

用于多标记学习的分类器圈方法^{*}

王少博, 李宇峰

(计算机软件新技术国家重点实验室(南京大学), 江苏 南京 210023)

通讯作者: 李宇峰, E-mail: liyf@nju.edu.cn, http://lamda.nju.edu.cn/liyf/

摘要: 如何利用标记间关系来提高学习性能,是多标记学习领域的一个重要问题.分类器链方法及其变型是解决这类问题的一个有效技术.然而,它的学习过程需要预先给定标记的学习次序,这个信息真实情况难以获得.次序选择不当会导致学习性能提高受限.针对这个问题,提出用于多标记学习的分类器圈方法.该方法随机生成标记的学习次序,通过圈结构依次迭代地更新每个标记的分类器.实验结果表明,该方法在多个数据集上取得了比分类器链方法以及一系列经典多标记学习方法更好的性能.

关键词: 多标记学习;标记关系;分类器圈;分类器链

中图法分类号: TP181

中文引用格式: 王少博,李宇峰.用于多标记学习的分类器圈方法.软件学报,2015,26(11):2811-2819. <http://www.jos.org.cn/1000-9825/4908.htm>

英文引用格式: Wang SB, Li YF. Classifier circle method for multi-label learning. Ruan Jian Xue Bao/Journal of Software, 2015, 26(11):2811-2819 (in Chinese). <http://www.jos.org.cn/1000-9825/4908.htm>

Classifier Circle Method for Multi-Label Learning

WANG Shao-Bo, LI Yu-Feng

(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China)

Abstract: Exploiting label relationship to help improve learning performance is important for multi-label learning. The classifier chain method and its variants are shown to be a powerful solution to such a problem. However, its learning process requires the ordering of labels, which is hard to obtain in real-world situations, and incorrect label ordering may cause a suboptimal performance. To overcome the drawback, this paper presents a classifier circle method for multi-label learning. It initializes the label ordering randomly, and then subsequently and iteratively updates the classifier for each label by connecting the labels as a circle. Experimental results on a number of data sets show that the proposal outperforms classifier chains method as well as many state-of-the-art multi-label methods.

Key words: multi-label learning; label relationship; classifier circle; classifier chain

传统监督学习假定每个学习对象只隶属于一个概念标记^[1],然而在现实问题中,学习对象通常可同时与多个概念标记相关.例如,一篇报道亚洲杯的新闻可能既与“体育”相关,又与“经济”、“博弈”等相关;一幅摄于巴黎的自然场景照片可能既与“铁塔”相关,又与“人群”、“花鸟”等相关.多标记学习(multi-label learning)^[2,3]对多标记对象进行学习是这方面的一个主流研究领域,目前已在文本分类^[4,5]、图像标注^[6]、生物信息学^[7-9]、个性化推荐^[10,11]等很多领域取得了应用.

多标记对象的标记存在相关性,否则,多个标记不会同时隶属于同一个对象.事实上,标记之间关系的发现与利用是多标记学习领域的一个关键问题^[12].理由主要有两个方面:一是当处理多标记对象时,预测的输出结果不再是一个二元值(+1 或-1)而变为标记的集合,因此,输出结果的数量与标记数量呈指数增长,必须要通过发现

* 基金项目: 国家自然科学基金(61403186); 国家高技术研究发展计划(863)(2015AA015406); 江苏省基础研究计划(自然科学基金)(BK20140613)

收稿时间: 2015-06-08; 修改时间: 2015-07-14; 定稿时间: 2015-08-26

和利用标记关系来缩减输出结果的空间;二是若不考虑标记关系,将多个标记视为独立的成分进行分类,则将失去丰富的标记关系信息,性能通常不够理想.

分类器链方法(classifier chains method)^[13,14]及其变型是利用标记关系方面的一类有效技术,它将标记按照某种次序排成链,然后依次对链上的标记构建分类器,其中,链后面的标记可利用链前面标记所学习到的分类结果.虽然分类器链方法实现简单,但已在很多文献的实验报道中^[12-15]指出,分类器链方法利用标记关系取得了比不用标记关系的方法更好的性能.然而,分类器链方法的学习过程需要预先给定标记的学习次序,这个信息在真实情况中难以获得.此外,若次序选择不当,分类器链方法或许性能提高有限^[16].一种改进的办法是集成分类器链方法(ensemble classifier chains method)^[13,14].它生成多个标记学习次序,然后对多个次序的结果加以集成.集成分类器链虽然在一定程度上缓解了分类器链性能受限的问题,但是标记次序的数量庞大(与标记数目呈阶乘增长),集成分类器链方法仍面临次序选择的困难,性能仍然受限.本文针对以上问题,提出分类器圈方法(classifier circle method)用于多标记学习.与分类器链方法不同,该方法随机生成标记的学习次序,通过构建分类器圈结构,依次迭代地对每个标记的分类器进行更新.本文方法实现简单.实验结果表明,本文方法在多个数据集、多种不同数量的训练样本设置下,均取得比分类器链方法及一系列经典多标记方法更好的性能.

本文第1节介绍多标记学习及相关工作.第2节给出本文分类器圈方法.第3节汇报实验结果.最后,第4节总结全文.

1 多标记学习及相关工作

由于多标记学习适用于处理很多现实复杂对象,自多标记学习提出以来,得到很多学者的关注,发展出一系列方法.从利用标记关系的视角,这些方法大致可以分为以下3类^[3]:

- 第1类为一阶方法.这类方法不考虑标记之间的关系,将多标记学习问题分解为多个独立的二类分类问题^[6,8,17]来解决.这类方法实现简单,但其忽略标记之间的关系信息,很多时候性能并不理想.
- 第2类为二阶方法.这类方法考虑了标记对之间的关系,如相关标记与非相关标记的排序关系^[7,18,19].这类方法考虑了标记关系,通常可取得比一阶方法更好的性能,但它无法利用更复杂的标记关系.
- 第3类为高阶方法.这类方法通过考虑多个^[14,20]乃至全部^[21,22]标记之间的关系来构建多标记决策函数.这类方法有时可得到更好的学习性能^[15].

从发现标记关系的视角,多标记学习方法也可以分为如下两类:

- 第1类是利用相关领域知识作为先验构造出标记之间的关系^[23-26],如层次关系等.这种方法得到的标记关系往往准确且有助于提高性能,但现实情况中,这种准确的关系通常难以获得.
- 第2类是通过训练数据自动发现标记关系,如统计规律^[27,28]等.这种方法更贴合现实任务的需求,但有时标记关系选择不当会导致性能提高受限.

值得一提的是,分类器链方法^[13,14]是利用标记关系方面的一类简单而有效的高阶多标记学习方法.该方法的思想是:构造一条二类分类器链,链上每个节点对应于一个标记.该方法随后依次对链上的标记构建分类器,其中,链后面的标记可以结合链前面标记的学习结果用于学习当前标记.显然,分类器链方法利用到了标记之间的关系,而且分类器链方法实现简单,在不少实验报道中取得了良好的性能^[12-15].然而,其学习过程需要事先给定标记的学习次序,这个信息现实情况下通常难以准确得到.此外,实验结果表明,学习次序的选择会严重影响最终的学习性能^[16].为了缓解这个困难,Read等人提出了集成分类器链方法^[13,14].该方法构造多条分类器链,其中每条分类器链对应一个随机选取的标记次序.最终,每个标记上的预测结果由多个分类器链的投票结果产生.但标记可能的学习次序数量很大(与标记数目呈阶乘函数增长),集成分类器链方法很难有效遍历所有甚至大部分次序,因此它仍存在因次序选择不当而导致性能受限的情况.为了克服以上困难,本文提出分类器圈方法.

2 本文分类器圈方法

首先给出多标记学习的形式化描述.在多标记学习中,用户输入一组训练样本 $D = \{(x_i, y_i) | i=1, 2, \dots, m\}$,其中, x

为一个训练示例, $\mathbf{y} \in \{+1, -1\}^q$ 为该示例的标记向量(label vector), +1(或-1)分别指对应位置的标记为该示例的相关(或不相关)标记; m 表示训练样本数量, q 为标记总数量. 记 $\mathcal{X} = \mathbb{R}^d$ (或 \mathbb{Z}^d) 为示例输入空间, 多标记学习的目标是: 学到一个决策函数 $h: \mathcal{X} \rightarrow \{+1, -1\}^q$, 使得对于任意未见示例 \mathbf{z} , 预测其相关标记的集合.

本文的分类器圈方法与分类器链方法最大的不同是: 本文方法使用圈结构, 而不是链. 具体来说, 分类器链方法沿着链结构遍历每个标记一次, 分类器圈方法考虑沿着圈结构遍历每个标记若干次. 图 1 分别给出了分类器链方法与本文分类器圈方法的示意图. 直观上, 分类器圈方法沿着标记或分类器的圈反复修正每个标记的分类器, 充分利用到了它与每个标记或分类器之间的关系, 避免分类器链方法次序选择不当时, 链前面的标记难以利用它与链后面标记的关系的情况. 此外, 由于分类器圈方法具有这个优势, 它对学习次序不敏感, 本文方法可随机生成标记次序. 第 3 节给出实验结果说明随机次序与固定次序的性能没有显著差异.

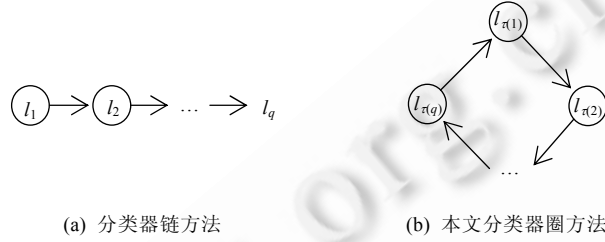


Fig. 1 Illustration for classifier chains method and the proposed classifier circle method

图 1 分类器链方法和本文分类器圈方法示意图

具体来说, 本文方法随机地将 q 个标记 $\{l_1, l_2, \dots, l_q\}$ 连成一个圈, 其中, $l_j (1 \leq j \leq q)$ 表示一个标记. 不失一般性, 记产生的圈为 $\{l_{\tau(1)}, l_{\tau(2)}, \dots, l_{\tau(q)}\}$, 其中, $\tau: \{1, 2, \dots, q\} \rightarrow \{1, 2, \dots, q\}$ 是一个随机排列函数. 本文方法依次对 $\{l_{\tau(1)}, l_{\tau(2)}, \dots, l_{\tau(q)}\}$ 构建二类分类器(如 Logistic 回归、支持向量机等), 该过程循环 T 次, 每次分类器更新利用之前的学习结果. 形式化看, 记 \mathcal{B} 为二类分类学习算法, 沿着标记圈循环 T 次, 得到 $q \times T$ 个二类分类器 $h_{k,j}$, 即

$$h_{k,j} \leftarrow \mathcal{B}(\mathcal{D}_{k,j}), k=1, 2, \dots, T, j=1, 2, \dots, q \quad (1)$$

利用之前的学习结果存在多种方式, 例如, 利用前面所有分类器的结果; 利用前面 q 个分类器的结果(如果已得分类器的数量小于 q , 则利用之前所有分类器的结果). 为了防止过拟合现象, 本文采用后者, 即, 仅利用之前 q 个分类器的结果. 具体来说, 第 k 轮的第 j 个标记上的分类器所用的单标记训练集 $\mathcal{D}_{k,j}$ 按如下方式构造:

$$\mathcal{D}_{k,j} = \{([\mathbf{x}_i, \mathbf{pre}_{k,j}^i], y_i^{\tau(j)}) \mid i=1, 2, \dots, m\} \quad (2)$$

其中, \mathbf{pre} 表示分类器的预测结果:

$$\mathbf{pre}_{k,j}^i = [h_{k,1}([\mathbf{x}_i, \mathbf{pre}_{k,1}^i]), \dots, h_{k,j-1}([\mathbf{x}_i, \mathbf{pre}_{k,j-1}^i]), h_{k,j}([\mathbf{x}_i, \mathbf{pre}_{k,j}^i]), \dots, h_{k-1,q}([\mathbf{x}_i, \mathbf{pre}_{k-1,q}^i])] \quad (3)$$

为简便符号, 记第 0 轮所有标记上的预测结果为空, 即 $h_{0,j}(\mathbf{x}) \equiv 0$, 则对于 $k=1, 2, \dots, T$ 和 $j=1, 2, \dots, q$, 有 $|\mathbf{pre}_{k,j}|=q$. 预测阶段, 对于未见示例 \mathbf{z} , 依下式预测其标记向量:

$$\hat{\mathbf{y}} = \mathbf{h}(\mathbf{z}) = [h_{T,\tau^{-1}(1)}([\mathbf{z}, \mathbf{pre}_{T,\tau^{-1}(1)}]), h_{T,\tau^{-1}(2)}([\mathbf{z}, \mathbf{pre}_{T,\tau^{-1}(2)}]), \dots, h_{T,\tau^{-1}(q)}([\mathbf{z}, \mathbf{pre}_{T,\tau^{-1}(q)}])] \quad (4)$$

算法 1 和算法 2 分别给出本文方法的训练和预测过程.

算法 1. 本文 CCE 方法训练过程.

输入: 训练数据集 $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^d \times \{+1, -1\}^q \mid i=1, 2, \dots, m\}$; 轮数 T .

输出: $q \times T$ 个二类分类器 $h_{k,j}, k=1, 2, \dots, T, j=1, 2, \dots, q$; 排列函数 τ .

1. 初始化 $\mathcal{D}' = \{([\mathbf{x}_i, \mathbf{0}], \mathbf{y}_i) \mid i=1, 2, \dots, m\}$, 其中, $\mathbf{0}$ 是长度为 q 的零向量
2. 随机初始化一个排列函数 $\tau: \{1, 2, \dots, q\} \rightarrow \{1, 2, \dots, q\}$
3. **for** $k \leftarrow 1$ **to** T
4. **for** $j \leftarrow 1$ **to** q

5. $\mathcal{D}'_{k,j} \leftarrow \{(\mathbf{x}'_i, y_i^{\tau(j)}) \mid i=1,2,\dots,m\}$
6. $h_{k,j} \leftarrow \mathcal{B}(\mathcal{D}'_{k,j})$
7. **for all** \mathbf{x}' **in** \mathcal{D}'
8. $\mathbf{x}'^{(d+j)} \leftarrow h_{k,j}(\mathbf{x}')$
9. **end for**
10. **end for**
11. **end for**

算法 2. 本文 CCE 方法预测过程.

输入: $q \times T$ 个二类分类器 $h_{k,j}, k=1,2,\dots,T, j=1,2,\dots,q$; 排列函数 τ ; 待预测示例 $\mathbf{z} \in \mathbb{R}^d$.

输出: 预测结果 $\hat{\mathbf{y}}$.

1. 初始化 $\mathbf{z}' = [\mathbf{z}, \mathbf{0}]$, 其中 $\mathbf{0}$ 是长度为 q 的零向量
2. **for** $k \leftarrow 1$ **to** T
3. **for** $j \leftarrow 1$ **to** q
4. $\mathbf{z}'^{(d+j)} \leftarrow h_{k,j}(\mathbf{z}')$
5. **end for**
6. **end for**
7. $\hat{\mathbf{y}} \leftarrow [\mathbf{z}'^{(d+\tau^{-1}(1))}, \mathbf{z}'^{(d+\tau^{-1}(2))}, \dots, \mathbf{z}'^{(d+\tau^{-1}(q))}]$

分类器圈方法与集成分类器链方法的时间复杂度相当. 具体而言, 分类器圈方法训练和预测的时间复杂度分别为 $\mathcal{O}(T \cdot q \cdot \mathcal{F}_{\mathcal{B}}(m, d+q))$ 和 $\mathcal{O}(T \cdot q \cdot \mathcal{F}'_{\mathcal{B}}(d+q))$, 其中, $\mathcal{F}_{\mathcal{B}}(\cdot, \cdot)$ 和 $\mathcal{F}'_{\mathcal{B}}(\cdot)$ 分别为二类分类器用于训练和测试的时间复杂度; 集成分类器链的训练和预测时间复杂度分别为 $\mathcal{O}(K \cdot q \cdot \mathcal{F}_{\mathcal{B}}(m, d+q))$ 和 $\mathcal{O}(K \cdot q \cdot \mathcal{F}'_{\mathcal{B}}(d+q))$, 其中, K 是集成个数. 实验中, K 的取值与 T 相当. 因此, 与集成分类器链方法相比, 分类器圈方法没有增加时间开销. 但后面的实验结果表明, 分类器圈方法取得了更好的性能.

3 实验结果

3.1 实验数据

本文在 6 个多标记基准数据集上进行实验比较. 这些数据集来自多标记学习不同应用领域: Medical^[29]来自于文本分类, Enron^[30]来自于电子邮件分析, Scene^[6]来自于图片分类, Emotions^[31]来自于音乐标注, Yeast^[7]来自于基因功能预测, Genbase^[32]来自于蛋白质分类. 表 1 给出数据集的统计信息.

Table 1 Experimental datasets

表 1 实验数据集

数据集	示例数目 m	示例维度 d	标记数目 q	标记基数 $LCard$	标记密度 $LDen$	标记多样性 $LDiv$
Emotions	593	72	6	1.869	0.311	27
Enron	1 702	1001	53	3.378	0.064	753
Genbase	662	1186	27	1.252	0.046	32
Medical	978	1449	45	1.245	0.028	94
Scene	2 407	294	6	1.074	0.179	15
Yeast	2 417	103	14	4.237	0.303	198

- $LCard$ 指标记基数(label cardinality), 即, 每个样本相关标记的平均个数:

$$LCard(\mathcal{D}) = \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^q \llbracket y_i^{(j)} = +1 \rrbracket \quad (5)$$

其中, $\llbracket \cdot \rrbracket$ 表示:

$$\llbracket \pi \rrbracket = \begin{cases} 1, & \text{if } \pi \text{ holds} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

- $LDen$ 指标记密度(label density),即,标记基数相对于标记数目的比例:

$$LDen(\mathcal{D}) = \frac{1}{q} LCard(\mathcal{D}) \quad (7)$$

- $LDiv$ 指标记多样性(label diversity),即,样本相关标记集合的总数:

$$LDiv(\mathcal{D}) = |\{y \mid \exists x: (x, y) \in \mathcal{D}\}| \quad (8)$$

3.2 评价指标

本文采用多标记学习领域两个常用的评价指标 F_{macro}^1 和 F_{micro}^1 来衡量方法的预测性能,值越大性能越好.

- F_{macro}^1 指每个标记上 F^1 指标的平均,计算方式如下:

$$F_{macro}^1 = \frac{1}{q} \sum_{j=1}^q \frac{2 \cdot TP_j}{2 \cdot TP_j + FN_j + FP_j} \quad (9)$$

其中,

$$TP_j = |\{\mathbf{x}_i \mid y_i^{(j)} = +1 \wedge h^{(j)}(\mathbf{x}_i) = +1, i = 1, 2, \dots, m\}| \quad (10)$$

$$FN_j = |\{\mathbf{x}_i \mid y_i^{(j)} = +1 \wedge h^{(j)}(\mathbf{x}_i) = -1, i = 1, 2, \dots, m\}| \quad (11)$$

$$FP_j = |\{\mathbf{x}_i \mid y_i^{(j)} = -1 \wedge h^{(j)}(\mathbf{x}_i) = +1, i = 1, 2, \dots, m\}| \quad (12)$$

分别为 true positive(真阳性)、false negative(假阴性)和 false positive(假阳性)样本总数.

- F_{micro}^1 指所有样本在所有标记上的总体 F^1 指标,计算方式如下:

$$F_{micro}^1 = \frac{2 \cdot \sum_{j=1}^q TP_j}{2 \cdot \sum_{j=1}^q TP_j + \sum_{j=1}^q FN_j + \sum_{j=1}^q FP_j} \quad (13)$$

关于 F_{macro}^1 和 F_{micro}^1 的详细介绍可参阅文献[3,20].

3.3 比较方法

本文比较如下多标记方法:

- 1) BR(binary relevant)方法^[2,21]:该方法不考虑标记关系,独立训练每个标记的二类分类器,是一阶方法.
- 2) ML- k NN(multi-label k -nearest neighbor)方法^[17]:该方法拓展 k 近邻方法用于处理多标记学习问题.它没有考虑标记间的关系,属于一阶方法,近邻个数采用文献的推荐参数 $k=10$.
- 3) CLR(calibrated label rank)方法^[18]:该方法考虑相关标记与不相关标记的排序关系,属于二阶方法.
- 4) CC(classifier chains)方法^[13,14],即,分类器链方法:该方法考虑多个标记间的关系,属于高阶方法.
- 5) ECC(ensemble classifiers chains)方法^[13,14],即,集成分类器链方法:与 CC 类似,属于高阶方法.分类器链的集成个数采用文献推荐参数 $K=10$.
- 6) CCE(classifier circle)方法,即,本文的分类器圈方法:本文方法考虑了全部标记之间的关系,属于高阶方法.轮数 T 固定为 5.

以上方法的二类分类器采用线性 Logistic 回归,线性 Logistic 回归采用 Liblinear 工具包^[33]实现.其中, Logistic 回归模型的正则化系数 C 根据 5 折交叉验证确定.

3.4 实验结果

对每个数据集,本文随机选取 80% 样本做训练,余下 20% 做测试.实验重复 30 次,表 2 给出平均结果和标准差,•(○)表示 CCE 方法显著优(劣)于比较方法(成对 t 检验根据 95% 置信度),括号里的数字为对应方法性能序(越小性能越好).表中还给出每个方法的平均序.可见,CCE 方法取得了不错的性能,在 F_{macro}^1 与 F_{micro}^1 两个指标上都取得了最小的平均序.CCE 方法在 3 个数据集上取得了最好的 F_{macro}^1 性能,在 4 个数据集上取得了最好的 F_{micro}^1 性

能.然而,ECC方法只分别在1个和2个数据集上取得相应的性能.这一结果验证了本文方法的有效性.

Table 2 Average performance for CCE method and compared methods (using 80% instances for training)
表2 本文CCE方法与比较方法的平均性能(采用80%训练样本)

指标	数据集	CCE	BR	ML-kNN	CLR	CC	ECC
F_{macro}^1	Emotions	.6491±.0310 ₍₁₎	.6039±.0264 ₍₄₎ *	.3863±.0226 ₍₆₎ *	.6115±.0267 ₍₃₎ *	.6001±.0358 ₍₅₎ *	.6374±.0300 ₍₂₎ *
	Enron	.1620±.0072 ₍₁₎	.1454±.0061 ₍₄₎ *	.0876±.0079 ₍₆₎ *	.1371±.0059 ₍₅₎ *	.1456±.0074 ₍₃₎ *	.1594±.0077 ₍₂₎
	Genbase	.6503±.0349 ₍₂₎	.6504±.0350 ₍₁₎	.5441±.0256 ₍₆₎ *	.5702±.0308 ₍₅₎ *	.6446±.0393 ₍₃₎ *	.6195±.0362 ₍₄₎ *
	Medical	.3322±.0236 ₍₂₎	.3313±.0196 ₍₃₎	.2406±.0153 ₍₆₎ *	.2743±.0160 ₍₅₎ *	.3070±.0167 ₍₄₎ *	.3472±.0216 ₍₁₎ °
	Scene	.7639±.0185 ₍₁₎	.7402±.0177 ₍₅₎ *	.7603±.0160 ₍₃₎	.7457±.0196 ₍₄₎ *	.7322±.0187 ₍₆₎ *	.7604±.0180 ₍₂₎
	Yeast	.3634±.0088 ₍₂₎	.3350±.0082 ₍₅₎ *	.3718±.0133 ₍₁₎ °	.3326±.0075 ₍₆₎ *	.3420±.0088 ₍₄₎ *	.3465±.0072 ₍₃₎ *
	avg. rank	1.50±0.55	3.67±1.51	4.67±2.16	4.67±1.03	4.17±1.17	2.33±1.03
F_{micro}^1	Emotions	.6706±.0274 ₍₁₎	.6327±.0252 ₍₄₎ *	.4905±.0233 ₍₆₎ *	.6355±.0277 ₍₃₎ *	.6279±.0350 ₍₅₎ *	.6578±.0267 ₍₂₎ *
	Enron	.5772±.0138 ₍₂₎	.5750±.0124 ₍₃₎	.4808±.0188 ₍₆₎ *	.5736±.0129 ₍₄₎ *	.5708±.0145 ₍₅₎ *	.5836±.0139 ₍₁₎ °
	Genbase	.9820±.0083 ₍₁₎	.9819±.0080 ₍₂₎	.9455±.0156 ₍₆₎ *	.9624±.0154 ₍₅₎ *	.9739±.0083 ₍₄₎ *	.9790±.0074 ₍₃₎ *
	Medical	.8200±.0241 ₍₃₎	.8235±.0187 ₍₂₎	.7010±.0245 ₍₆₎ *	.7955±.0228 ₍₄₎ *	.7949±.0217 ₍₅₎ *	.8270±.0208 ₍₁₎ °
	Scene	.7553±.0192 ₍₁₎	.7318±.0174 ₍₅₎ *	.7520±.0165 ₍₂₎	.7368±.0201 ₍₄₎ *	.7232±.0203 ₍₆₎ *	.7517±.0186 ₍₃₎ *
	Yeast	.6424±.0119 ₍₁₎	.6302±.0105 ₍₅₎ *	.6422±.0126 ₍₂₎	.6293±.0104 ₍₆₎ *	.6337±.0113 ₍₄₎ *	.6375±.0106 ₍₃₎ *
	avg. rank	1.50±0.84	3.50±1.38	4.67±2.07	4.33±1.03	4.83±0.75	2.17±0.98

表3、表4进一步给出了60%和40%训练样本下的实验结果.可以看出,CCE方法同样在两个指标上都取得了比ECC及其他多标记方法更好的性能.总体上讲,在一共18种情况中,CCE方法均在11种情况下取得了最好的 F_{macro}^1 和 F_{micro}^1 性能,而ECC方法只分别在3种和5种情况下取得相应的性能.这一结果进一步验证了本文方法的有效性适用于不同数量的训练样本.

图2给出了模型参数对CCE方法和ECC方法的预测性能的影响.CCE方法的参数为迭代轮数,ECC方法的参数为集成分类器个数.实验采用80%训练样本,重复20轮,汇报平均 F_{macro}^1 和 F_{micro}^1 结果.可见,CCE方法对模型参数不敏感.当迭代轮数大于3时,CCE方法通常取得比较稳定的结果.ECC方法在有些情况下对模型参数敏感(如Genbase数据集).在相同参数情况下,CCE方法的性能通常优于ECC方法.

Table 3 Average performance for CCE method and compared methods (using 60% instances for training)
表3 本文CCE方法与比较方法的平均性能(采用60%训练样本)

指标	数据集	CCE	BR	ML-kNN	CLR	CC	ECC
F_{macro}^1	Emotions	.6314±.0236 ₍₁₎	.5814±.0290 ₍₄₎ *	.3857±.0236 ₍₆₎ *	.5830±.0224 ₍₃₎ *	.5726±.0277 ₍₅₎ *	.6170±.0202 ₍₂₎ *
	Enron	.1561±.0054 ₍₁₎	.1394±.0054 ₍₄₎ *	.0776±.0064 ₍₆₎ *	.1313±.0056 ₍₅₎ *	.1396±.0058 ₍₃₎ *	.1517±.0058 ₍₂₎ *
	Genbase	.6416±.0249 ₍₂₎	.6422±.0242 ₍₁₎	.5325±.0237 ₍₅₎ *	.5198±.0301 ₍₆₎ *	.6369±.0233 ₍₃₎ *	.6134±.0257 ₍₄₎ *
	Medical	.3246±.0136 ₍₂₎	.3245±.0143 ₍₃₎	.2330±.0139 ₍₆₎ *	.2491±.0168 ₍₅₎ *	.2994±.0152 ₍₄₎ *	.3382±.0154 ₍₁₎ °
	Scene	.7461±.0129 ₍₂₎	.7277±.0113 ₍₅₎ *	.7470±.0115 ₍₁₎	.7315±.0113 ₍₄₎ *	.7194±.0138 ₍₆₎ *	.7425±.0124 ₍₃₎
	Yeast	.3620±.0077 ₍₁₎	.3308±.0043 ₍₅₎ *	.3602±.0085 ₍₂₎	.3272±.0046 ₍₆₎ *	.3362±.0062 ₍₄₎ *	.3442±.0046 ₍₃₎ *
	avg. rank	1.50±0.55	3.67±1.51	4.33±2.25	4.83±1.17	4.17±1.17	2.50±1.05
F_{micro}^1	Emotions	.6523±.0228 ₍₁₎	.6160±.0237 ₍₃₎ *	.4766±.0217 ₍₆₎ *	.6117±.0209 ₍₄₎ *	.6035±.0244 ₍₅₎ *	.6387±.0199 ₍₂₎ *
	Enron	.5729±.0067 ₍₂₎	.5700±.0066 ₍₃₎ *	.4692±.0136 ₍₆₎ *	.5691±.0078 ₍₄₎ *	.5665±.0091 ₍₅₎ *	.5804±.0086 ₍₁₎ °
	Genbase	.9769±.0083 ₍₁₎	.9769±.0087 ₍₂₎	.9402±.0134 ₍₆₎ *	.9468±.0155 ₍₅₎ *	.9751±.0089 ₍₃₎ *	.9724±.0106 ₍₄₎ *
	Medical	.8135±.0166 ₍₃₎	.8169±.0150 ₍₁₎ °	.6806±.0225 ₍₆₎ *	.7783±.0133 ₍₅₎ *	.7880±.0170 ₍₄₎ *	.8168±.0134 ₍₂₎
	Scene	.7371±.0131 ₍₂₎	.7193±.0116 ₍₅₎ *	.7381±.0116 ₍₁₎	.7224±.0117 ₍₄₎ *	.7111±.0144 ₍₆₎ *	.7340±.0130 ₍₃₎
	Yeast	.6412±.0048 ₍₁₎	.6275±.0054 ₍₅₎ *	.6357±.0077 ₍₃₎ *	.6267±.0053 ₍₆₎ *	.6309±.0061 ₍₄₎ *	.6367±.0051 ₍₂₎ *
	avg. rank	1.67±0.82	3.17±1.60	4.67±2.16	4.67±0.82	4.50±1.05	2.33±1.03

Table 4 Average performance for CCE method and compared methods (using 40% instances for training)

表 4 本文 CCE 方法与比较方法的平均性能(采用 40%训练样本)

指标	数据集	CCE	BR	ML-kNN	CLR	CC	ECC
F^1_{macro}	Emotions	.5991±.0238 ₍₁₎	.5530±.0228 ₍₄₎ •	.3647±.0305 ₍₆₎ •	.5541±.0236 ₍₃₎ •	.5416±.0223 ₍₅₎ •	.5863±.0208 ₍₂₎ •
	Enron	.1438±.0059 ₍₁₎	.1291±.0056 ₍₄₎ •	.0690±.0038 ₍₆₎ •	.1182±.0046 ₍₅₎ •	.1298±.0057 ₍₃₎ •	.1384±.0058 ₍₂₎ •
	Genbase	.6017±.0458 ₍₁₎	.6015±.0460 ₍₂₎	.4741±.0267 ₍₅₎ •	.4629±.0381 ₍₆₎ •	.5868±.0467 ₍₃₎ •	.5499±.0467 ₍₄₎ •
	Medical	.2900±.0199 ₍₂₎	.2873±.0206 ₍₃₎ •	.2049±.0134 ₍₆₎ •	.2169±.0140 ₍₅₎ •	.2722±.0150 ₍₄₎ •	.3162±.0172 ₍₁₎ °
	Scene	.7379±.0087 ₍₁₎	.7213±.0066 ₍₄₎ •	.7372±.0086 ₍₂₎	.7182±.0085 ₍₅₎ •	.7082±.0126 ₍₆₎ •	.7363±.0104 ₍₃₎
	Yeast	.3537±.0080 ₍₁₎	.3211±.0065 ₍₅₎ •	.3447±.0081 ₍₂₎ •	.3169±.0055 ₍₆₎ •	.3268±.0066 ₍₄₎ •	.3369±.0048 ₍₃₎ •
	avg. rank	1.17±0.41	3.67±1.03	4.50±1.97	5.00±1.10	4.17±1.17	2.50±1.05
F^1_{micro}	Emotions	.6273±.0177 ₍₁₎	.5907±.0150 ₍₃₎ •	.4566±.0225 ₍₆₎ •	.5852±.0200 ₍₄₎ •	.5750±.0157 ₍₅₎ •	.6129±.0161 ₍₂₎ •
	Enron	.5600±.0077 ₍₂₎	.5576±.0073 ₍₃₎ •	.4469±.0158 ₍₆₎ •	.5545±.0059 ₍₄₎ •	.5528±.0076 ₍₅₎ •	.5682±.0056 ₍₁₎ °
	Genbase	.9677±.0122 ₍₁₎	.9676±.0128 ₍₂₎	.9174±.0148 ₍₆₎ •	.9240±.0228 ₍₅₎ •	.9591±.0142 ₍₃₎ •	.9560±.0129 ₍₄₎ •
	Medical	.7918±.0167 ₍₃₎	.7918±.0153 ₍₂₎	.6407±.0180 ₍₆₎ •	.7471±.0159 ₍₅₎ •	.7577±.0160 ₍₄₎ •	.7949±.0119 ₍₁₎
	Scene	.7273±.0087 ₍₁₎	.7123±.0067 ₍₄₎ •	.7272±.0085 ₍₂₎	.7080±.0085 ₍₅₎ •	.6981±.0137 ₍₆₎ •	.7270±.0109 ₍₃₎
	Yeast	.6366±.0058 ₍₁₎	.6216±.0056 ₍₅₎ •	.6298±.0052 ₍₃₎ •	.6202±.0052 ₍₆₎ •	.6247±.0060 ₍₄₎ •	.6328±.0046 ₍₂₎ •
	avg. rank	1.50±0.84	3.17±1.17	4.83±1.83	4.83±0.75	4.50±1.05	2.17±1.17

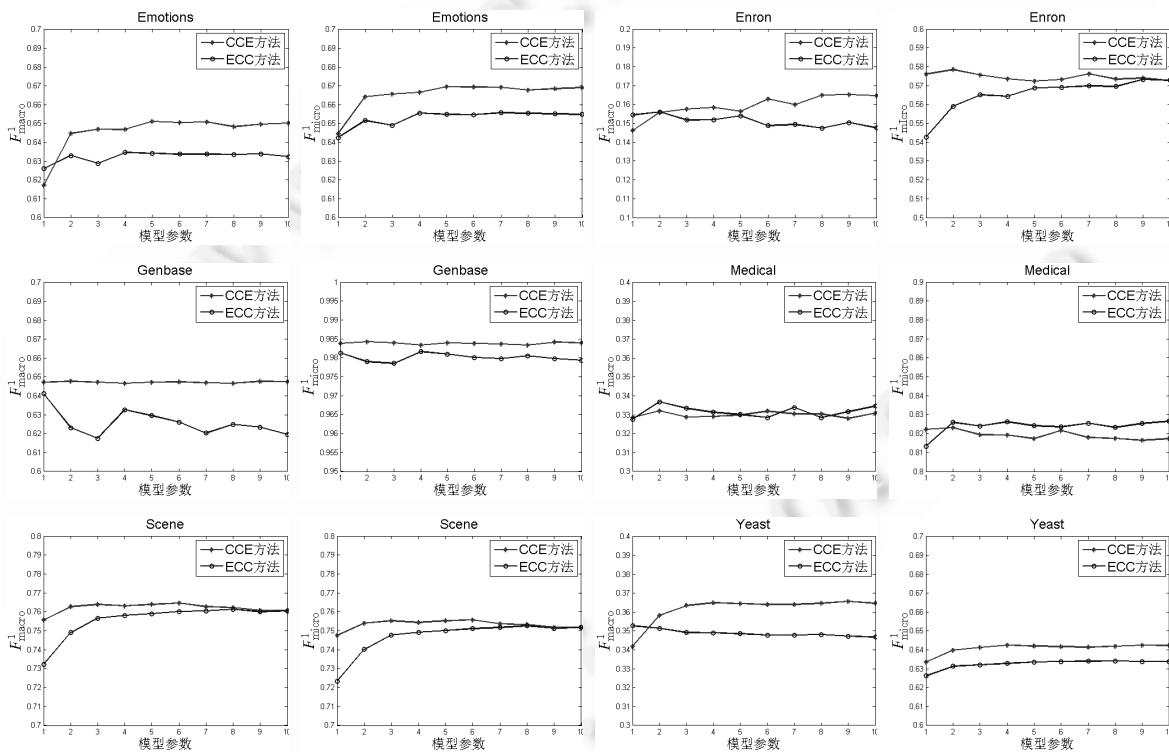


Fig.2 Parameter influence on the predictive performance of CCE and ECC method

图 2 CCE 方法和 ECC 方法的参数对预测性能的影响

图 3、图 4 进一步给出了 CCE 方法在随机标记序和固定标记序的实验结果(采用 80%样本做训练,重复 30 轮).成对 t 检验(95%置信度)的结果表明,随机序与固定序的性能没有显著差异.即,CCE 方法的性能对学习次序不敏感.

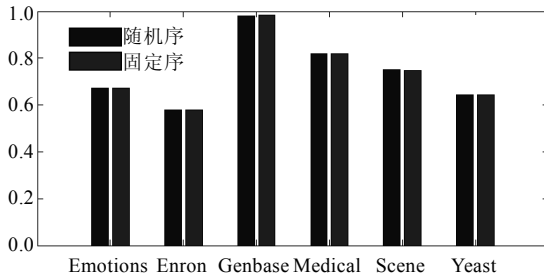


Fig.3 F_{macro}^1 performance result for CCE with random order and fixed order

图3 CCE方法分别采用随机序和固定序的 F_{macro}^1 性能结果

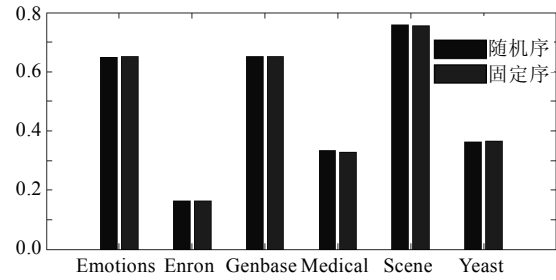


Fig.4 F_{micro}^1 performance result for CCE with random order and fixed order

图4 CCE方法分别采用随机序和固定序的 F_{micro}^1 性能结果

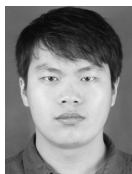
4 结束语

利用标记关系来提高学习性能,是多标记学习的关键.分类器链方法及其变型是这方面一类经典做法,但它依赖于标记学习次序的选择.这个信息现实情况难以可靠得到,使用不当会使性能提高有限.针对这个问题,本文提出了分类器圈方法 CCE.CCE 通过圈结构迭代训练每个标记的分类器,避免了分类器链方法学习次序选择的问题.实验结果表明,CCE 方法取得了比分类器链方法及一系列经典多标记学习方法更好的性能.未来将研究更精确设置 CCE 的迭代轮数(比如经验风险不再变化)、将 CCE 用于大规模数据以及 CCE 工作的理论基础.

References:

- [1] Mitchell TM. Machine Learning. New York: McGraw-Hill, 1997.
- [2] Tsoumakas G, Katakis I. Multi-Label classification: An overview. Int'l Journal of Data Warehousing and Mining, 2007,3(3):1-13. [doi: 10.4018/jdwm.2007070101]
- [3] Zhang ML, Zhou ZH. A review on multi-label learning algorithms. IEEE Trans. on Knowledge and Data Engineering, 2014,26(8): 1819-1837. [doi: 10.1109/TKDE.2013.39]
- [4] Schapire RE, Singer Y. BoosTexter: A boosting-based system for text categorization. Machine Learning, 2000,39(2/3):135-168. [doi: 10.1023/A:1007649029923]
- [5] McCallum A. Multi-Label text classification with a mixture model trained by EM. In: Proc. of the Working Notes of the AAAI'99 Workshop on Text Learning. 1999.
- [6] Boutell MR, Luo J, Shen X, Brown CM. Learning multi-label scene classification. Patter Recognition, 2004,37(9):1757-1771. [doi: 10.1016/j.patcog.2004.03.009]
- [7] Elisseeff A, Weston J. A kernel method for multi-labelled classification. In: Proc. of the Advances in Neural Information Processing Systems 14. Cambridge: MIT Press, 2002. 681-687.
- [8] Clare A, King RD. Knowledge discovery in multi-label phenotype data. Lecture Notes in Computer Science, 2001,2168:42-53. [doi: 10.1007/3-540-44794-6_4]
- [9] Barutcuoglu Z, Schapire RE, Troyanskaya OG. Hierarchical multi-label prediction of gene function. Bioinformatics, 2006,22(7): 830-836. [doi: 10.1093/bioinformatics/btk048]
- [10] Song Y, Zhang L, Giles LC. A sparse Gaussian processes classification framework for fast tag suggestions. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management. 2008. 93-102. [doi: 10.1145/1458082.1458098]
- [11] Ozonat K, Young D. Towards a universal marketplace over the web: Statistical multi-label classification of service provider forms with simulated annealing. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2009. 1295-1303. [doi: 10.1145/1557019.1557158]
- [12] Zhang ML, Zhou ZH. Multi-Label learning by exploiting label dependency. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2010. 999-1007. [doi: 10.1145/1835804.1835930]

- [13] Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. *Lecture Notes in Artificial Intelligence*, 2009,5782:254–269. [doi: 10.1007/978-3-642-04174-7_17]
- [14] Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. *Machine Learning*, 2011,85(3):333–359. [doi: 10.1007/s10994-011-5256-5]
- [15] Madjarov G, Dragi K, Gjorgjevikj D, Džeroski S. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 2012,45:3084–3104. [doi: 10.1016/j.patcog.2012.03.004]
- [16] Cheng W, Dembczyński K, Hüllermeier E. Bayes optimal multilabel classification via probabilistic classifier chains. In: *Proc. of the 27th Int'l Conf. on Machine Learning*. 2010. 279–286.
- [17] Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 2007,40(7):2038–2048. [doi: 10.1016/j.patcog.2006.12.019]
- [18] Fürnkranz J, Hüllermeier E, Mencia EL, Brinker K. Multilabel classification via calibrated label ranking. *Machine Learning*, 2008, 73(2):133–153. [doi: 10.1007/s10994-008-5064-8]
- [19] Zhang ML, Zhou ZH. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans. on Knowledge and Data Engineering*, 2006,18(10):1338–1351. [doi: 10.1109/TKDE.2006.162]
- [20] Tsoumakas G, Vlahavas I. Random k -labelsets: An ensemble method for multilabel classification. *Lecture Notes in Artificial Intelligence*, 2007,4701:406–417. [doi: 10.1007/978-3-540-74958-5_38]
- [21] Godbole S, Sarawagi S. Discriminative methods for multi-labeled classification. In: *Proc. of the 18th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*. Berlin: Springer-Verlag, 2004. 22–30. [doi: 10.1007/978-3-540-24775-3_5]
- [22] Cheng WW, Hüllermeier E. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 2009,76(2/3):211–225. [doi: 10.1007/s10994-009-5127-5]
- [23] Cai L, Hofmann T. Hierarchical document categorization with support vector machines. In: *Proc. of the 13th ACM Int'l Conf. on Information and Knowledge Management*. 2004. 78–87. [doi: 10.1145/1031171.1031186]
- [24] Hariharan B, Zelnik-Manor L, Vishwanathan SVN, Varma M. Large scale max-margin multi-label classification with priors. In: *Proc. of the 27th Int'l Conf. on Machine Learning*. 2010. 423–430.
- [25] Rousu J, Saunders C, Szedmak S, Shawe-Taylor J. Learning hierarchical multi-category text classification models. In: *Proc. of the 22nd Int'l Conf. on Machine Learning*. 2005. 774–751. [doi: 10.1145/1102351.1102445]
- [26] Cesa-Bianchi N, Gentile C, Zaniboni L. Hierarchical classification: Combining Bayes with SVM. In: *Proc. of the 23rd Int'l Conf. on Machine Learning*. 2006. 177–184. [doi: 10.1145/1143844.1143867]
- [27] Ghamrawi N, Mccallum A. Collective multilabel classification. In: *Proc. of the 14th ACM Int'l Conf. on Information and Knowledge Management*. 2005. 195–200. [doi: 10.1145/1099554.1099591]
- [28] Tsoumakas G, Dimou A, Spyromitros E, Mezaris V. Correlation-Based pruning of stacked binary relevance models for multi-label learning. In: *Proc. of the 1st Int'l Workshop on Learning from Multi-Label Data*. 2009. 101–116.
- [29] Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, Duch W. A shared task involving multi-label classification of clinical free text. In: *Proc. of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Stroudsburg: Association for Computational Linguistics, 2007. 97–104.
- [30] Klimt B, Yang Y. Introducing the enron corpus. In: *Proc. of the 1st Conf. on Email and Anti-Spam*. 2004.
- [31] Trohidis K, Tsoumakas G, Kalliris G, Vlahavas I. Multilabel classification of music into emotions. In: *Proc. of the 9th Int'l Conf. on Music Information Retrieval*. 2008. 325–330.
- [32] Diplaris S, Tsoumakas G, Mitkas P, Vlahavas I. Protein classification with multiple algorithms. In: *Proc. of the 10th Panhellenic Conf. on Informatics*. 2005. 448–456. [doi: 10.1007/11573036_42]
- [33] Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 2008,9:1871–1874.



王少博(1990—),男,河南郑州人,硕士生,主要研究领域为机器学习。



李宇峰(1983—),男,博士,助理研究员,CCF会员,主要研究领域为机器学习,数据挖掘。