DOI:10.13718/j.cnki.xsxb.2019.03.021

基于转换策略的多标记学习改进算法①

任 翀

广西广播电视大学 文理学院,南宁 530001

摘要:针对传统转换多标记学习算法较难确定最低阈值的问题,该文对传统转换策略的多标记学习算法进行改进,提出了一种基于最低阈值的学习算法(TFEL),该方法根据类别标记学习为每个类别得到一个最低阈值.当分类器将一个测试示例预测为某个类别标记的分值大于为该类别标记学得的最低阈值时,则将该类别标记添加到该测试示例的最终分类结果中.实验结果表明,TFEL 方法能够得到较好的分类效果,证明了该方法的可行性和有效性.

关 键 词:多标记学习;最低阈值;转换策略;分类

中图分类号:TP391 **文献标志码**:A **文章编号**:1000-5471(2019)03-0124-06

传统监督学习框架下多种算法对单一语义具有较好性能[1],但是现实中对象常表现出多义性,因此需要建立多个标签子集,多目标学习也变得越来越受关注.多目标学习在数据挖掘中的应用日益成为广大学者研究的焦点[2].多目标学习指其中的一个示例有多个类别标记与之相对应,其最终实现示例预测多个类别的标记[3].这样的例子有:某个基因序列[4],其可能具有多个功能,如"新陈代谢"和"合成白细胞"等.

有关多标记学习的研究已有很多,目前提出的学习方法和策略主要有两类:①提出新的算法或改进的算法,文献[5]中给出一种 k 近邻方法的多标记分类方法,并进行改进使得其性能更优.文献[6]提出了基于概率隐语义分析(Probabilistic latent semantic analysis,PLSA)模型的多标记假设重用文本分类算法,解决了多标记文本分类时文本标记关系不明确以及特征维数过大的问题.文献[7]中提出了一种用于多标签学习的多层 ELM-RBF 神经网络方法,在单标签和多标签数据集上都有较好性能.该类方法在多标记分类上局限性比较大.②基于转换的 PT(Problem Transformation)策略[8-9],总共包含有PT1~PT5 等 5 种方法.其中,PT5 的实现思路是首先把(x_i, Y_i)的多标记示例经过一定的处理后将其分成$|Y_i|$个单标记数据集,接着再将得到的$|Y_i|$个单标记数据集经过一定的处理后形成与之对应的一个单标记分类模型.确定一个合适的阈值是该方法实现的关键,阈值的取值大小直接影响分类器的性能.文献[10]中提出了一种标签优先标记集合(LPP)转换方法,根据标签的重要性排序来解决标签依赖性问题.文献[11]提出了一种基于标签间相关性的多标签分类方法,它使用了问题变换方法和算法适应方法,该方法分类准确性更高.

上述基于问题转换方法的关键是确定最低阈值,然而阈值的设定还没有一个准确的原则,如设置过高,类别标记会被漏判,如设置过低,则会出现多判.如何确定最低阈值还是目前的一个难题,针对这个问题,本文提出了一种基于最低阈值的多标记学习算法(TFEL,Threshold For Each Label),根据类别标记学习为每个类别得到一个最低阈值.当分类器将一个测试示例预测为某个类别标记的分值大于为该类别标记学得的最低阈值时,则将该类别标记添加到该测试示例的最终分类结果中.实验结果表明,本文提出的 TFEL 方法具有较好的分类效果.

1 PT5 算法的缺陷分析

PT5 是将(x_i, Y_i)的多标记示例经过一定的处理后形成$|Y_i|$个单标记数据集,比如,$(x_1, \{y_1, y_2\})$

① 收稿日期:2018-02-01
作者简介:任 翀(1981-),女,硕士,讲师,主要从事算法分析研究.

可以转换成(x_1, y_1)和(x_1, y_2)2 个单标记示例，接着分配一个单标记分类器给上述各个数据集. 同时，所有的单标记分类器都会有一个对应的分布，表示每个对象属于相应类别的概率，并根据分布为每个对象输出一组类标记集合. 通常情况下取 $threshold = 0.5$ 为最低阈值. 进行类别 y_l 包含示例 x_i 的概率预测时，如果该数值大于最低阈值，则将类别 y_l 合并到示例 x_i 预测类别集合. 公式表示形式为

$$f_{PT5}(x) = \bigcup_{y_l \in Y} \{y_l\}: f(x, y_l) > threshold \tag{1}$$

在 PT5 方法中，为其设置一个适当的最低阈值对该方法非常重要，在现有的多标记学习算法中设置一个适合的最低阈值也十分必要. 阈值设置得过高或过低都会影响到预测结果. 当我们设置的阈值过高时，得到的预测结果就可能不全；当我们设置的阈值过低时，就会得到大量的无用类别. 因此，在所有的类别中设置同一个阈值是不恰当的.

为解决设置最低阈值的难题，基于最低阈值多目标学习，本文提出了 TFEL 学习算法. 在使用该算法时，对于类别 $y_l(1 \leqslant l \leqslant |Y|)$，都将会有一个阈值与之对应，在本文中将其记为 $threshold_l$，同时所有类别的标记集合记为 Y. 如果一个示例 x_i 预测为类别 y_l，可能性 $f(x_i, y_l) > threshold_l$，将其代入式(2)，并将 y_l 的类别标记添加到对 x_i 预测的类别标记集合中，即

$$f_{TFEL}(x) = \bigcup_{y_l \in Y} \{y_l\}: f(x, y_l) > threshold_l \tag{2}$$

TFEL 方法对传统单标记学习算法和现有多标记算法中的阈值确定具有通用性.

2　基于最低阈值的多标记学习算法 TFEL(Threshold For Each Label)

对于多标记学习中如何设置适合的最低阈值，本文提出了一种新的学习算法 —— 基于最低阈值的多标记学习算法(TFEL, Threshold For Each Label).

2.1　TFEL 方法中的阈值确定

TFEL 方法通过对训练数据集学习，根据标记学习为每个类别得到一个阈值.

2.1.1　TFEL 方法阈值分析

在 TFEL 算法中，首先对训练数据集进行训练得到每个类别 y_l 的概率 $f(x_i, y_l)$. 然后将 $f(x_i, y_l)$ 存储在相应集合中，如式(3)、式(4) 所示. 再对概率 $f(x_i, y_l)$ 进行学习，为每个标记 y_l 得一个最低阈值.

$$\Lambda_l^+ = \{f(x_i, y_l) \mid x_i \in D_l^+\}, 1 \leqslant l \leqslant |Y| \tag{3}$$

$$\Lambda_l^- = \{f(x_i, y_l) \mid x_i \in D_l^-\}, 1 \leqslant l \leqslant |Y| \tag{4}$$

在上式中，对于类别 y_l 中所有示例的数据集记为 D_l^+；对于 D_l^+ 示例中属于 y_l 的概率 $f(x_i, y_l)$ 的集合记为 Λ_l^+. 同理，对于类别 y_l 中所有示例的数据集记为 D_l^-；对于 D_l^- 示例中属于 y_l 的概率 $f(x_i, y_l)$ 的集合记为 Λ_l^-. 如式(5) 和式(6) 所示，其分别属于 Λ_l^+ 中的最小值与最大值区间和 Λ_l^- 中的最小值与最大值区间.

$$\Xi_l^- = [\Lambda_{Min_l}^+, \Lambda_{Max_l}^+], \Lambda_{Min_l}^+, \Lambda_{Max_l}^+ \in \Lambda_l^+ \tag{5}$$

$$\Xi_l^- = [\Lambda_{Min_l}^-, \Lambda_{Max_l}^-], \Lambda_{Min_l}^-, \Lambda_{Max_l}^- \in \Lambda_l^- \tag{6}$$

在上式中，由集合 Λ_l^+ 中的最小概率 $\Lambda_{\mathrm{Min}_l}^+$ 和最大概率 $\Lambda_{\mathrm{Max}_l}^+$ 构成的区间记为 Ξ_l^+. 同理，由集合 Λ_l^- 中的最小概率 $\Lambda_{\mathrm{Min}_l}^-$ 和最大概率 $\Lambda_{\mathrm{Max}_l}^-$ 构成的区间记为 Ξ_l^-.

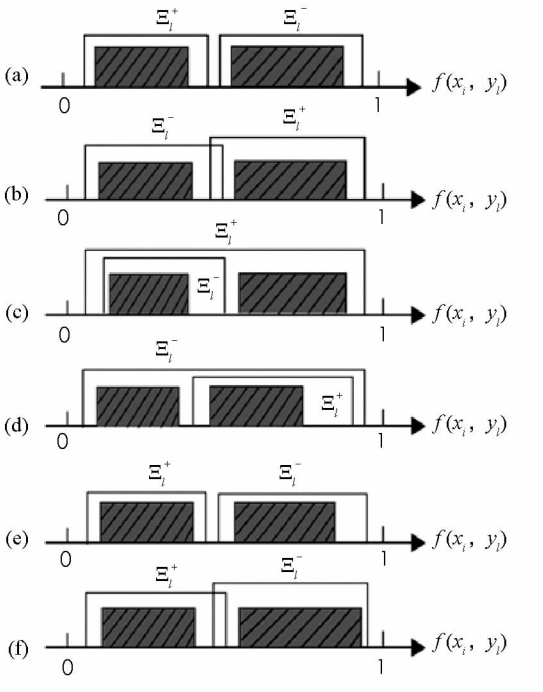

图 1　区间 Ξ_l^+ 和 Ξ_l^- 的位置关系

如图 1 所示，区间 Ξ_l^+ 和 Ξ_l^- 存在 6 种位置关系. 图 1(a) 表示所有正例比负例被预测为 y_l 的概率 $f(x_i, y_l)$ 都高，图 1 表示的是最理想状态. 2 个区间有交集的情况如图 1(b) — 图 1(d) 所示，最常见的情况如图 1(b) 所示. 正确的可能性大小和 $f(x_i, y_l)$ 存在的位置有关系，对于一个测试示例，当其被预测为类别 y_l 的 $f(x_i, y_l)$ 值位于交集中时，那么预测结果很有可能是错误的. 当其被预测为类别 y_l 的 $f(x_i, y_l)$ 值位于交

集两侧时，那么预测结果很有可能是正确的. 其他情况如图1(e)－图1(f)所示，在训练集中对于本应该属于类别 y_l 的示例，其被预测为不属于类别 y_l 的概率要比预测为属于类别 y_l 的概率要大，可以看出该分类器性能较差.

由最大值 $\Lambda^{\pm}_{\mathrm{Max}_l}$ 和最小值 $\Lambda^{\pm}_{\mathrm{Min}_l}$ 可以确定区间，从图1(a)－图1(f)中的阴影部分可以看出，Λ_l^+ 和 Λ_l^- 中的分值集中分布在某一段区间上. 通过 χ^2 拟合度检验得知，Λ_l^+ 和 Λ_l^- 集合中的值 $f(x_i, y_l)$ 均近似服从正态分布，如图2所示.

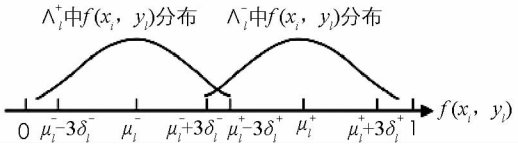

图2 Λ_l^+ 和 Λ_l^- 分值 $f(x_i, y_l)$ 的分布情况

2.1.2 TFEL方法阈值计算

确定正态分布中的参数 μ 和参数 δ 是求解TFEL方法中最低阈值的前提. 因此，求得参数 μ 和 δ 的准确值对阈值确定很关键. 本文根据无偏估计法求出其近似值，求解方法如式(7)和式(8)所示.

$$\hat{u}_l^{\pm} = \frac{1}{|D_l^{\pm}|}\sum_{i=1}^{|D_l^{\pm}|} f(x_i, y_l),\ x_i \in D_l^{\pm} \tag{7}$$

$$\hat{\delta}_l^{\pm} = \sqrt{\frac{1}{|D_l^{\pm}|-1}\sum_{i=1}^{|D_l^{\pm}|}(f(x_i, y_l) - \hat{u}_l^{\pm})^2},\ x_i \in D_l^{\pm} \tag{8}$$

对 Λ_l^+ 中所有概率值求标准差和均值，分别记为 $\hat{\delta}_l^+$ 和 $\hat{u}_l^+$. 同理，对 Λ_l^- 中所有概率值求标准差和均值，分别记为 $\hat{\delta}_l^-$ 和 $\hat{u}_l^-$. 由正态分布图中的 3δ 标准可知，阈值 $\mathrm{threshold}_l$ 可能会有以下3种：

$$\mathrm{Min}_l = u_l^+ - i * \delta_l^+ \qquad 1 \leqslant i \leqslant 3 \tag{9}$$

$$\mathrm{Max}_l = u_l^- + i * \delta_l^- \qquad 1 \leqslant i \leqslant 3 \tag{10}$$

$$\mathrm{Mid}_l = (\mathrm{Min}_l + \mathrm{Max}_l)/2 \tag{11}$$

式9－式10中参数均利用式7－式8中计算得到的估计值.

2.2 TFEL算法实现

TFEL算法伪码的实现过程如下：

In：

集合 D 用于表示多标记数据集的集合，则有 $D = \{(x_1, Y_1), (x_2, Y_2), \cdots, (x_{|D|}, Y_{|D|})\}$；

集合 Y 表示所有类别标记的集合，则有 $Y = \{y_1, y_2, \cdots, y_{|Y|}\}$；

$f(\ ,\ )$：计算概率的函数；

thrType：阈值类型(Min_l, Max_l, Mid_l)以及 i 的值；

t：测试示例；

Out：$f_{TFEL}(t)$ 对测试示例 t 预测的类别标记集合.

Process：

1) 计算 Λ_l^+ 和 Λ_l^- －，$1 \leqslant l \leqslant |Y|$；

2) 根据式7－式8计算 $u_i, \delta_i, u_l^{\pm}$ 和 $\delta_l^{\pm}$，$1 \leqslant l \leqslant |Y|$；

3) 根据上一步计算得到的结果和给定的 thrType 计算 $\mathrm{threshold}_l$，$1 \leqslant l \leqslant |Y|$；

4) $f_{TFEL}(t) = \bigcup_{y_l \in Y}\{y_l\}$：$f(t, y_l) > \mathrm{threshold}_l$

3 实验与分析

3.1 算法评估标准

多标记学习算法常用的评估标准有 Hamming Loss, Average Precision, Average Precision, Coverage 和 α-Evaluation 等. 本文选取2个最适合的评估标准 Hamming Loss 和 α-Evaluation.

3.1.1 Hamming Loss

根据 Hamming Loss 评估标准，用计算器分类器可以得出对象预测出的类别标记集合和实际对应的类

别标记集合差异个数. D 为多标记数据集，M 表示数据集中对象的总数，$f(x)$ 为多标记分类器，x_i 为测试数据对象，Y_i 为 x_i 对应的类标记集合，该评估标准可表示为式(12) 所示.

$$HLoss_D = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|Y|}|f(x_i)\Delta Y_i| \tag{12}$$

其中，$f(x_i)\Delta Y_i = (f(x_i)-Y_i)(Y_i-f(x_i))$. Hamming loss 的值越小说明分类器的性能越好，反之则差.

3.1.2　α-Evaluation

假定多标记数据集 D，类标记集合 $Y=\{y_1, y_2, \cdots, y_L\}$，示例 x_i 对应的类标记集合为 Y_i，分类器得到的类标记集合为 P_i. 没有预测到的类标记集合用 $M_i = Y_i - P_i$ 表示；预测的错误类标记集合用 $F_i = P_i - Y_i$ 表示.

$$score(P_i) = \left(1 - \frac{|\beta M_i + \gamma F_i|}{|Y_i \cup P_i|}\right)^{\alpha} \tag{13}$$

其中，参数 $\alpha \geqslant 0$，$\beta \geqslant 0$，$\gamma \leqslant 1$，$\beta = 1 \mid \gamma = 1$，该值越大，表明分类器的性能越好.

3.2　实验设计

为验证 TFEL 算法的有效性，从 UCI 下载数据集 iris 和 diabetes. 这 2 个数据集中均存在属性相同，而对应类别标记不相同的示例. 实验中将这些属性相同，对应类别标记不相同的示例类别标记预处理成他们所属多个类别标记的集合. 例如，存在示例 (x_1, y_1)，(x_1, y_2) 和 (x_1, y_3)，则将示例的类别标记预处理成 $\{y_1, y_2, y_3\}$，即 $(x_1, \{y_1, y_2, y_3\})$. 预处理后的数据集信息如表 1 所示.

表 1　数据集信息概要

类别标记	iris	diabetes
{1}	50	95
{2}	45	276
{3}	6	0
{1, 2}	0	394
{2, 3}	49	0
总数	150	765

为了方便获得一个示例属于每个类别标记的概率 $f(x_i, y_l)$，基于贝叶斯算法对分类模型进行建模. 为确定阈值大小对分类结果的影响，本文根据 TFEL 和朴素贝叶斯算法，为每个类别设置一个最低阈值，同时为每个类别标记选取 9 组值. 将 i 取 1,2 和 3 分别代入公式(9) 和公式(10) 中，threshold_l 产生 3 组不同的取值(Min_l,Max_l和 Mid_l)，然后通过 TFEL 方法对测试数据集进行分类. 设定 β 和 γ 参数的取值均为 1，通过 Hamming Loss 和 α-Evaluation 评估标准对分类结果进行评估.

3.3　TFEL 方法对 iris 数据集分类

利用 TFEL 方法对数据集 iris 分类后的评估结果如图 3 所示.

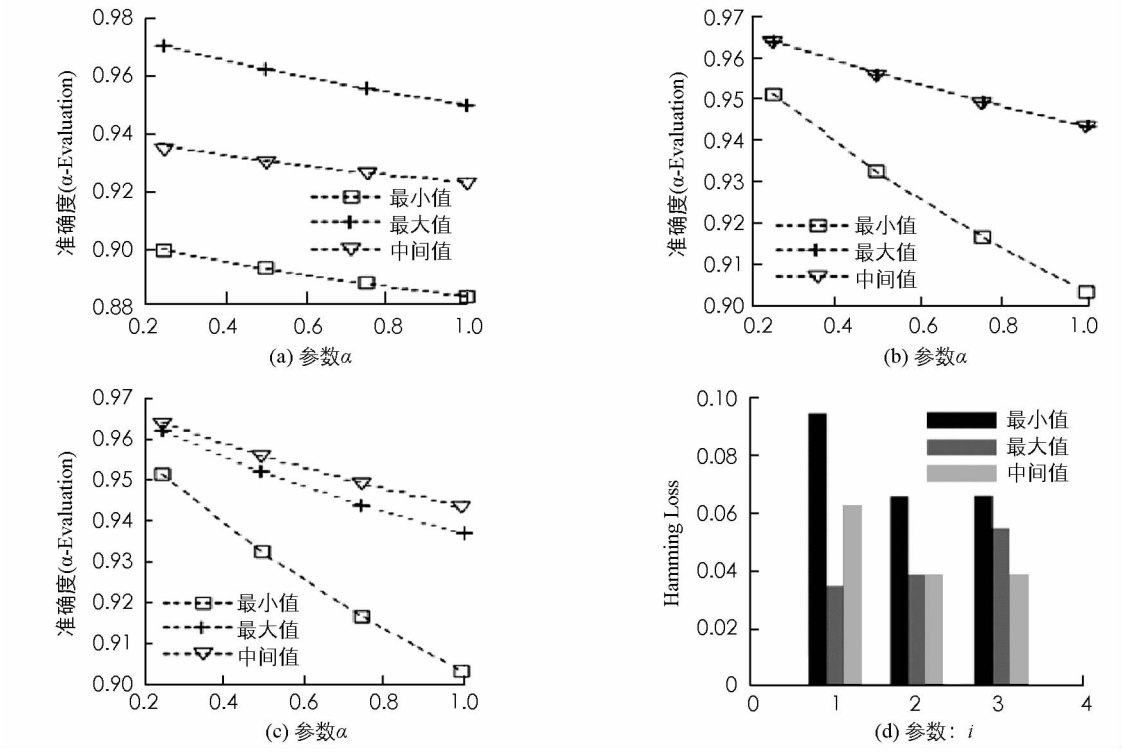

图 3　利用 TFEL 方法对数据集 iris 分类后的评估结果

其中，图3(a)、图3(b)和图3(c)分别是式9－式10中参数i取1,2和3时的分类结果，图3(d)是当i取不同值时，对Hamming loss评估值的比较. 通过对比图3(a)－图3(d)得出，当threshold_l和Max_l相等时，分类器性能达到最佳，当$i=1$时，准确率达到最高，Hamming loss评估值达到最低.

3.4 TFEL方法对diabetes数据集分类

为了确定当阈值$\text{threshold}_l = \text{Max}_l$时，是否对每个数据集分类器都能得到最佳性能，实验利用TFEL方法对diabetes数据集进行分类，分类后的评估结果如图4所示.

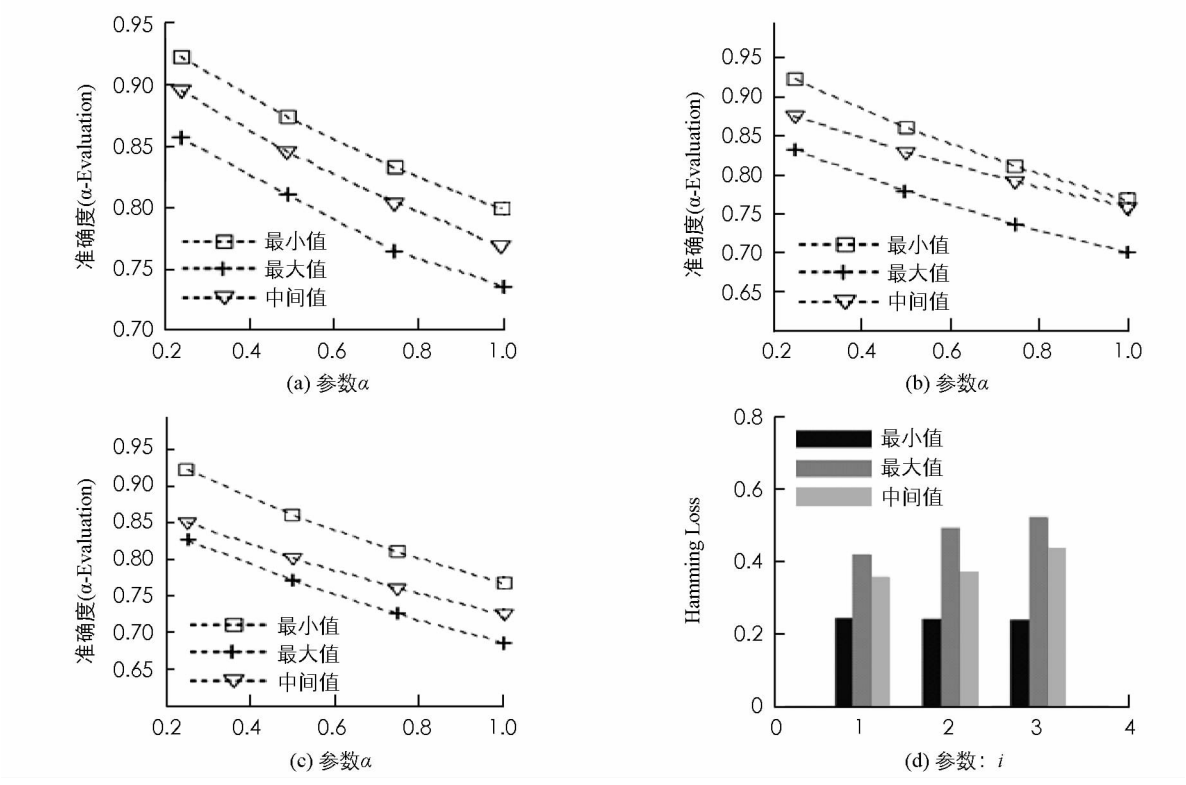

图4 利用TFEL方法对数据集diabetes分类后的评估结果

其中，图4(a)、图4(b)和图4(c)分别是式9－式10中的参数i取1,2和3时的分类结果，当$\text{threshold}_l = \text{Min}_l$时，分类器的准确率均高于其他2组取值. 图4(d)中，当$\text{threshold}_l = \text{Min}_l$时，Hamming loss评估值均小于其他2组评估值. 通过总结图4(a)－图4(d)得出，当分类器对diabetes数据集的整体分类性能达到最优时，$\text{threshold}_l = \text{Min}_l$.

实验结果表明，TFEL方法能有效的对多标记数据集进行分类，同时对于不同的数据集，在取不同的阈值时，分类器都可以表现出良好的性能.

4 结 论

通过对多标记学习详细的理论研究，本文结合最低阈值知识提出了多标记学习TFEL算法. 通过训练数据集可以得到每一个类别标记的最佳最低阈值，这样能够使分类器的分类性能达到最佳. 通过对该算法进行大量实验，结果表明TFEL算法具有较好的分类效果. 但本文算法也具有一定的不足，后续工作需要：①对单标记数据进行预处理得到数据集，将TFEL方法应用于现有的多标记数据集；②将TFEL方法中的阈值确定方法应用于现有的多标记学习算法，并进行分析比较；③进一步研究数据集中每一个类别正负例个数的分布对最低阈值的影响.

参考文献：

[1] 晏 勇. 基于SKLLE和SVM的人脸表情识别[J]. 西南师范大学学报(自然科学版)，2014，39(1)：55-60.

[2] 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.

[3] ZHANG M L, WU L. Lift: Multi-Label Learning with Label-Specific Features [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1): 107-120.
[4] 姜海燕，刘昊天，舒　欣，等. 基于最大均值差异的多标记迁移学习算法 [J]. 信息与控制，2016，45(4)：463-470，478.
[5] KANJ S, ABDALLAH F, DENCEUX T, et al. Editing Training Data for Multi-Label Classification with the K-nearest Neighbor Rule [J]. Pattern Analysis and Applications, 2016, 19(1): 145-161.
[6] 蒋铭初，潘志松，尤　峻. 基于 PLSA 主题模型的多标记文本分类 [J]. 数据采集与处理，2016，31(3)：541-547.
[7] ZHANG N, DING S, ZHANG J. Multi Layer ELM-RBF for Multi-Label Learning [J]. Applied Soft Computing, 2016, 43: 535-545.
[8] 余　鹰. 多标记学习研究综述 [J]. 计算机工程与应用，2015，51(17)：20-27.
[9] 梁新彦，钱宇华，郭　倩，等. 面向多标记学习的局部粗糙集 [J]. 南京大学学报(自然科学)，2016，52(2)：270-279.
[10] ABDALLAH Z, EL-ZAART A, OUEIDAT M. An Improvement of Label PowerSet Method Based on Priority Label Transformation [J]. International Journal of Applied Engineering Research, 2016, 11(16): 9079-9087.
[11] ALAZAIDAH R, THABTAH F, AL-RADAIDEH Q. A Multi-Label Classification Approach Based on Correlations Among Labels [J]. International Journal of Advanced Computer Science and Applications, 2015, 6(2): 52-59.

On Improved Multi-label Learning Algorithm Based on the Strategy of Transformation

REN　Chong

Department of Arts and Sciences, Guangxi Open University, Nanning 530001, China

Abstract: To solve the problem that the traditional problem transformation method, multi-label learning algorithm is difficult to determine the lowest threshold, the traditional multi-label learning algorithm has been improved in this paper and a learning algorithm based on the lowest threshold for each label (TFEL) been proposed. The method deals with learning a minimum threshold for each category label. When the score for a test instance to one label is bigger then threshold which is learned for the label, the label will be added to the last classifying result for the test instance. The method of programming, experimental results show that TFEL method can achieve better classification results, which proves the feasibility and effectiveness of this method.

Key words: multi-label learning; minimum threshold; strategy of transformation; classification

责任编辑　夏　娟