# 基于粒子群的多标记阈值自适应极限学习机

许二戗,于化龙
(江苏科技大学 计算机学院,江苏 镇江 212003)

**摘 要**:多标记学习考虑单个样例与多个类别标记相关联的情况,类别不平衡主要研究样本不均衡带给算法的影响,两者均是当前机器学习研究领域的热点。在多标记数据集中普遍存在类别不平衡现象,虽然目前已经提出了大量的多标记学习,但对于数据集的内在特点却鲜有研究。针对这一问题,提出了一种基于粒子群的多标记阈值自适应极限学习机算法(MLTA-ELM)。该算法充分结合了极限学习机学习速度快、泛化性能好的优点及类别不平衡学习中的阈值自适应选择策略。首先利用极限学习机构建一个单隐层前馈神经网络模型,其次利用该模型实现多标记初步预测,然后采用粒子群优化算法作为阈值自适应选择策略,以此获得判断标记类别的最优阈值组合。最后,通过 12 个基准的多标记数据集,对 MLTA-ELM 算法的可行性及有效性进行了验证。实验结果表明,该算法与其他几种流行的方法相比,具有更好的预测能力。

**关键词**:多标记分类;类别不平衡;粒子群优化;极限学习机;阈值技术
**中图分类号**:TP181 **文献标识码**:A **文章编号**:1673-629X(2019)04-0047-06
doi:10.3969/j.issn.1673-629X.2019.04.010

# An Extreme Learning Machine of Multi-label Threshold Adaptation Based on Particle Swarm Optimization

XU Er-qiang, YU Hua-long
(School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

**Abstract**: Multi-label learning investigates the case of single object related to multiple labels, while class imbalanced learning mainly studies the impact of unbalancedly distrubuted samples on the algorithm. Both of them are the hot spots in the field of machine learning research. Class imbalance is the common phenomenon in multi-label datasets. Though a large number of multi label learning algorithms have been put forward, there is little research on the intrinsic characteristics of dataset. To address the problem, we present a PSO-based multi-label threshold adaptation extreme learning machine (MLTA-ELM). This algorithm fully combines the advantages of extreme learning machine such as fast learning speed, strong generalization and the adaptive selection strategy of threshold value in class unbalance learning. First, a single hidden layer feed forward neural network is built by extreme learning machine, and the multi labels are predicted preliminarily by this model. Then the particle swarm optimization algorithm is taken as the threshold adaptive selection strategy to obtain the optimal threshold combination for label prediction. Lastly, we conduct experiments on 12 baseline multi-label datasets to verify the feasibility and effectiveness of the proposed algorithm. The experiment indicates that the proposed algorithm outperforms several state-of-the-art ones.

**Key words**: multi-label classification; class imbalance; particle swarm optimization; extreme learning machine; threshold technique

## 0 引 言

众所周知,在传统的监督学习框架中,数据集中的每个样本通常只关联于一个标记,但在现实应用场景中,一个样本则通常可能关联多个标记,此种类型数据被称为多标记数据。在近十几年中,多标记学习已逐渐发展成为机器学习领域的研究热点之一,吸引了大量研究者的关注,并在多媒体内容自动标注[1]、信息检索[2]、个性化推荐[3]、生物信息学[4]等多个领域得到了实际的应用。

在多标记数据中,普遍存在着类别不平衡的现象,其表现为在绝大多数或全部标记中的正类样本个数远少于负类样本个数。类别不平衡问题往往会导致所训

**收稿日期**:2018-06-13 **修回日期**:2018-10-16 **网络出版时间**:2018-12-20
**基金项目**:国家自然科学基金(61305058,61572242);中国博士后特别资助计划项目(2015T80481);中国博士后科学基金(2013M540404);江苏省自然科学基金(BK20130471);江苏省博士后基金(1401037B)
**作者简介**:许二戗(1987-),男,硕士研究生,研究方向为机器学习、数据挖掘;于化龙,博士,副教授,研究方向为机器学习、数据挖掘。

练的分类超平面产生严重偏倚,从而降低多标记算法的最终分类性能。为解决上述问题,Charte 等[5-6]将单标记类别不平衡学习中的 ROS、RUS 及 SMOTE 等采样技术扩展到多标记数据中,分别提出了 ML-ROS、ML-RUS、ML-SMOTE 等算法;Zhang 等[7]则在算法层面进行了改进,通过结合样本相关性及集成学习技术提出了 COCOA 算法。但上述算法仍存在着分类性能差或时间复杂度高等诸多缺点。

极限学习机(extreme learning machine,ELM)是 2006 年黄广斌等[8]提出的一种单隐层前馈神经网络训练方法,具有训练速度快、泛化性能好等优点。ELM 在回归、聚类、二类分类和多类分类等领域都有不错的表现[9],但目前在多标记领域的应用仍相对较少,同时也未考虑到多标记数据中的不平衡现象。

鉴于 ELM 技术的诸多优点,拟结合其与类别不平衡学习中常用的阈值选择技术,提出一种适用于多标记不平衡数据的自适应阈值极限学习机(PSO-based multi-label threshold adaptation extreme learning machine,MLTA-ELM)算法。首先,该算法通过建立 ELM 模型来获得样本标记的预测输出值;然后,选定合适的阈值组合对其进行标记判别。在进行阈值选择时,原有问题转化为一个多变量优化问题,故文中利用粒子群优化算法作为阈值选择器。当然,也可以尝试采用其他随机优化算法来替换 PSO 算法。最后,利用 12 个基准的多标记数据集对该算法的性能进行了验证,并与 5 种基准或流行的算法进行了比较。

## 1 相关工作

### 1.1 多标记中的类别不平衡问题

在多标记学习领域中,已存在多种成熟的算法,如 ML-KNN[10]、IMLLA[11]、BP-MLL[12]、RAkEL[13] 等,但大多算法仍主要关注于如何挖掘标记间的相关性,而忽略了多标记数据中往往存在类别不平衡问题这一特点。因此,下面将以标记密度与不平衡比率这两个评价指标来简单介绍多标记数据中存在的类别不平衡问题。

标记基数(label card,LCard)表示每个样本所对应正类标的均数,而标记密度(label density,LDen)则表示每个样本所对应正类标在所有类标中所占的比例,如一个多标记数据集的 LDen 测度值为 0.2,则表示每 10 个类标中平均有 2 个被标记为正类,上述测度可通过如下两个公式计算得出:

$$\mathrm{LCard}(D)=\frac{1}{N}\sum_{i=1}^{N}|Y_i==1| \tag{1}$$

$$\mathrm{LDen}(D)=\frac{1}{|y|}\cdot\mathrm{LCard}(D) \tag{2}$$

其中,$N$ 表示样本数;$|Y_i==1|$ 表示第 $i$ 个样本所对应类标被标记为 1 的数量;$|y|$ 表示类标的个数。

表 1 统计了在后续实验中使用的 12 个数据集的特征信息,从中可以看出:仅有 flags 数据集的标记密度接近 0.5,其余的均在 0.33 以下,且大部分在 0.2 左右。这说明多标记数据集中的正类标记所占比例均相对较低。

表 1 所用数据集及其不平衡测度

| 数据集 | 样本数 | 特征数 | 类标数 | LDen | $\mathrm{ImR}_{\mathrm{avg}}$ |
|---|---|---|---|---|---|
| image | 600 | 294 | 5 | 0.247 | 3.154 |
| emotions | 593 | 72 | 6 | 0.311 | 2.320 |
| scene | 2 407 | 294 | 6 | 0.179 | 4.662 |
| flags | 194 | 19 | 7 | 0.485 | 2.753 |
| yeast | 2 417 | 103 | 14 | 0.303 | 8.954 |
| birds | 645 | 260 | 19 | 0.053 | 32.859 |
| tmc2007 | 28 596 | 500 | 22 | 0.101 | 27.996 |
| mirflickr | 25 000 | 150 | 24 | 0.155 | 15.285 |
| genbase | 662 | 1 186 | 27 | 0.046 | 143.458 |
| llog | 1 460 | 1 004 | 75 | 0.212 | 6.844 |
| bibtex | 7 395 | 1 836 | 159 | 0.015 | 87.699 |
| cal500 | 502 | 68 | 174 | 0.150 | 22.345 |

除标记密度外,不平衡比率(imbalance ratio,ImR)也是用于描述类别不平衡程度的一项重要指标[7],可表示为负类数与正类数的比值。针对第 $j$ 个类标,用 $D_j^+$ 表示标记为正类的样本数,$D_j^-$ 表示标记为负类的样本数,对应的 ImR 可由下式计算得出:

$$\mathrm{ImR}_j=\max(|D_j^+|,|D_j^-|)/\min(|D_j^+|,|D_j^-|) \tag{3}$$

对于多标记数据集,不平衡比率的算术平均值 $\mathrm{ImR}_{\mathrm{avg}}$ 能够直观地反映出其类别偏倚的程度。从表 1 可明显看出,所有数据集的不平衡比率均处于 2.2 ~ 143 之间,其中 8 个数据集不平衡比率大于 5,6 个数据集的不平衡比率在 10 以上。总体而言,类别不平衡普遍存在于多标记数据中,且类标越多,极度不平衡现象出现的可能性也通常越高。

### 1.2 极限学习机

极限学习机是一种单隐层前馈神经网络(single hidden-layer feedback network,SLFN)训练方法[8]。其完全摒弃了传统的迭代误差调整策略,改为随机设置隐层权重与偏置,然后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。

不妨假设训练集包含 $N$ 个样本,且这些样本能被分入 $m$ 个类中,第 $i$ 个训练样本表示为 $(\boldsymbol{x}_i,\boldsymbol{t}_i)$,其中 $\boldsymbol{x}_i$ 是一个 $n$ 维的输入向量,而 $t_i$ 则对应于一个 $m$ 维的输出向量。另假设 ELM 中包括 L 个隐藏层节点,该层上的权重 $w$ 与偏置 $b$ 在[-1,1]区间完全随机生成,

那么对于样本 $x_i$,其对应的隐藏层输出可以表示为一个行向量 $h(x_i)=[h_1(x_i),h_2(x_i),\cdots,h_L(x_i)]$。ELM 的数学模型可以表示为:

$$\boldsymbol{H\beta}=\boldsymbol{T} \tag{4}$$

其中,$\boldsymbol{H}=[h(x_1),h(x_2),\cdots,h(x_N)]^{\mathrm{T}}$ 为所有样本对应的隐藏层输出矩阵;$\boldsymbol{\beta}$ 为待求解的输出层权重矩阵;$\boldsymbol{T}=[t_1,t_2,\cdots,t_N]$ 为样本类标所对应的期望输出矩阵。

利用最小二乘法,$\boldsymbol{\beta}$ 可通过下式进行求解:

$$\boldsymbol{\beta}=\boldsymbol{H}^{\mathrm{T}}=\begin{cases}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{HH}^{\mathrm{T}})^{-1}\boldsymbol{T}, & \text{when } N\leqslant L\\ (\boldsymbol{HH}^{\mathrm{T}})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{T}, & \text{when } N>L\end{cases} \tag{5}$$

其中,$\boldsymbol{H}$ 为 $\boldsymbol{H}$ 的 Moore-Penrose 广义逆,可以保证所求得解为式 4 的最小范数最小二乘解。因此,极限学习机可通过一步计算得到,无需迭代,使得训练时间大幅缩短。

也可从优化角度来描述和求解 ELM。为最小化训练误差且同时提升模型的泛化能力,需同时对 $\|\boldsymbol{H\beta}-\boldsymbol{T}\|^2$ 和 $\|\boldsymbol{\beta}\|^2$ 做最小化处理,故该问题可描述为如下形式:

$$\begin{cases}\text{Minimize: } \mathrm{Lp}_{\mathrm{ELM}}=\dfrac{1}{2}\|\boldsymbol{\beta}\|^2+C\dfrac{1}{2}\sum\limits_{i=1}^{N}\|\boldsymbol{\xi}_i\|^2\\ \text{Subject to: } \boldsymbol{h}(x_i)\boldsymbol{\beta}=\boldsymbol{t}_i^{\mathrm{T}}-\boldsymbol{\xi}_i^{\mathrm{T}},i=1,2,\cdots,N\end{cases} \tag{6}$$

其中,$\boldsymbol{\xi}_i=[\xi_{i,1},\xi_{i,2},\cdots,\xi_{i,m}]$ 表示样本 $\boldsymbol{x}_i$ 在所有输出节点上对应的训练误差向量;$C$ 表示惩罚因子,用于调控模型训练准确性与泛化性二者之间的均衡关系。

式 6 可通过求解得到,给定一个具体样例 $x$,其对应的实际输出向量可由下式求得:

$$\boldsymbol{f}(x)=\begin{cases}h(x)\boldsymbol{H}^{\mathrm{T}}\left(\dfrac{\boldsymbol{I}}{C}+\boldsymbol{HH}^{\mathrm{T}}\right)^{-1}\boldsymbol{T}, & \text{when } N<L\\ h(x)\left(\dfrac{\boldsymbol{I}}{C}+\boldsymbol{HH}^{\mathrm{T}}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{T}, & \text{when } N\geqslant L\end{cases} \tag{7}$$

其中,$\boldsymbol{f}(x)=[f_1(x),f_2(x),\cdots,f_m(x)]$ 表示样例 $x$ 的实际输出向量,而该样例的预测类标为向量 $\boldsymbol{f}(x)$ 中元素最大的值对应的类别。

## 2　文中算法

### 2.1　极限学习机的多标记应用

ELM 的网络结构不仅适用于单标记学习,也同样可用于多标记学习[9]。在多标记学习中,式 6、式 7 依然有效,输出节点个数不再代表类别的个数,而是多标记数据类标的个数,即 $m$ 个输出节点代表每个样例关联 $m$ 个标记。

标记判别时,单标记中,单个样例仅关联一个标记,仅需求出输出向量 $f(x)$ 中元素最大值的对应标记即可;而对于多标记问题,单个样本可能关联多个标记,此时,需要设定一个阈值函数 $\mathrm{th}(x)$,并通过下式预测类标:

$$\mathrm{label}(i)=\begin{cases}+1, & \text{when } f_i(x)\geqslant \mathrm{th}(x)\\ -1, & \text{when } f_i(x)<\mathrm{th}(x)\end{cases} \tag{8}$$

因此,阈值函数 $\mathrm{th}(x)$ 的确定成为了解决该问题的关键。

### 2.2　阈值自适应选取策略

类别不平衡问题中常用的阈值选择方式有[14]:根据经验来设定阈值[15],即 $\mathrm{th}(x)$ 等于一个常数 $\theta$;采用优化技术来确定阈值[16],即 $\mathrm{th}(x)$ 等于一个向量 $[\theta_1,\theta_2,\cdots,\theta_m]$。对于多标记分类问题,类标空间维度往往较高,因而阈值选择也会更加困难,故简单的由经验来设定阈值的方式通常不会取得理想的分类效果,所以文中关注如何通过优化技术来设定最优阈值,则问题就转变成了一个多变量的最优化问题。

首先选取不平衡问题的常用性能度量指标 Macro F-measure(Macro-F)为优化目标。首先基于统计量求得在各个类标上的分类性能,然后再将所有类上的测度均值作为最终结果。计算公式如下:

$$\text{Macro-F}=\frac{1}{|y|}\sum_{i=1}^{|y|}\text{F-measure}(\mathrm{TP}_i,\mathrm{FP}_i,\mathrm{TN}_i,\mathrm{FN}_i) \tag{9}$$

其中,$|y|$ 表示类标数。

$$\text{F-measure}=2*\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{10}$$

$$\text{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{11}$$

$$\text{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{12}$$

其中,TP 表示真正类;FP 表示假正类;TN 表示真负类;FN 表示假负类。

其次选用 PSO 粒子群优化算法[17-18]。在 PSO 中,每个粒子有适应性,能够与环境及其他粒子进行交流,并根据交流的过程学习来改变自己的结构与行为,以此达到最优。在 PSO 算法优化过程中,每个粒子通过学习其自身经验(pbest)和种群其他成员的经验(gbest),动态改变各自的位置和速度。其每轮的更新方式如下:

$$\begin{cases}v_{id}^{k+1}=v_{id}^{k}+c_1\times r_1\times(\mathrm{pbest}-x_{id}^{k})+c_2\times r_2\times(\mathrm{gbest}-x_{id}^{k})\\ x_{id}^{k+1}=x_{id}^{k}+v_{id}^{k+1}\end{cases} \tag{13}$$

其中,$v_{id}^{k}$ 和 $v_{id}^{k+1}$ 分别表示第 $i$ 个粒子中第 $d$ 个维度在第 $k$ 与第 $k+1$ 轮次更新的速度;$x_{id}^{k}$ 和 $x_{id}^{k+1}$ 表示其

在这两个轮次所对应的位置；$c_1$ 和 $c_2$ 则是两个非负常量，表示加速因子；$r_1$ 和 $r_2$ 为随机因子，其取值范围在[0,1]之间。在实验中，根据经验将粒子种群个数设置为 20，迭代次数设置为 50，加速因子 $c_1$ 和 $c_2$ 均设置为 1。此外，考虑到 ELM 的实际输出值通常不会过小或过大，因此将 $x$ 的位置限制在[0,2]之间，速度 $v$ 限制在[-1,1]之间。

### 2.3 MLTA-ELM 算法流程

综上所述，下面给出了 MLTA-ELM 算法的整体流程。

输入：多标记训练样本 $S$：$\{(x_i,Y_i) \mid i=1,2,\cdots,n\}$，隐层节点数 $L$，惩罚因子 $C$；

输出：所训练的多标记分类器 MLTA-ELM。

步骤 1：训练多标记的 ELM 分类器。

(1)根据输入节点数，隐层节点数 $L$，惩罚因子 $C$ 与多标记类别个数，随机生成网络模型的隐藏层权重和偏置，设置激活函数为 sigmoid 函数；

(2)在训练集 $S$ 上根据式 6 训练 ELM 分类器 $M$；

(3)获得训练集 $S$ 在模型 $M$ 上的实值输出的矩阵 $f(x)$。

步骤 2：最优阈值组合选取$[\theta_1,\theta_2,\cdots,\theta_m]$。

(1)种群初始化，包括初始位置、速度等；

(2)计算每个微粒的适应度；

(3)计算粒子所经历的最好位置 pbest，并计算群体中所有粒子经历的最好位置；

(4)根据式 13 进行速度和位置更新；

(5)反复执行步骤 2～4，直到达到最大进化迭代次数；

(6)最大适应度对应种群中的位置，即所求最优阈值组合。

步骤 3：标记预测。

对于一个样例 $x$，首先通过步骤 1 获得输出矩阵 $f(x)$，将其与最优阈值组合$[\theta_1,\theta_2,\cdots,\theta_m]$根据式 8 进行比较，获得判别标记。

## 3 实验与结果分析

### 3.1 数据集与实验设置

实验主要在 12 个基准的多标记数据集上完成，这些数据集涵盖了文本、音频、生物等不同场景。各数据集具有不同的样本数、类标数、标记密度及不平衡比率。有关这些数据集的具体信息见表 1。

硬件环境：Intel 酷睿 i7-555U 处理器，CPU 主频 3.1 GHz，内存 8 GB，硬盘 1 TB，操作系统为 Windows 8.1；编程环境为 Matlab2015b。

为验证提出算法的有效性与优越性，将其与几种经典的多标记不平衡分类算法进行实验比较，比较算法包括 COCOA[7]、ML-SMOTE[5]、ML-ROS[6]、ML-RUS[6]以及标准 ELM 等。各类算法所特有的参数均按照代码中的原始最优设置而设定。COCOA 算法中特有的参数 $K$，ML-ROS、ML-RUS 中的特有参数 $P$，根据对应参考文献分别设置如下：$K=\min(q-1,10)$，$P=10\%$。在标准的 ELM 算法中，各类标对应的阈值均为缺省值 0。同时，为了保证实验的公正性，除 COCOA 采用对应文献自带的分类器外，其他算法均采用 ELM 作为基分类器。ELM 算法中的两个参数，隐层节点数 $L$ 及惩罚因子 $C$，则通过内部五折交叉验证的 grid search 方法进行选取，选取范围为：$L \in \{50,100,\cdots,1\,000\}$，$C \in \{2^1,2^2,\cdots,2^{20}\}$。此外，考虑到实验中各种算法均存在一定的随机性，故实验结果以 50 次随机 5 折交叉验证所计算得到的均值形式给出。性能测度指标分别采用 Macro F-measure (Macro-F)及 Micro F-measure (Micro-F)。

### 3.2 结果与讨论

表 2 及表 3 分别给出了各算法在各个数据集上的 Macro-F 及 Micro-F 性能测度值。

表 2 各算法在各数据集上的 Macro-F 结果

| 数据集 | MLTA-ELM | COCOA | ML-SMOTE | ML-ROS | ML-RUS | ELM |
|---|---|---|---|---|---|---|
| image1 | 0.595 5 | 0.534 1 | 0.509 1 | 0.494 4 | 0.514 9 | 0.504 6 |
| emotions | 0.668 7 | 0.667 2 | 0.612 2 | 0.593 5 | 0.600 9 | 0.609 2 |
| scene | 0.729 7 | 0.716 7 | 0.657 6 | 0.651 3 | 0.651 2 | 0.656 7 |
| flags | 0.678 7 | 0.672 3 | 0.645 6 | 0.598 1 | 0.608 1 | 0.618 0 |
| yeast | 0.471 7 | 0.437 3 | 0.320 3 | 0.288 2 | 0.310 6 | 0.318 7 |
| birds | 0.388 9 | 0.418 2 | 0.360 8 | 0.349 8 | 0.319 1 | 0.331 1 |
| tmc2007 | 0.540 5 | 0.690 1 | 0.378 6 | 0.348 3 | 0.319 7 | 0.339 1 |
| mirflickr | 0.273 2 | 0.162 7 | 0.096 0 | 0.050 0 | 0.087 1 | 0.092 6 |
| genbase | 0.735 4 | 0.847 8 | 0.722 6 | 0.682 5 | 0.609 6 | 0.684 0 |
| llog | 0.379 4 | 0.418 1 | 0.181 1 | 0.182 6 | 0.223 2 | 0.221 0 |
| bibtex | 0.227 0 | 0.196 4 | 0.091 5 | 0.066 1 | 0.068 8 | 0.073 4 |
| cal500 | 0.212 0 | 0.169 4 | 0.111 0 | 0.086 4 | 0.104 1 | 0.104 9 |

表 3　各算法在各数据集上的 Micro-F 结果

| 数据集 | MLTA-ELM | COCOA | ML-SMOTE | ML-ROS | ML-RUS | ELM |
|---|---|---|---|---|---|---|
| image1 | 0.594 1 | 0.552 5 | 0.511 3 | 0.498 6 | 0.518 8 | 0.508 8 |
| emotions | 0.674 6 | 0.685 0 | 0.636 3 | 0.611 7 | 0.618 2 | 0.633 4 |
| scene | 0.715 6 | 0.709 3 | 0.643 1 | 0.638 8 | 0.638 1 | 0.644 2 |
| flags | 0.739 7 | 0.733 9 | 0.725 2 | 0.704 8 | 0.715 7 | 0.721 2 |
| yeast | 0.638 3 | 0.650 8 | 0.617 9 | 0.575 3 | 0.609 0 | 0.616 5 |
| birds | 0.479 9 | 0.547 0 | 0.473 6 | 0.457 2 | 0.435 6 | 0.443 9 |
| tmc2007 | 0.672 0 | 0.778 1 | 0.636 5 | 0.614 6 | 0.628 1 | 0.638 4 |
| mirflickr | 0.402 0 | 0.311 6 | 0.234 6 | 0.125 7 | 0.224 5 | 0.236 1 |
| genbase | 0.958 1 | 0.949 7 | 0.940 2 | 0.928 9 | 0.914 8 | 0.934 6 |
| llog | 0.515 0 | 0.587 7 | 0.440 2 | 0.418 4 | 0.475 2 | 0.469 4 |
| bibtex | 0.311 7 | 0.339 8 | 0.266 4 | 0.245 0 | 0.253 0 | 0.259 3 |
| cal500 | 0.437 5 | 0.373 9 | 0.379 9 | 0.318 7 | 0.363 8 | 0.365 6 |

从这些实验结果中,可以得出如下结论:

(1)从两种性能测度的结果来看,无论采用采样技术、集成学习技术还是文中采用的阈值技术,均可或多或少地缓解样本不平衡分布对分类器性能所产生的负面影响。这一结论主要体现在各类算法与基准 ELM 分类器的结果比较上。

(2)在几乎全部数据集上,MLTA-ELM 与 COCOA 算法均显著优于 ML-SMOTE、ML-ROS 及 ML-RUS 算法。究其原因,前两种算法属于算法适应型,其在算法模型上进行了针对性的改动以适应多标记数据中的不平衡现象,而后三种算法则采用了采样的策略,是立足于通过调整数据分布以弥补数据的不平衡分布,具有一定的随机性,同时也容易出现过拟合与欠拟合的现象。

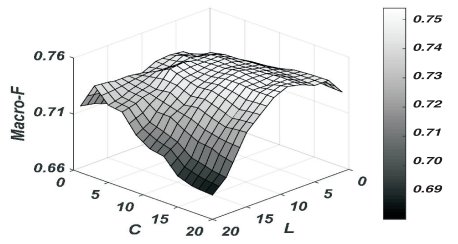
(3)相较于 ML-ROS 与 ML-RUS,ML-SMOTE 算法在绝大数据集上都有不同程度的性能提升,这是因为该算法不再简单地对少数类样本进行复制,而是通过一定策略生成大量新样本的方式来谋求训练样本集类分布的平衡,因此采样结果更具泛化性。这一结论也可通过比较 ML-ROS、ML-RUS 与基准 ELM 算法的结果而得出:在不平衡比率较大的数据集上,ROS 与 RUS 算法的性能往往低于基准 ELM 算法,而 ML-SMOTE 相较于基准 ELM 算法则通常会有一定的性能提升,这也再次证明了对多标记数据进行随机采样往往会造成过适应,而 ML-SMOTE 算法则可有效规避该问题。

(4)与除 COCOA 算法外的其他多标记不平衡学习算法相比,MLTA-ELM 算法在性能上均有较大幅度的提升。具体而言,在两个性能测度上,MLTA-ELM 算法分别在 8 个和 6 个数据集上获得了最优的性能,充分说明了 MLTA-ELM 算法能够根据不同的数据分布自适应地选择最优阈值组合。至于为何其在 Marco-F 测度上效果要更优,相信原因在于 PSO 是以该测度为寻优目标相关。

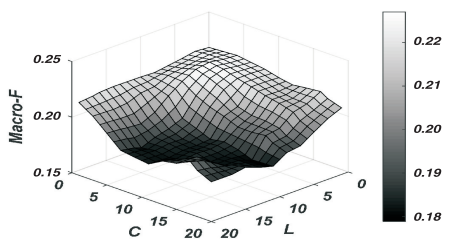
(5)相比于 COCOA 算法,文中算法并未体现出显著的优势。究其原因,不难发现:COCOA 算法利用了标记间的相关性信息;COCOA 算法采用了集成学习模式来提升分类模型的泛化性与分类性能,而这也是文中算法所欠缺的。当然,在实验中也发现,文中算法的时间开销往往远小于 COCOA 算法,尤其在类标规模较大的数据集上,这一优势通常会体现得更加明显。

### 3.3　参数分析

最后,分析参数对模型的重要程度。选取了标记小于 10 的数据集 scene 和标记大于 100 的数据集 cal500。通过实验获取了不同参数 $L$、$C$ 时对应的模型指标 Macro-F。




(a) scene



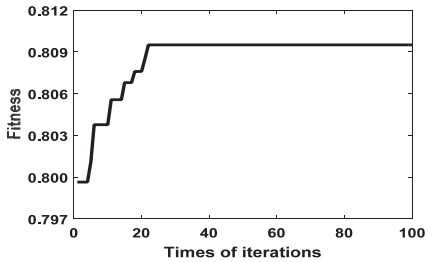

(b) cal500

图 1　不同 $L$ 与 $C$ 下的 Macro-F

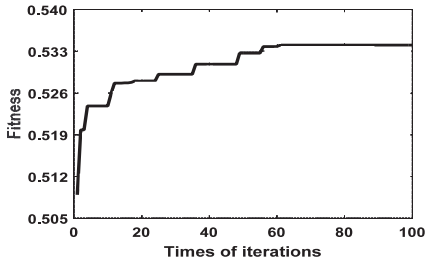
由图 1 可见,在不同参数 $L$、$C$ 下,其结果会随着参数的变化而较为平滑地上升或下降。可以看出,两个数据集中,在选定的参数范围内,均存在最小值与最大值,且最大值不处于边缘状态,也就是说,该参数范

围是包含了最大值范畴的,也证明了该参数范围是有效的。

此外,实验分析了粒子群算法迭代次数与标记个数的关系,理论上,标记数的大小,表明标记空间维度的大小,在高维空间中搜索的范围会更大,需要的迭代次数也越多。通过图 2 可以看出,在 scene 与 cal500 上的收敛迭代次数分别为 20 多次与 60 多次。由此可以得出,标记数越大,其迭代次数会越大。




(a)scene 标记数 6




(b)cal500 标记数 174

图 2 粒子群算法 100 次迭代过程的适应度变化曲线

## 4 结束语

针对多标记数据中广泛存在的类别不平衡问题,提出了一种基于粒子群的多标记自适应阈值极限学习机(MLTA-ELM)算法。该算法以 Macro F-measure 为优化目标,将多标记阈值选择问题转化为一个多维连续空间的优化问题,并通过粒子群优化算法进行求解,以自适应地构建较优的多标记分类模型。在 12 个多标记数据集上的实验结果表明,与诸多同类算法相比,该算法极大地提升了多标记分类的性能,可以满足各种实际应用的需求。但该算法未考虑类标间的相关性,若将该信息融合进分类模型,相信可以进一步提升分类性能;由于引入了随机优化过程,故该算法的时间复杂度仍然较高。对于这些问题,该算法还有待进一步的改进。

**参考文献:**

[1] 吴 磊,张敏灵. 基于类属属性的多标记学习算法[J]. 软件学报,2014,25(9):1992-2001.

[2] 付 博,刘 挺. 社会媒体中用户的隐式消费意图识别[J]. 软件学报,2016,27(11):2843-2854.

[3] 纪亚亮,郑 啸. 慢性乙型肝炎用药推荐系统的设计与实现[J]. 医疗卫生装备,2017,38(7):48-51.

[4] 胡海峰,郑 茂,吴伟坚,等. 基于多示例多标记迁移学习的蛋白质功能预测[J]. 中国科学:信息科学,2017,47(11):1538-1550.

[5] CHARTE F,RIVERA A J,JESUS M J D,et al. MLSMOTE:approaching imbalanced multilabel learning through synthetic instance generation[J]. Knowledge-Based Systems,2015,89:385-397.

[6] CHARTE F,RIVERA A J,JESUS M J D,et al. Addressing imbalance in multilabel classification:measures and random resampling algorithms[J]. Neurocomputing,2015,163:3-16.

[7] ZHANG Minling, LI Yukun, LIU Xuying. Towards class-imbalance aware multi-label learning[C]//International conference on artificial intelligence. Buenos Aires, Argentina:AAAI Press,2014:4041-4047.

[8] HUANG Guangbin,ZHU Qinyu,SIEW C K. Extreme learning machine:theory and applications[J]. Neurocomputing,2006,70(1-3):489-501.

[9] HUANG Guangbin, WANG Dianhui, LAN Yuan. Extreme learning machines:a survey[J]. International Journal of Machine Learning & Cybernetics,2011,2(2):107-122.

[10] ZHANG Minling,ZHOU Zhihua. ML-KNN:a lazy learning approach to multi-label learning[J]. Pattern Recognition,2007,40(7):2038-2048.

[11] 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.

[12] ZHANG Minling,ZHOU Zhihua. Multilabel neural networks with applications to functional genomics and text categorization[J]. IEEE Transactions on Knowledge & Data Engineering,2006,18(10):1338-1351.

[13] TSOUMAKAS G, VLAHAVAS I. Random k-labelsets:an ensemble method for multilabel classification[C]//European conference on machine learning. [s. l. ]:[s. n. ],2007:406-417.

[14] 于化龙. 类别不平衡学习:理论与算法[M]. 北京:清华大学出版社,2017.

[15] ZHOU Zhihua, LIU Xuying. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Transactions on Knowledge & Data Engineering,2006,18(1):63-77.

[16] YU Hualong, SUN Changyin, YANG Xibei, et al. ODOC-ELM:optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data[J]. Knowledge-Based Systems,2016,92:55-70.

[17] KENNEDY J, EBERHART R. Particle swarm optimization[C]//IEEE international conference on neural networks. [s. l. ]:IEEE,1995:1942-1948.

[18] 纪 震. 粒子群算法及应用(计算机理论基础与应用丛书)[M]. 北京:科学出版社,2009.