y \in y_i$ 比例的平均分数:

$$\text{AvePrec} = \frac{1}{n} \sum_{i=1}^n \frac{1}{|Y_i^l|} \sum_{y_j \in Y_i^l} \frac{|\{y_q \in Y_i^l : R_i(y_q) \leq R_i(y_j)\}|}{R_i(y_j)} \quad (14)$$

其中, $R_i(y_j)$ 是样本 x_i 的类标签 y_j 预测等级。

Ranking Loss 描述样本的标签对被反向排序的平均比例:

$$\text{RankingLoss} = \frac{1}{n} \sum_{i=1}^n \frac{|\{(y_a, y_b) : R_i(y_a) > R_i(y_b), (y_a, y_b) \in Y_i^l \times Y_i^l\}|}{|Y_i^l| |Y_i^l|} \quad (15)$$

Micro F1 将标签向量的每个条目视为单独实例, 不考虑标签的区别:

$$\text{Micro F1} = \frac{2 \sum_{j=1}^l \sum_{i=1}^n Y_{ij} h(x_{ij})}{\sum_{j=1}^l \sum_{i=1}^n Y_{ij} + \sum_{j=1}^l \sum_{i=1}^n h(x_{ij})} \quad (16)$$

Macro F1 为各标签的精度和召回率的综合:

$$\text{Macro F1} = \frac{1}{l} \sum_{i=1}^l \frac{2p_i^l r_i^l}{p_i^l + r_i^l} \quad (17)$$

其中, p_i^l , r_i^l 分别为第 i 个标签的精度和召回率。

F1 为每个样本的精度和召回率的综合:

$$\text{F1} = \frac{1}{n} \sum_{i=1}^n \frac{2p_i^s r_i^s}{p_i^s + r_i^s} \quad (18)$$

其中, p_i^s , r_i^s 分别为第 i 个样本的精度和召回率。

Hamming Loss 评估样本对应标签分类结果错误的频率, 包括标签预测错误或漏预测:

$$\text{hLoss}(h) = \frac{1}{n} \sum_{i=1}^n \frac{1}{l} |h(x_i) \Delta Y_i| \quad (19)$$

Δ 表示两组之间的对称差。

为分析算法之间显著性差异使用 Friedman 检验, 表 2 给出这些指标 Friedman 统计量 F_F 。为验证所提方法是否取得优异性能, 将所提方法作为控制算法, 比较任意一对算法间平均排名差与临界差 $CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$, 显著性水平 $\alpha = 0.05$ 时对于 Nemenyi 检测 $q_\alpha = 3.164$, $CD = 1.0614$ ($k = 9, N = 9$), 各指标 CD 图见图 1。如果所提方法与某对比算法平均排名相差一个或一个以上 CD 值且未用线连接, 认为性能间存在显著差异。

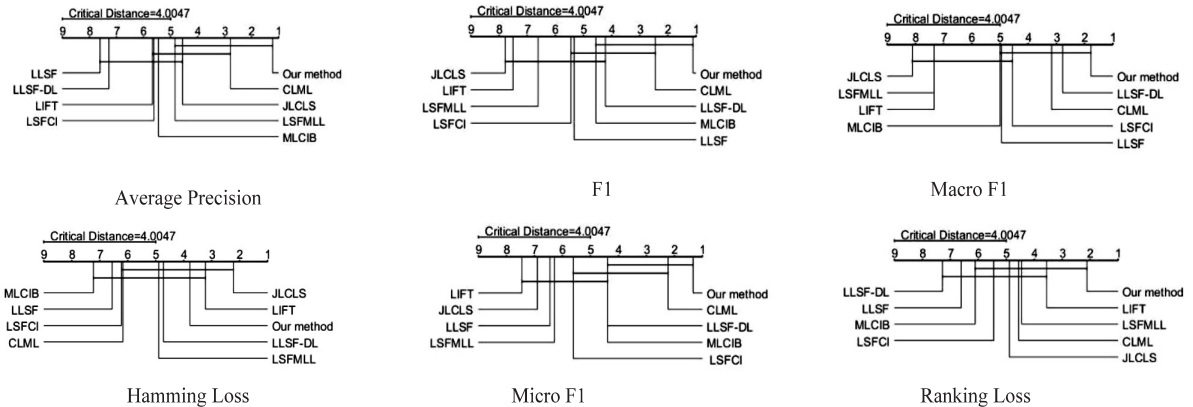


图 1 所提方法与对比算法的 Nemenyi 检验比较结果

根据图 1 可知, 在实验指标下所提方法优于对比算法。Hamming Loss 指标与标签相关性无明显关系, 除 LIFT 是一阶以外, 其余对比算法均为二阶或高阶, 实验结果也显示此指标下 LIFT 算法相对最优, 所提方法非最优但在 Hamming Loss 指标对比其他算法无显著差异。在其余实验指标下所提方法在实验精度上明显高于 LLSF, LSFCI, CLML, 体现了类不平衡处理和公共特征与标签特定特征的有效性。因为在构建模型过程中根据标签信息学习标签特定特征, 类不平衡问题将影响标签特定特征的比重, 预测过程倾向大类信息。而使用重采样策略单独处理每个少数标签出现的实例集, 每个少数样本视为新的合成样本, 可以有效平

衡标签特定特征的比重。

表 2 在 0.05 显著性水平条件下 F_F 每种评估方法的临界值

| Metric | F_F | Critical value($\alpha = 0.05$) |
|-------------------|----------|-----------------------------------|
| Average Precision | 8.729 4 | 1.955 |
| Hamming Loss | 11.13 | |
| F1 | 60.392 7 | |
| Ranking Loss | 8.800 0 | |
| Macro F1 | 32.571 1 | |
| Micro F1 | 35.700 2 | |

此外, 在其他单指标上 (如 Micro F1 和 Macro

F1), 所提方法显著优于 LIFT, LLSF, LSFMLL。原因在于类不平衡情况下对标签特定特征的选择非最优, 且这些方法忽略公共特征带来的优势。CLML 指标排名均靠前, 原因在于其他方法的假设条件并非总是成立, 该文假设相似标签之间具有相似输出, 相关实例可共享对应标签分布。通过约束模型的系数可有效解决此类问题, 且 CLML 和所提方法均考虑标签特定特征和公共特征, 引入实例相关性和标签相关性, 取得了较好的实验结果。

在类不平衡的特殊环境下, 不平衡的标签空间给特征筛选过程带来误导, 降低了分类精度。例如在 MeanIR 平均不平衡度较高的 Corel5k, Medical 和 Education 数据集上, 可以看出所提方法均优于 CLML, 在低 MeanIR 值如: 数据集 Cal500, Genbase 类不平衡手段未取得明显优势。虽然在 Nemenyi 排名上

所提方法相比 CLML 没有取得显著优势, 但所有排名均靠前。

为了验证类不平衡处理第二步少数实例筛选完成后选择一个与其最近邻居的集合, 集合大小由参数 K 对实验结果的影响, 该文在 4 个代表数据集 Medical, Enron, Science, Education 上设置不同参数 K , 大小在区间 $[3, 4, \dots, 10]$, 步长为 1 上调整。图 2 为 4 个数据集上 6 个评价指标下的实验结果折线图, 由图可知实验结果随 K 的变化而变化, 其中 K 值取 3, 4, 9, 10 时效果最差, 此时筛选出来的集合大小结果要么信息缺失要么信息冗余, 取值为 5 时在 Education, Medical, Science 数据集上各实验指标结果最优, 取 8 时实验结果在 Enron 数据集上最优, 且在其他数据集上实验结果为次优。实验结果与文献[30]给定建议参数 K 取值为 5 吻合。

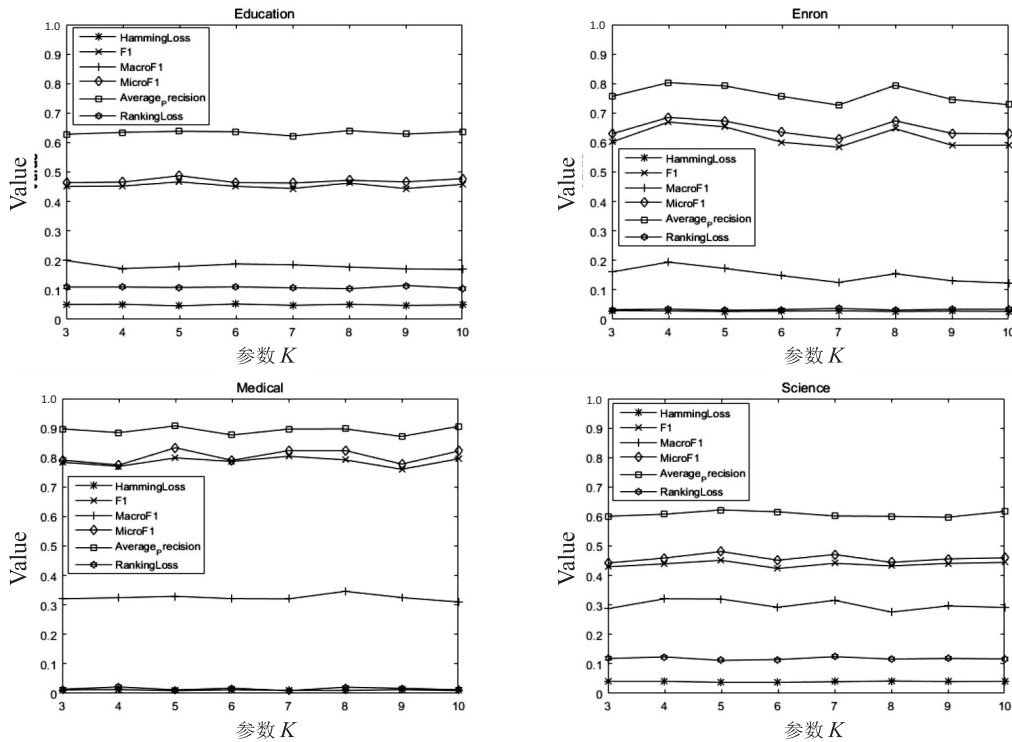


图2 在4个代表数据集上调整 K 所得到的实验结果

4 结束语

该文提出类不平衡的公共和标签特定特征多标签分类方法, 在类不平衡环境中考虑实例间的相关性及公共特征问题。利用重采样策略, 通过找到种子实例的最近邻居结合插值技术得到合成实例的特征。然后综合标签公共特征和标签特定特征的优势进行数据筛选, 不仅找出对所有标签都有意义的公共特征集合, 还为每一个标签找出最具代表意义的特定特征。最后采用标签相关性实现关联标签的相似模型输出, 实例相关性保证关联特征共享对应标签分布信息, 提高分类

精度。实验结果表明, 通过算法对比, 所提方法在精度上取得明显优势。未来将针对缺失特征与缺失标签的不平衡多标签数据展开研究。

参考文献:

- [1] 邹一章. 面向多标签分类问题的特征选择方法研究[D]. 合肥: 合肥工业大学, 2022.
- [2] 武红鑫, 韩 萌, 陈志强, 等. 监督和半监督学习下的多标签分类综述[J]. 计算机科学, 2022, 49(8): 12-25.
- [3] WICKRAMASINGHE N L, ATHIF M. Multi-label classification of reduced-lead ECGs using an interpretable deep convolutional neural network[J]. Physiological Measure-

- ment, 2022, 43(6):064002.
- [4] JIN Y, LU H, ZHU W, et al. Deep learning based classification of multi-label chest X-ray images via dual-weighted metric loss[J]. *Computers in Biology and Medicine*, 2023, 157:106683.
 - [5] DEVASIA J, GOSWAMI H, LAKSHMINARAYANAN S, et al. Deep learning classification of active tuberculosis lung zones wise manifestations using chest X-rays; a multi label approach[J]. *Scientific Reports*, 2023, 13(1):887.
 - [6] BHATI A, GOUR N, KHANNA P, et al. Discriminative kernel convolution network for multi-label ophthalmic disease detection on imbalanced fundus image dataset[J]. *Computers in Biology and Medicine*, 2023, 153:106519.
 - [7] 程玉胜, 曹天成. 基于分类间隔增强的不平衡多标签学习算法[J]. *数据采集与处理*, 2021, 36(3):519–528.
 - [8] 廖晨. 基于 GAN 的多类不平衡数据过采样方法研究[D]. 桂林: 桂林理工大学, 2022.
 - [9] 李昂, 韩萌, 穆栋梁, 等. 多类不平衡数据分类方法综述[J]. *计算机应用研究*, 2022, 39(12):3534–3545.
 - [10] CHARTE F, RIVERA A J, DEL JESUS M J, et al. Addressing imbalance in multilabel classification; measures and random resampling algorithms[J]. *Neurocomputing*, 2015, 163:3–16.
 - [11] HU L, LI Y, GAO W, et al. Multi-label feature selection with shared common mode[J]. *Pattern Recognition*, 2020, 104:107344.
 - [12] RASTOGI R, MORTAZA S. Imbalance multi-label data learning with label specific features[J]. *Neurocomputing*, 2022, 513:395–408.
 - [13] 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.
 - [14] LI J, LI P, HU X, et al. Learning common and label-specific features for multi-Label classification with correlation information[J]. *Pattern Recognition*, 2022, 121:108259.
 - [15] CASTELLANOS F J, VALERO-MAS J J, CALVO-ZARAGOZA J, et al. Oversampling imbalanced data in the string space[J]. *Pattern Recognition Letters*, 2018, 103:32–38.
 - [16] CHARTE F, RIVERA A, DEL JESUS M J, et al. A first approach to deal with imbalance in multi-label datasets[C]//Hybrid artificial intelligent systems; 8th international conference. Salamanca: Springer, 2013:150–160.
 - [17] LUO F F, GUO W Z, CHEN G L. Addressing imbalance in weakly supervised multi-label learning[J]. *IEEE Access*, 2019, 7:37463–37472.
 - [18] TAHIR M A, KITTLER J, YAN F. Inverse random under sampling for class imbalance problem and its application to multi-label classification[J]. *Pattern Recognition*, 2012, 45(10):3738–3750.
 - [19] DANIELS Z, METAXAS D. Addressing imbalance in multi-label classification using structured hellinger forests[C]//Proceedings of the AAAI conference on artificial intelligence. San Francisco: AAAI, 2017:1826–1832.
 - [20] ZHU P, XU Q, HU Q, et al. Multi-label feature selection with missing labels[J]. *Pattern Recognition*, 2018, 74:488–502.
 - [21] JIAN L, LI J, SHU K, et al. Multi-label informed feature selection[C]//International joint conference on artificial intelligence (IJCAI). New York: [s. n.], 2016:1627–1633.
 - [22] ZHAN W, ZHANG M L. Multi-label learning with label-specific features via clustering ensemble[C]//2017 IEEE international conference on data science and advanced analytics (DSAA). Tokyo: IEEE, 2017:129–136.
 - [23] GUO Y, CHUNG F, LI G, et al. Leveraging label-specific discriminant mapping features for multi-label learning[J]. *ACM Transactions on Knowledge Discovery from Data*, 2019, 13(2):1–23.
 - [24] HUANG J, LI G, HUANG Q, et al. Learning label specific features for multi-label classification[C]//2015 IEEE international conference on data mining. New Jersey: IEEE, 2015:181–190.
 - [25] HAN H, HUANG M, ZHANG Y, et al. Multi-label learning with label specific features using correlation information[J]. *IEEE Access*, 2019, 7:11474–11484.
 - [26] HUANG J, LI G, HUANG Q, et al. Learning label-specific features and class-dependent labels for multi-label classification[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(12):3309–3323.
 - [27] BRAYTEE A, LIU W, ANAISSI A, et al. Correlated multi-label classification with incomplete label space and class imbalance[J]. *ACM Transactions on Intelligent Systems and Technology*, 2019, 10(5):1–26.
 - [28] JIA X Y, ZHU S S, LI W W. Joint label-specific features and correlation information for multi-label learning[J]. *Journal of Computer Science and Technology*, 2020, 35:247–258.
 - [29] WANG Y, ZHENG W, CHENG Y, et al. Joint label completion and label-specific features for multi-label learning algorithm[J]. *Soft Computing*, 2020, 24:6553–6569.
 - [30] CHARTE F, RIVERA A J, DEL JESUS M J, et al. MLSTMOTE: Approaching imbalanced multi label learning through synthetic instance generation[J]. *Knowledge-Based Systems*, 2015, 89:385–397.