DOI:10.11918/j.issn.0367-6234.201811004

结合标签相关性和不均衡性的多标签学习模型

姜思羽[1],钟晓玲[1],邱少健[2],宋恒杰[1]

(1.华南理工大学 软件学院,广州 510006;2.华南理工大学 计算机科学与工程学院,广州 510006)

摘 要:针对现有多标签学习算法较少兼顾标签间关联性和不平衡性的问题,提出一种同时考虑多标签间相关性与多标签不平衡问题的学习模型(A Multi-label Learning Model based on Label Correlation and Imbalance,MLCI).该学习模型针对每个标签类别,通过耦合其他标签类别以考量标签间的关联性,并降低缓解标签间不均衡比率,MLCI是一个将当前标签的二类不平衡学习器和多个与其他标签耦合的多类不平衡学习器结合的集成分类器.采用7种常用的多标签算法作为对比算法,针对yeast、scene、emotions和CAL500这4个开放数据集进行分类处理.实验结果表明,MLCI相比其他对比算法,在精度均值(Average- Precision)、排序损失(Ranking-Loss)、宏观平均AUC(Macro-Averaging AUC)和微观平均AUC(Micro-Averaging AUC)4个性能评估指标上总体占明显优势.

关键词:多标签学习;标签相关性;类不平衡;不平衡分类器,集成分类器

中图分类号:TP181 **文献标志码**:A **文章编号**:0367-6234(2019)01-0142-08

Amulti-label learning model based on label correlation and imbalance

JIANG Siyu[1], ZHONG Xiaoling[1], QIU Shaojian[2], SONG Hengjie[1]

(1.School of Software Engineering, South China University of Technology, Guangzhou 510006, China;
2. School of Computer Science & Engineering, South China University of Technology, Guangzhou 510006, China)

Abstract: Since the existing multi-label learning algorithms pay less attention to the problem of correlation and imbalance between label sets, this paper proposes a multi-label learning model based on label correlation and imbalance (MLCI). By coupling other label categories to consider the correlation among labels and reducing the imbalance ratio between labels of instances, the learning model is for each label category, and it is an ensemble classifier that combines the current class of binary imbalance classifier with multiple imbalanced classifiers coupled to other labels. In this paper, seven commonly used multi-label algorithms are used as comparison algorithms to classify the four open datasets of yeast, scene, emotions and CAL500. The experimental results show that the MLCI has obvious advantages in accuracy precision, ranking-Loss, macro-averaging AUC and micro-averaging AUC.

Keywords: Multi-label learning; label correlation; label imbalance; imbalance classifier, ensemble classifier

传统监督学习假设一个实例只属于一个标签类别[1].虽然传统监督学习在许多学习任务中得到广泛的应用,但是对于某些学习任务,其假设并不成立.因为随着数据量的不断增大,数据标签的复杂度也在增加,单一的样本可能同时具有多个语义含义.例如,在语义场景分类中,一篇新闻报道可能同时分类到社会、体育和娱乐等多个主题;在音乐类型分类中,一首歌可能同时属于摇滚和抒情类别;在图片标注中,一张图片可能同时被标注为建筑和城市.为了解决一个实例同时具有多个语义含义的问题,一个直接的方法是对一个实例赋予多个标签,以显式表达它的多个语义.基于以上考虑,多标签学习已经成为当代研究的一个热点.在多标签学习中,由于考虑了一个实例可能同时被标记为多种标签类别的情况,它与传统监督学习相比较,更适合用于实际生活中.多标签学习的早期研究主要集中在多标签文本分类问题[2].但随着多标签学习的发展,如今多标签学习已被广泛应用在自动注释的各种问题上,比如图像分类[3]、生物信息学[4]、网络挖掘[5]、规则挖掘[6]、信息检索[7]以及标记推荐[8]等.

多标签学习算法主要分为问题转换方法和算法适应方法.问题转换方法将多标签问题转换成其他已知的学习问题进行求解.问题转换方法的代表性算法有二元关联(Binary Relevance, BR)、组合分类器链[9](Ensembles of Classifier Chains, ECC)、校准的标签排序算法[10](Calibrated Label Ranking, CLR)和随机k标签集[11](Random K-labelsets, RAKEL)等.算法适应方法改进已有的流行算法,直接处理多标签数据.算法适应方法的代表性算法有

收稿日期:2018-11-01
作者简介:姜思羽(1992—),女,博士研究生;
宋恒杰(1976—),男,教授,博士生导师
通信作者:宋恒杰,sehjsong@scut.edu.cn

多标签 k 最近邻分类算法[12-15]（ML-KNN）、多类多标记组合分类[16]（MMAC）、条件随机场[17-18]（Conditional Random Fields，CRF）、反向传播多标签学习法[19]（BP-MLL）等方法.

现今，多标签学习热点研究中存在两个问题. 其一，在多标签学习中，随着标签类别数量的增加，实例预测的可能标签集合的数量会指数增长. 例如，在多标签学习中，标签空间有 20 个标签类别，则实例可能的标签集合将超过 1 百万（即 2^{20}）个. 为了解决该问题，现有的算法主要研究标签类别间的联系[20]. 其二，由于类的正样本/负样本数量可能远远少于负样本/正样本数量，多标签学习存在类的不平衡问题，该问题也导致大多数分类方法性能的下降[21]. 例如在预测肺癌的场景中，肺癌患者在所有来诊病人中所占的比例非常低. 在该场景下如果直接构建模型，会导致数据集以负样本为主，只有很少的正样本. 由于分类模型的目标是希望准确率尽可能达到最高，因此该场景问题最终会导致预测的结果全部偏向负样本，但在实际生活中，医生和病人往往更加关注肺癌患者的情况，对非肺癌患者的关注度并没有那么高. 所以类不平衡问题会严重影响到分类的效果. 长期以来，类不平衡一直被视为危及机器学习算法性能的一个基本威胁. 在多标签学习中，已有的算法基本没有充分考虑类不平衡问题.

本文在学习多标签问题时，提出了同时考虑多标签间相关性与多标签不平衡问题的学习模型. 该学习模型针对每个标签类别，通过耦合其他标签类别以考量标签间的关联性，并缓解标签间不均衡比率，是一个将当前标签的二类不平衡学习器和多个与其他标签耦合的多类不平衡学习器结合的集成分类器（multi-label learning model based on label correlation and imbalance，MLCI）.

1　方法描述

1.1　多标签问题的定义

在多标签学习中，一个实例被赋予多个标签. 假设 $\boldsymbol{X}=\mathbf{R}^n$ 表示 n 维的实例空间，$\boldsymbol{Y}=\{y_1, y_2, \cdots, y_q\}$ 表示标签空间，该标签空间有 q 个可能的标签类别. 多标签学习的任务是，从多标签训练数据集 $\boldsymbol{D}=\{(\boldsymbol{x}_i,\boldsymbol{Y}_i)\mid 1\leqslant i\leqslant m\}$ 中学习一个函数 $h:\boldsymbol{X}\rightarrow 2^{\boldsymbol{Y}}$. 对于每个多标签实例 $(\boldsymbol{x}_i,\boldsymbol{Y}_i)$，$\boldsymbol{x}_i\in\boldsymbol{X}$ 是一个 n 维的特征向量 $(x_{i1},x_{i2},\cdots,x_{in})^{\mathrm{T}}$，$\boldsymbol{Y}_i\in\boldsymbol{Y}$ 是实例 $\boldsymbol{x}_i$ 的标签集合. 对于任何一个测试实例 $\boldsymbol{x}\in\boldsymbol{X}$，多标签分类器 $h(\cdot)$ 预测 $\boldsymbol{x}$ 的标签集合 $h(\boldsymbol{x})\in\boldsymbol{Y}$. 在一般的情况下，多标签学习系统学习的实数值函数一般返回一个实数：$f_k:\boldsymbol{X}\rightarrow\mathbf{R}$，$f_k(\boldsymbol{X})$ 表示 $\boldsymbol{X}$ 被 y_k 标记的信心，每个标签类别伴随着一个阈值函数 $t_k:\boldsymbol{X}\rightarrow\mathbf{R}$. 因此，预测标签集合的函数 $h(\boldsymbol{x})$ 一般等价于

$$h(\boldsymbol{x})=\{y_k\mid f_k(\boldsymbol{x})>t_k(\boldsymbol{x}),1\leqslant k\leqslant q\}.$$

1.2　算法

该算法不仅考虑了标签类别之间的相关性，也考虑了标签的类不平衡问题. 令 $\boldsymbol{D}_k^+$ 和 $\boldsymbol{D}_k^-$ 分别表示根据标签类别 y_k，对数据集 $\boldsymbol{D}$ 进行划分得到的正训练样本数据集和负训练样本数据集.

$$\boldsymbol{D}_k^+=\{(\boldsymbol{x}_i,+1)\mid y_k\in\boldsymbol{Y}_i,1\leqslant i\leqslant m\},$$

$$\boldsymbol{D}_k^-=\{(\boldsymbol{x}_i,+1)\mid y_k\notin\boldsymbol{Y}_i,1\leqslant i\leqslant m\}.$$

标签 y_k 的类不平衡率为

$$\mathrm{IR}_k=\max(|\boldsymbol{D}_k^+|,|\boldsymbol{D}_k^-|)/\min(|\boldsymbol{D}_k^+|,|\boldsymbol{D}_k^-|).$$

式中 IR_k 表示标签 y_k 的类不平衡率. 由于类的不平衡，IR_k 的值将会很大. 例如，在本文实验中所采用的 4 个多标签数据集中，标签空间的最大类不平衡率从 3.0 到 24.4.

为了同时解决标签类别之间存在关联性和类不平衡问题，本文对于每个标签类别，构建一个对应于当前标签的二类不平衡学习器，以及多个与其他标签耦合的多类不平衡学习器. 之后，通过聚合二类学习器和多类学习器产生的输出，获得对每个类标签的最终预测.

1.2.1　二类不平衡学习器

该算法对多标签数据集进行处理，采用问题转换方法，将多标签问题转换为 q 个独立互不影响的二元分类问题.

令 $\boldsymbol{D}_k$ 表示根据第 k 个类标签 y_k，划分数据集 $\boldsymbol{D}$ 而得到的二类训练数据集，则 $\boldsymbol{D}_k$ 为

$$\boldsymbol{D}_k=\{(\boldsymbol{x}_i,\varnothing(\boldsymbol{Y}_i,y_k))\mid 1\leqslant i\leqslant m\}. \quad (1)$$

式中 $\varnothing(\boldsymbol{Y}_i,y_k)=\begin{cases}+1,\text{if } y_k\in\boldsymbol{Y}_i;\\ -1,\text{otherwise.}\end{cases}$

因此，$\boldsymbol{D}_k^+$ 表示了二类训练数据集 $\boldsymbol{D}_k$ 的正样本数据集，$\boldsymbol{D}_k^-$ 表示了二类训练数据集 $\boldsymbol{D}_k$ 的负样本数据集，IR_k 表示了二类训练数据集的类不平衡率.

对于分解后的二元分类问题，为了解决 $\boldsymbol{D}_k^+$ 和 $\boldsymbol{D}_k^-$ 之间的数据量不平衡问题，可以采用已有的二元类不平衡学习方法解决，比如过采样、欠采样方法. 由于本文针对 yeast、scene、emotions 和 CAL500 这 4 个开放数据集进行分类处理，而这 4 个开放数据集的样本数量过少，因此本文采用了合成少数样本过采用算法[22]（Synthetic Minority Over-sampling Technique，SMOTE），添加部分样本的副本.

SMOTE 算法对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中. SMOTE 算法流程如下：

1)以欧式距离为基准,根据样本到少数类样本集中所有样本的距离,计算得到少数类样本集中每一个样本 $\boldsymbol{x}$ 的 k 个近邻.

2) 根据类不平衡比例设置采样倍率 r,对于少数类中的每一个样本 $\boldsymbol{x}$,从其 K 近邻中随机选择 r 个样本.

3) 对于每一个随机选出的近邻 $\boldsymbol{xk}$,分别与原样本 $\boldsymbol{x}$ 构建新的样本添加到数据集中:

$$\boldsymbol{x}^{'} = \boldsymbol{x} + \mathrm{rand}(0,1) * | \boldsymbol{x} - \boldsymbol{xk} |.$$

本文对数据集 $\boldsymbol{D}_k$ 采用二元类不平衡方法中的 SMOTE 算法来学习二元分类器 g_k,即 $g_k \leftarrow \S(\boldsymbol{D}_k)$. 令 $g_k(+1 | \boldsymbol{x})$ 表示对于标签类别 y_k,样本 $\boldsymbol{x}$ 被预测为正样本的信心. 换而言之,实数值函数 $f_k(\boldsymbol{x}) = g_k(+1 | \boldsymbol{x})$. 而阈值函数 $t_k(\boldsymbol{x})$ 可以被简单设置为常数函数,比如 $t_k(\boldsymbol{x}) = 0$.

1.2.2 多类不平衡学习器

考虑到标签类别之间存在关联性,本文随机选择 2 个标签 y_a, y_b 作为 y_k 的耦合标签.

令 $\boldsymbol{D}_{kab}$ 表示根据标签对 (y_k, y_a, y_b) 划分数据集 $\boldsymbol{D}$ 而得到的多类训练数据集,则 $\boldsymbol{D}_{kab}$ 为

$$\boldsymbol{D}_{kab} = \{(\boldsymbol{x}_i, \varphi(\boldsymbol{Y}_i, y_k, y_a, y_b)) | 1 \leqslant i \leqslant m\},$$

式中 $\varphi(\boldsymbol{Y}_i, y_k, y_a, y_b)$ 表示数据集 $\boldsymbol{D}_{kab}$ 中实例的类别,即

$$\varphi(\boldsymbol{Y}_i, y_k, y_a, y_b) = \begin{cases} 0, & \text{if } y_k \notin \boldsymbol{Y}_i \text{ and } y_a \notin \boldsymbol{Y}_i \text{ and } y_b \notin Y_i; \\ +1, & \text{if } y_k \notin \boldsymbol{Y}_i \text{ and } y_a \notin \boldsymbol{Y}_i \text{ and } y_b \in \boldsymbol{Y}_i; \\ +2, & \text{if } y_k \notin \boldsymbol{Y}_i \text{ and } y_a \in \boldsymbol{Y}_i \text{ and } y_b \notin \boldsymbol{Y}_i; \\ +3, & \text{if } y_k \notin \boldsymbol{Y}_i \text{ and } y_a \in \boldsymbol{Y}_i \text{ and } y_b \in \boldsymbol{Y}_i; \\ +4, & \text{if } y_k \in \boldsymbol{Y}_i \text{ and } y_a \notin \boldsymbol{Y}_i \text{ and } y_b \notin \boldsymbol{Y}_i; \\ +5, & \text{if } y_k \in \boldsymbol{Y}_i \text{ and } y_a \notin \boldsymbol{Y}_i \text{ and } y_b \in \boldsymbol{Y}_i; \\ +6, & \text{if } y_k \in \boldsymbol{Y}_i \text{ and } y_a \in \boldsymbol{Y}_i \text{ and } y_b \notin \boldsymbol{Y}_i; \\ +7, & \text{if } y_k \in \boldsymbol{Y}_i \text{ and } y_a \in \boldsymbol{Y}_i \text{ and } y_b \in \boldsymbol{Y}_i. \end{cases}$$

在学习过程中,尽管利用 $\boldsymbol{D}_{kab}$ 考虑了标签类别间的关联性,但是联合考虑 y_k, y_a, y_b 将会放大类的不平衡问题. 假设在二类训练数据集 $\boldsymbol{D}_k(\boldsymbol{D}_a, \boldsymbol{D}_b)$ 中,$\boldsymbol{D}_k^+(\boldsymbol{D}_a^+, \boldsymbol{D}_b^+)$ 是相应的少数类样本集,则在多类训练数据集 $\boldsymbol{D}_{kab}$ 中,第一个类 $\varphi(\boldsymbol{Y}_i, y_k, y_a, y_b) = 0$ 的样本数将会是最多的,而第八个类 $\varphi(\boldsymbol{Y}_i, y_k, y_a, y_b) = 7$ 的样本数将会是最少的.

令 N_{kabj} 表示数据集 $\boldsymbol{D}_{kab}$ 中,$\varphi(\boldsymbol{Y}_i, y_k, y_a, y_b) = j$ 的样本数,即 N_{kabj} 为

$$N_{kabj} = | \{\boldsymbol{x}_i | 1 \leqslant i \leqslant m, \varphi(\boldsymbol{Y}_i, y_k, y_a, y_b) = j\} |. \tag{2}$$

$\mathrm{IR}_k, \mathrm{IR}_a, \mathrm{IR}_b$ 分别表示二类训练数据集 $\boldsymbol{D}_k, \boldsymbol{D}_a, \boldsymbol{D}_b$ 的类不平衡率,因此,在假设样本分布均匀的条件下,在 $\boldsymbol{D}_{kab}$ 数据集中,第一个类 $\varphi(\boldsymbol{Y}_i, y_k, y_a, y_b) = 0$ 的样本数与第八个类的样本数的比例 IR_{kab07} 为

$$\mathrm{IR}_{kab07} = \frac{N_{kab0}}{N_{kab7}} = \frac{N_{kab0}}{N_{kab1}} \times \frac{N_{kab1}}{N_{kab3}} \times \frac{N_{kab3}}{N_{kab7}}. \tag{3}$$

根据式(2)和式(3),可知 IR_{kab07} 为

$$\mathrm{IR}_{kab07} \approx \mathrm{IR}_b \times \mathrm{IR}_a \times \mathrm{IR}_k.$$

为了解决 $\mathrm{IR}_b \times \mathrm{IR}_a \times \mathrm{IR}_k$ 即 IR_{kab07} 过大的问题,本文将第五、六、七、八类合并成一个新类,将 $\boldsymbol{D}_{kab}$ 转换成一个新的数据集

$$\boldsymbol{D}_{kab}^{'} = \{(\boldsymbol{x}_i, \varphi^{'}(\boldsymbol{Y}_i, y_k, y_a, y_b)) | 1 \leqslant i \leqslant m\},$$

式中,

$$\varphi^{'}(\boldsymbol{Y}_i, y_k, y_a, y_b) = \begin{cases} 0, & \text{if } y_k \notin \boldsymbol{Y}_i \text{and } y_a \notin \boldsymbol{Y}_i \text{and } y_b \notin \boldsymbol{Y}_i; \\ +1, & \text{if } y_k \notin \boldsymbol{Y}_i \text{and } y_a \notin \boldsymbol{Y}_i \text{and } y_b \in \boldsymbol{Y}_i; \\ +2, & \text{if } y_k \notin \boldsymbol{Y}_i \text{and } y_a \in \boldsymbol{Y}_i \text{and } y_b \notin \boldsymbol{Y}_i; \\ +3, & \text{if } y_k \notin \boldsymbol{Y}_i \text{and } y_a \in \boldsymbol{Y}_i \text{and } y_b \in \boldsymbol{Y}_i; \\ +4, & \text{if } y_k \in \boldsymbol{Y}_i. \end{cases}$$

令 $\mathrm{IR}_{kabzw}^{'}$ 表示在新的数据集 $\boldsymbol{D}_{kab}^{'}$ 中,类 $\varphi^{'}(\boldsymbol{Y}_i, y_k, y_a, y_b) = z$ 的样本数与类 $\varphi^{'}(\boldsymbol{Y}_i, y_k, y_a, y_b) = w$ 的样本数的比例. 因此第一类 $\varphi^{'}(\boldsymbol{Y}_i, y_k, y_a, y_b) = 0$ 的样本数与第五类类 $\varphi^{'}(\boldsymbol{Y}_i, y_k, y_a, y_b) = 4$ 的样本数的比例 $\mathrm{IR}_{kab04}^{'}$ 为:

$$\mathrm{IR}_{kab04}^{'} = \frac{N_{kab0}}{N_{kab4} + N_{kab5} + N_{kab6} + N_{kab7}},$$

$$\frac{1}{\mathrm{IR}_{kab04}^{'}} = \frac{N_{kab4} + N_{kab5} + N_{kab6} + N_{kab7}}{N_{kab0}}$$

$$\frac{1}{\mathrm{IR}_{kab04}^{'}} = \frac{N_{kab4}}{N_{kab0}} + \frac{N_{kab5}}{N_{kab0}} + \frac{N_{kab6}}{N_{kab0}} + \frac{N_{kab7}}{N_{kab0}}$$

$$\frac{1}{\mathrm{IR}_{kab04}^{'}} = \frac{N_{kab4}}{N_{kab0}} + \frac{N_{kab5}}{N_{kab1}} \times \frac{N_{kab1}}{N_{kab0}} + \frac{N_{kab6}}{N_{kab2}} \times \frac{N_{kab2}}{N_{kab0}} + \frac{N_{kab7}}{N_{kab3}} \times \frac{N_{kab3}}{N_{kab2}} \times \frac{N_{kab2}}{N_{kab0}}. \tag{4}$$

根据式(2)和式(4),可知

$$\frac{1}{\mathrm{IR}_{kab04}^{'}} \approx \frac{1}{\mathrm{IR}_k} + \frac{1}{\mathrm{IR}_k} \times \frac{1}{\mathrm{IR}_b} + \frac{1}{\mathrm{IR}_k} \times \frac{1}{\mathrm{IR}_a} + \frac{1}{\mathrm{IR}_k} \times \frac{1}{\mathrm{IR}_b} \times \frac{1}{\mathrm{IR}_a}.$$

$$\mathrm{IR}_{kab04}^{'} = \frac{\mathrm{IR}_b \times \mathrm{IR}_a \times \mathrm{IR}_k}{\mathrm{IR}_b \times \mathrm{IR}_a + \mathrm{IR}_a + \mathrm{IR}_b + 1}. \tag{5}$$

同理可得,第二类 $\varphi^{'}(\boldsymbol{Y}_i, y_k, y_a, y_b) = 1$、第三类 $\varphi^{'}(\boldsymbol{Y}_i, y_k, y_a, y_b) = 2$、第四类 $\varphi^{'}(\boldsymbol{Y}_i, y_k, y_a, y_b) = 3$ 的样本数与第五类类 $\varphi^{'}(\boldsymbol{Y}_i, y_k, y_a, y_b) = 4$ 的样本数的比例分别为

$$\mathrm{IR}'_{kab14} \approx \frac{\mathrm{IR}_a \times \mathrm{IR}_k}{\mathrm{IR}_b \times \mathrm{IR}_a + \mathrm{IR}_a + \mathrm{IR}_b + 1}, \quad (6)$$

$$\mathrm{IR}'_{kab24} \approx \frac{\mathrm{IR}_b \times \mathrm{IR}_k}{\mathrm{IR}_b \times \mathrm{IR}_a + \mathrm{IR}_a + \mathrm{IR}_b + 1}, \quad (7)$$

$$\mathrm{IR}'_{kab34} \approx \frac{\mathrm{IR}_k}{\mathrm{IR}_b \times \mathrm{IR}_a + \mathrm{IR}_a + \mathrm{IR}_b + 1}. \quad (8)$$

由式(5)~(8)可知，IR'_{kab04}、IR'_{kab14}、IR'_{kab24}、IR'_{kab34} 都远小于 IR_{kab07}，因此该算法在一定程度上减轻了多标签学习过程中的类不平衡性.

该算法对数据集 $\boldsymbol{D}'_{kab}$ 采用多类不平衡方法£来学习多类分类器 g_{kab}，即 $g_{kab} \leftarrow £(\boldsymbol{D}'_{kab})$. 相应地，令 $g_{kab}(+4 \mid \boldsymbol{x})$ 表示对于标签类别 y_k，样本 $\boldsymbol{x}$ 被预测为正样本的信心.

1.2.3　二类学习器和多类学习器的聚合

对于每一个标签类别 y_k，该算法随机选取 r 个标签对 $(y_a, y_b)(a \neq k, b \neq k, a \neq b)$，分别与 y_k 耦合成标签对 (y_k, y_a, y_b). 数据集 $\boldsymbol{D}$ 根据标签对 (y_k, y_a, y_b) 的值构建新的数据集 $\boldsymbol{D}'_{kab}$，进而根据多类不平衡方法£来学习 r 个多类分类器 g_{kab}，基于集成学习的线性组合思想，实数值函数 $f_k(x)$ 线性集成了二类分类器和 r 个多类分类器的预测结果，即 $f_k(x) = g_k(+1 \mid \boldsymbol{x}) + \sum_{(y_a, y_b)} g_{kab}(+4 \mid \boldsymbol{x})$. 而对于阈值函数 $t_k(\boldsymbol{x})$，该算法将它设置为常数函数 $t_k(\boldsymbol{x}) = l_k$. 因此，对于测试实例 $\boldsymbol{x}$，当 $f_k(\boldsymbol{x}) > l_k$ 时，实例 $\boldsymbol{x}$ 被预测为标签类别 y_k 的正样本，否则，实例 $\boldsymbol{x}$ 被预测为标签类别 y_k 的负样本. 通过将 l_k 作为二分阈值，本文采用综合评价指标(F-measure)来衡量 f_k 对数据集 $\boldsymbol{D}_k$ 的分类性能. 根据分类性能最优化原则，选择合适的 l_k 值. 令 $F(f_k, l, \boldsymbol{D}_k)$ 表示将 l 作为二分阈值，分类器 f_k 对二元数据集 $\boldsymbol{D}_k$ 的分类性能 F-measure 的值，因此常数阈值 $l_k = \arg\max_{l \in R} F(f_k, l, \boldsymbol{D}_k)$.

表 1 总结了本文提出的算法的完整过程. 对于每一个标签类别 y_k，该算法通过处理多标签训练数据集 $\boldsymbol{D}$，学习一个二类不平衡分类器和 r 个耦合的多类不平衡分类器. 之后，该算法通过集成二类分类器和多类分类器的预测信心，构建针对标签类别 y_k 的预测模型. 最后，对测试实例，分别查询每个标签类别的预测模型，从而获取测试实例的预测标签集.

对于每一个标签类别 y_k，该算法在一个多类分类器中考察了 2 个标签与该标签 y_k 的关联性，该算法针对一个标签类别，学习 r 个多类分类器，因此实际上，该算法考察了多个标签与该标签 y_k 的关联性，属于高阶策略思想.

表 1　MLCI 算法伪代码

Tab.1　The pseudo-code of MLCI

输入：

$\boldsymbol{D}$：多标签训练数据集
§：二类不平衡分类器
£：多类不平衡分类器
r：耦合类标签对的个数
$\boldsymbol{x}$：测试实例($\boldsymbol{x} \in \boldsymbol{X}$)

输出：

$\boldsymbol{Y}$：测试实例 $\boldsymbol{x}$ 的预测标签集

过程：

```
1) for k = 1 to q do
2:   根据式(1)构建二类训练数据集 D_k
3)   g_k ← §(D_k)
4)   随机选取 r 个标签对 (y_a, y_b)
5)   for 每个标签对 (y_a, y_b) do
6)     根据式构建多类数据集 D'_kab
7)     g_kab ← £(D'_kab)
8)   end for
9)   根据式设置实数值函数 f_k(·)
10)  根据式学习常数 l_k，从而设置常数阈值函数 t_k(·)
11)  end for
12)  根据式返回 Y = h(x)
```

2　实验及分析

通过对 4 个真实多标签数据集进行广泛的实验来验证方法的有效性. 在实验中，将 MLCI 与几个先进的多标签学习方法做对比，并且以 20 次交叉的方式进行验证，10 次实验的平均结果表明，所提的 MLCI 优于其他算法.

2.1　数据集和评价标准

2.1.1　数据集

为了测评算法的性能，本文在 Mulan (Mulan 网址：http://mulan.sourceforge.net/datasets-mlc.html)平台上调用了 4 个真实多标签数据集，分别来自图像、音乐和生物信息等方面，对于每个数据集 $\boldsymbol{D}$，我们采用 $T(\boldsymbol{D})$、$I(\boldsymbol{D})$、$L(\boldsymbol{D})$、$F(\boldsymbol{D})$、$C(\boldsymbol{D})$ 分别表示数据集 $\boldsymbol{D}$ 的域类型、样本数、标签类别的数量、样本的特征数量以及基数. 各个数据集具体的参数详见表 2.

表 2　数据集的相关信息

Tab.2　Characteristics of the data sets

Name	$T(\boldsymbol{D})$	$I(\boldsymbol{D})$	$L(\boldsymbol{D})$	$F(\boldsymbol{D})$	$C(\boldsymbol{D})$
Yeast	biology	2417	14	2417	4.237
Scene	images	2407	6	2407	1.074
Emotions	music	593	6	593	1.869
CAL500	images	502	174	502	26.044

基于此,分析每个数据集中的数据不均衡问题,各个数据集具体的参数详见表 3,其中,min_IR 表示数据集中最小的类不平衡率 ($\min\_IR = \min\limits_{1\leqslant k\leqslant q} IR_k$),max_IR 表示数据集中最大的类不平衡率 ($\max\_IR = \max\limits_{1\leqslant k\leqslant q} IR_k$),ave_IR 表示数据集中的平均类不平衡率 ($ave\_IR = \frac{1}{q}\sum\limits_{k=1}^{q} IR_k$).

表 3 数据集的类不平衡信息

Tab.3 Class-imbalance of the data sets

Name	min_IR	max_IR	ave_IR
Yeast	1.328	12.500	2.778
Scene	3.521	5.618	4.566
Emotions	1.247	3.003	2.146
CAL500	1.040	24.390	3.846

2.1.2 度量方法

在实验中,为了更全面度量本算法与其他算法的效果,本文采用了 2 种度量方法:基于样本的度量方法和基于标签的度量方法[23]. 将 $(\boldsymbol{x}_i, \boldsymbol{Y}_i)(i = 1,..,m)$ 表示为测试数据集中的样本,其中 $\boldsymbol{x}_i$ 是第 i 个测试样本的特征向量,$\boldsymbol{Y}_i$ 是测试集中第 i 个样本的标签集合,$\gamma_i(\lambda)$ 表示测试集合中第 i 个样本标签 λ 在 $\boldsymbol{Y}_i$ 中的排位.

在基于样本的度量方法中,主要选取了 Average-Precision 和 Ranking-Loss.

Average-Precision 方法衡量了分类器对测试集中所有样本的预测标签排序中,排在相关标签前面的标签,也是相关标签概率的平均值.

$$\text{Average-Precision} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|\boldsymbol{Y}_i|}\sum_{\lambda\in\boldsymbol{Y}_i}\frac{|\{\lambda'\in\boldsymbol{Y}_i:\gamma_i(\lambda')\leqslant\}\gamma_i(\lambda)|}{\gamma_i(\lambda)}.$$

Ranking-Loss 方法衡量了分类器对测试集中所有样本的预测标签排序中,排在相关标签前面的标签,是不相关标签概率的平均值.

$$\text{Ranking-Loss} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|\boldsymbol{Y}_i||\bar{\boldsymbol{Y}}_i|}|\{(\lambda_a,\lambda_b):\gamma_i(\lambda_a)>\gamma_i(\lambda_b)(\lambda_b),(\lambda_a,\lambda_b)\in\boldsymbol{Y}_i\times\bar{\boldsymbol{Y}}_i\}|.$$

基于标签的度量方法是预测每个单独的标签,然后对所有标签的结果取平均值进行度量,在单标签度量方法中,将 $tp_\lambda, tn_\lambda, fp_\lambda, tp_\lambda$ 分别表示第 λ 标签的正阳例(true positives, TP),负阳例(true negatives, TN),正阴例(false positives, FP),负阴例(true negatives, TN)的数目. 在这种度量方法中有 2 种方法用以实现平均值:宏观平均(Macro-Averaging)和微观平均(Micro-Averaging):

$$\text{Macro-Averaging-B} = \frac{1}{m}\sum_{\lambda}^{m} B(tp_\lambda, tn_\lambda, fp_\lambda, tp_\lambda),$$

$$\text{Micro-Averaging-B} = B(\sum_{\lambda=1}^{m} tp_\lambda, \sum_{\lambda=1}^{m} tn_\lambda, \sum_{\lambda=1}^{m} fp_\lambda, \sum_{\lambda=1}^{m} tp_\lambda).$$

其中,B 代表的是二元分类器中 F1 和 AUC 评价标准.

2.2 对比算法及实验分析

为了全面的进行对比,本文选取了 7 个对比算法,分别是交叉耦合的聚合算法[24](Cross-coupling Aggregation, COCOA),ML-kNN[12],BP-MLL[19],结合实例和逻辑回归的多标签分类[25](Combining Instance-Based Learning and Logistic Regression for Multilabel classification, IBLR),ECC[9],CLR[10],RAKEl[11].

实验结果包括 MLCI 与对比算法在 4 组数据集上的性能测评结果,结果采用度量标准的平均值和标准差表示,在表中,符号 ↑ 表示测评结果的值越大,对应的算法性能越好;符号 ↓ 表示测评结果的值越小,对应的算法性能越好. 由于文章边幅有限,本文不详细描述每次实验的性能测试结果,仅描述 10 次实验的平均性能测评结果.

2.2.1 基于样本的度量方法上的性能评判

在基于样本的度量方法中,MLCI 与对比算法在 4 个数据集上的性能测评结果如表 4,表 5 所示. 在 Average-Precision 评价标准中,MLCI 在 4 个数据集中此评测标准整体高于其他对比算法. 具体分析中,MLCI 在 scene 和 CAL500 数据集上的平均差值要更优于在数据集 yeast 和 emotions 上的平均差值,说明数据集标签分布越不均衡,MLCI 得到的相关标签的概率越高,效果越好. 在 Ranking-Loss 评价标准中,MLCI 在 4 个数据集中此评测标准整体高于其他对比算法. 此外,MLCI 在 yeast 数据集和 CAL500 数据集和 yeast 数据集上的表现要更由于另 2 个数据集,说明数据集 Cardinality 值越大,MLCI 得到的不相关标签概率越低,效果越好.

2.2.2 基于标签的度量方法上的性能评判

在基于标签的度量方法中,MLCI 与对比算法在 4 个数据集上的性能测评结果如表 6~9 所示. 在 Macro-Averaging F1 评价标准中,MLCI 在 4 个数据集中此评测标准整体高于其他对比算法. 但是在 Micro-averaging F1 评价标准中,MLCI 的效果略有下降,特别是在数据集上 yeast 数据集上的表现并不理想. 在 Macro-Averaging AUC 和 Micro-Averaging AUC 评价标准中,MLCI 在 4 个数据集中此评测标准整体都高于其他对比算法,并且效果表现的较为平均.

表 4　在每个数据集上，每种算法的 Average-Precision 性能（↑）

Tab.4　Performance of each algorithm in terms of Average-Precision in each data set（↑）

算法	Yeast	scene	emotions	CAL500
MLCI	**0.773 2±0.017 0**	**0.888 4±0.016 6**	**0.827 2±0.036 2**	**0.521 2±0.020 2**
COCOA	0.767 6±0.020 9	0.873 7±0.017 4	0.804 9±0.022 7	0.491 8±0.019 2
ML-kNN	0.766 9±0.024 7	0.866 2±0.017 4	0.801 6±0.028 7	0.506 6±0.015 5
BP-MLL	0.751 6±0.021 7	0.668 5±0.040 7	0.798 5±0.036 2	0.509 5±0.001 6
IBLR	0.768 3±0.025 3	0.867 3±0.018 0	0.815 0±0.030 6	0.313 0±0.010 7
ECC	0.745 9±0.027 5	0.847 8±0.014 5	0.794 7±0.021 5	0.473 0±0.008 1
CLR	0.749 6±0.022 1	0.820 9±0.022 3	0.784 7±0.024 8	0.508 4±0.004 8
RAKEL	0.639 6±0.022 2	0.718 6±0.004 3	0.704 1±0.015 4	0.244 4±0.010 0
平均差值/%	4.0	7.9	4.0	8.6

表 5　在每个数据集上，每种算法的 Ranking-Loss 性能（↓）

Tab.5　Performance of each algorithm in terms of Ranking-Loss in each data set（↓）

算法	yeast	scene	emotions	CAL500
MLCI	**0.156 9±0.011 2**	**0.060 1±0.009 0**	**0.135 3±0.023 2**	**0.229 3±0.012 3**
COCOA	0.163 4±0.011 6	0.069 5±0.010 2	0.152 9±0.023 1	0.251 8±0.012 3
ML-kNN	0.165 1±0.016 0	0.077 4±0.011 6	0.161 1±0.022 1	0.235 6±0.009 3
BP-MLL	0.452 9±0.033 2	0.186 1±0.029 7	0.163 1±0.034 0	0.240 0±0.000 6
IBLR	0.164 3±0.016 5	0.076 0±0.011 6	0.148 1±0.025 5	0.329 0±0.008 8
ECC	0.183 1±0.016 3	0.089 0±0.011 4	0.164 1±0.017 4	0.266 0±0.003 9
CLR	0.178 4±0.015 7	0.101 1±0.013 5	0.175 9±0.025 5	0.236 8±0.003 9
RAKEL	0.334 0±0.022 7	0.209 9±0.001 7	0.294 8±0.022 1	0.553 5±0.019 8
平均差值/%	-7.7	-5.5	-4.5	-7.3

表 6　在每个数据集上，每种算法的 Macro-Averaging F1 性能（↑）

Tab.6　Performance of each algorithm in terms of Macro-Averaging F1 in each data set（↑）

算法	yeast	scene	emotions	CAL500
MLCI	0.446 3±0.004 0	0.739 2±0.019 4	0.677 9±0.035 6	0.313 0±0.013 0
COCOA	0.419 5±0.027 7	0.736 8±0.031 0	0.668 2±0.030 5	0.209 0±0.014 3
ML-kNN	0.395 3±0.036 5	0.735 5±0.021 4	0.629 5±0.050 7	0.087 9±0.010 7
BP-MLL	**0.452 9±0.033 2**	0.526 3±0.046 2	0.677 1±0.044 6	0.260 9±0.003 0
IBLR	0.397 2±0.044 7	**0.748 5±0.025 7**	0.657 3±0.038 4	0.214 4±0.012 0
ECC	0.379 6±0.039 1	0.710 2±0.012 2	0.619 7±0.029 5	0.131 5±0.003 9
CLR	0.394 8±0.039 4	0.644 2±0.018 3	0.599 6±0.019 5	0.086 3±0.011 4
RAKEL	0.390 4±0.037 6	0.609 0±0.006 9	0.576 2±0.027 5	0.171 5±0.003 4
平均差值/%	4.2	6.6	4.4	14.7

表 7 在每个数据集上,每种算法的 Micro-Averaging F1 性能(↑)

Tab.7 Performance of each algorithm in terms of Micro-Averaging F1 in each data set(↑)

算法	yeast	scene	emotions	CAL500
MLCI	0.546 5±0.005 0	0.719 7±0.021 9	0.679 9±0.035 8	0.424 7±0.013 2
COCOA	0.639 1±0.024 9	0.732 1±0.037 0	0.685 7±0.031 1	0.398 1±0.018 4
ML-kNN	0.648 8±0.028 1	0.734 3±0.025 8	0.665 2±0.047 0	0.330 5±0.009 2
BP-MLL	**0.657 0±0.022 9**	0.506 7±0.041 3	**0.691 6±0.041 4**	**0.757 8±0.001 0**
IBLR	0.650 9±0.029 7	**0.745 2±0.031 4**	0.684 4±0.039 3	0.332 4±0.015 3
ECC	0.621 9±0.032 0	0.705 0±0.015 5	0.646 9±0.030 9	0.358 6±0.000 8
CLR	0.623 0±0.027 2	0.627 6±0.023 4	0.614 7±0.019 5	0.313 9±0.021 0
RAKEL	0.581 4±0.026 6	0.597 2±0.007 1	0.599 0±0.023 1	0.355 4±0.015 9
平均差值/%	-8.5	5.5	2.5	1.8

表 8 在每个数据集上,每种算法的 Macro-Averaging AUC 性能(↑)

Tab.8 Performance of each algorithm in terms of Macro-Averaging AUC in each data set(↑)

算法	yeast	scene	emotions	CAL500
MLCI	**0.723 9±0.0041**	**0.951 4±0.0075**	**0.862 7±0.0221**	**0.572 2±0.0045**
COCOA	0.722 2±0.014 8	0.943 6±0.008 4	0.846 2±0.028 0	0.561 9±0.006 2
ML-kNN	0.679 0±0.004 0	0.934 4±0.011 2	0.837 7±0.028 5	0.509 4±0.005 0
BP-MLL	0.710 0±0.004 1	0.860 7±0.014 1	0.839 8±0.032 8	0.569 7±0.002 3
IBLR	0.698 0±0.004 1	0.942 2±0.010 2	0.853 9±0.026 9	0.510 1±0.006 0
ECC	0.689 0±0.004 0	0.925 3±0.012 0	0.830 7±0.015 4	0.542 2±0.005 1
CLR	0.650 0±0.004 0	0.901 9±0.010 4	0.810 2±0.020 8	0.566 4±0.001 3
RAKEL	0.640 0±0.004 0	0.756 5±0.005 5	0.700 3±0.018 5	0.510 4±0.000 3
平均差值/%	4	5.6	4.6	3.4

表 9 在每个数据集上,每种算法的 Micro-Averaging AUC 性能(↑)

Tab.9 Performance of each algorithm in terms of Micro-Averaging AUC in each data set(↑)

算法	yeast	scene	emotions	CAL500
MLCI	**0.855 8±0.011 2**	**0.957 0±0.006 9**	**0.879 5±0.018 6**	**0.767 7±0.010 5**
COCOA	0.851 5±0.012 6	0.950 2±0.008 4	0.863 5±0.023 3	0.745 5±0.011 2
ML-kNN	0.846 8±0.015 3	0.943 5±0.010 7	0.858 0±0.026 9	0.760 7±0.008 4
BP-MLL	0.825 2±0.015 3	0.838 0±0.025 5	0.851 7±0.028 2	0.757 8±0.001 0
IBLR	0.847 9±0.016 0	0.948 3±0.009 7	0.869 2±0.023 8	0.676 3±0.008 6
ECC	0.825 6±0.017 6	0.934 7±0.011 1	0.851 1±0.015 8	0.729 7±0.003 6
CLR	0.822 0±0.015 3	0.911 7±0.010 2	0.827 9±0.021 6	0.761 4±0.003 2
RAKEL	0.700 5±0.019 4	0.750 6±0.005 4	0.708 9±0.015 8	0.599 8±0.008 2
平均差值/%	3.9	6.0	4.7	4.9

3 结 论

为进一步提高多标签学习中的预测效果,本文提出一种改进多标签学习的算法,针对多标签中标签不平衡问题,考量标签与标签之间相互关联性,提出了一种二类不平衡分类器,与多类不平衡分类器结合的耦合分类器. 为验证本文提出的算法的有效性,在 4 种不同类型的数据集上进行不同评价标准的测评. 实验结果表明,本文提出的算法比其他 7 种对比算法优势明显. 下一阶段将围绕如何解决在线多标签学习中标签不均衡的问题开展研究.

参考文献

[1] ZHANG Minling, ZHOU Zhihua. A review on multi-label learning algorithms [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26 (8): 1819. DOI: 10.1109/TKDE.2013.39

[2] UEDA N, SAITO K. Parametric mixture models for multi-label text

[C]//Proceedings of the 15th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2002: 721

[3] ZHU Zijiang, CHEN Hang, HU Yi, et al. Age estimation algorithm of facial images based on multi-label sorting [J]. EURASIP Journal on Image and Video Processing. 2018, 2018(1): 114

[4] WANG Xiao, ZHANG Weiwei, ZHANG Qiuwe, et al. MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier [J]. Bioinformatics, 2015, 31(16): 2639

[5] AGRAWAL R, GUPTA A, PRABHU Y, et al. Multi-label learning with millions of labels: recommending advertiser bid phrases for web pages [C]//WWW'13 Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro: ACM, 2013: 13

[6] CHEN Chunling, TSENG F S C, LIANG T, et al. Editorial: an integration of WordNet and fuzzy association rule mining for multi-label document clustering [J]. Data & Knowledge Engineering, 2010, 69(11):1208

[7] MA Haiping, CHEN Enhong, XU Linli, et al. Capturing correlations of multiple labels: a generative probabilistic model for multi-label learning [J]. Neurocomputing, 2012, 96(1): 116

[8] WU Yu, WU Wei, ZHANG Xiang, et al. Improving recommendation of tail tags for questions in community question answering [C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 3066

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

[10] FURNKRANZ J, HULLERMEIER E, MENCIA E L, et al. Multi-label classification via calibrated label ranking [J]. Machine Learning, 2008, 73(2):133

[11] TSOUMAKAS G, VLAHAVAS I. Random k-labelsets: an ensemble method for multi-label classification [C]//Proceedings of the 18th European Conference on Machine Learning. Berlin Heidelberg: Springer Berlin Heidelberg, 2007:406

[12] ZHANG Minling, ZHOU Zhihua. Ml-kNN: a lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7):2038

[13] WIECZORKOWSKA A, SYNAK P, RAS Z. Multi-label classification of emotions in music [C]//Proceedings of the 2006 International Conference on Intelligent Information Processing and Web Mining (IIPWM). Berlin Heidelberg: Springer Berlin Heidelberg, 2006:307

[14] LUO Xiao, ZINCIR A N. Evaluation of two systems on multi-class multi-label document classification [C]// Proceedings of the 15th International Symposium on Methodologies for Intelligent Systems (ISMIS). Berlin Heidelberg: Springer-Verlag, 2005:161

[15] XU Jianhua. Multi-label weighted k-nearest neighbor classifier with adaptive weight estimation [C]// Proceedings of the 18th international conference on Neural Information Processing Volume Part II. Berlin Heidelberg: Springer Berlin Heidelberg, 2011:79

[16] THABTAH F, COWLING P, PENG Y. MMAC: a new multi-class, multi-label associative classification approach [C]//Proceedings of the 4th IEEE International Conference on Data Mining (ICDM). Brighton: IEEE, 2004:217

[17] XU Xinshun, JIANG Yuanp, PENG Liang, et al. Ensemble approach based on conditional random field for multi-label image and video annotation [C]//Proceedings of the 19th ACM international conference on Multimedia. Scottsdale: ACM, 2011:1377

[18] GHAMRAWI N, MCCALLUM A. Collective multi-label classification [C]//Proceedings of the 2005ACM Conference on Information and Knowledge Management (CIKM). Bremen Germany: ACM, 2005:195

[19] ZHANG Minling, ZHOU Zhihua. Multi-label neural networks with applications to functional genomics and text categorization [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1338

[20] TSOUMAKAS G, ZHANG Minling, ZHOU Zhihua. Learning from multi-label data [Z]. Bled Slovenia: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2009

[21] HE Haibo, GARCIA E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9):1263

[22] NITESH V C, KEVIN W B, LAWRENCE O H, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321

[23] LI Sinan, LI Ning, LI Zhanhuai. Multi-label data mining: a survey [J]. Computer Science, 2013, 40(4):14

[24] ZHANG Minling, LI Yukun, LIU Xuying. Towards class-imbalance aware multi-label learning [C]//Proceedings of the 24th International Joint Conference on Artificial Intelligence IJCAI 2015. Buenos Aires: AAAI Press, 2015: 4041

[25] CHENG Weiwei, HÜLLERMEIER E. Combining instance-based learning and logistic regression for multilabel classification [J]. Machine Learning. 2009, 76(2): 211

（编辑　王小唯）