

基于双标签集的标签匹配集成学习算法

张丹普^{1,2*}, 王莉莉^{1,2}, 付忠良¹, 李 昕^{1,2}

(1. 中国科学院 成都计算机应用研究所, 成都 610041; 2. 中国科学院大学, 北京 100049)

(* 通信作者电子邮箱 Linda_zdp@126.com)

摘要:当标识示例的两个标签分别来源于两个标签集时,这种多标签分类问题称之为标签匹配问题,目前还没有针对标签匹配问题的学习算法。尽管可以用传统的多标签分类学习算法来解决标签匹配问题,但显然标签匹配问题有其自身特殊性。通过对标签匹配问题进行深入的研究,在连续 AdaBoost(real Adaptive Boosting)算法的基础上,基于整体优化的思想,采用算法适应的方法,提出了基于双标签集的标签匹配集成学习算法,该算法能够较好地学习到标签匹配规律从而完成标签匹配。实验结果表明,与传统的多标签学习算法用于解决标签匹配问题相比,提出的新算法不仅缩小了搜索的标签空间的范围,而且最小化学习误差可以随着分类器个数的增加而降低,进而使得标签匹配分类更加快速、准确。

关键词:连续 AdaBoost; 多标签学习; 多标签集; 标签匹配; 集成学习

中图分类号: TP391.4 **文献标志码:** A

Ensemble learning algorithm for labels matching based on pairwise labelsets

ZHANG Danpu^{1,2*}, WANG Lili^{1,2}, FU Zhongliang¹, LI Xin^{1,2}

(1. Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu Sichuan 610041, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: It is called labels matching problem when two labels of an instance come from two labelsets respectively in multi-label classification, however there is no any specific algorithm for solving such problem. Although the labels matching problem could be solved by traditional multi-label classification algorithms, but this problem has its own particularity. After analyzing the labels matching problem, a new labels matching algorithm based on pairwise labelsets was proposed using adaptive method, which considered the real Adaptive Boosting (real AdaBoost) and the global optimization idea. This algorithm could learn the rule of labels matching well and complete matching. The experimental results show that, compared with the traditional algorithms, the new algorithm can not only reduce searching scope of the labels space, but also decrease the minimum learning error as the number of weak classifiers increases, and make the classification more accurate and faster.

Key words: real AdaBoost (Adaptive Boosting); multi-label classification; multi-label dataset; label matching; ensemble learning

0 引言

随着信息技术的发展,互联网数据及资源在数据量增大的同时,数据标注结构复杂程度也在增加,数据对象往往并不只是具有唯一的语义,这就使得只考虑明确的、易分辨的、单一语义的传统的监督学习框架难以取得好的效果。为了直观地反映多义性对象所具有的多种语义信息,一种很自然的方式就是为该对象显式地赋予一组合适的类别标记,即标签子集。所以,传统的单标签数据研究成果已经不能满足技术发展的需要,多标签学习的研究对于解决多义对象的学习问题有十分重要的意义,现已逐渐成为研究的一个热点。

多标签分类算法的研究已经取得了许多成果,目前已经有了大量的多标签学习算法,大致可以分为问题转换和算法适应两种方法^[1];问题转换的方法是通过改造数据将多标签学习问题转化为其他已知的单标签学习问题进行求解,该方法不受特定算法的限制,比较典型的算法有二分法(Binary

Relevance, BR)^[2]、校准标签排序(Calibrated Label Ranking, CLR)^[3]和随机 k -标记集(Random k -labelsets)^[4]等;算法适应方法是通过直接改造现存的单标签学习算法,使之能够适应多标签数据的处理,该类方法代表性的学习算法有 ML- k NN^[5]、Rank-SVM^[6]、AdaBoost.MH^[7]、BoosTexter^[8]以及基于标签依赖的多标签学习(multi-label Learning by Exploiting Label Dependency, LEAD)^[9]等。

上述多标签学习只是针对一个多标签集问题,即每个对象具有多个类别标记,且这些类别标记属于一组标签集。然而,在实际生活中,一个对象具有多个类别标记,但是不同的类别标记可能来自不同的标签集,这些标签集之间通常又存在两种关系:第一种是并列关系,例如音乐可以按照流派分为流行、摇滚、电子、爵士等,也可以依据情感分为惊奇、愉悦、放松、寂寞、悲伤和生气等;第二种是层级关系,例如一条新闻既可以按照第一层来源分为腾讯、新浪、搜狐、环球和百度等门户网站,又可以依据第二层网站新闻版块分为体育、财经、科技、社会和教

收稿日期:2014-04-02;修回日期:2014-06-08。 基金项目:四川省科技支撑计划项目(2011GZ0171;2012GZ0106)。

作者简介:张丹普(1986-),女,河南平顶山人,博士研究生,主要研究方向:机器学习、模式识别; 王莉莉(1987-),女,河南周口人,博士研究生,主要研究方向:机器学习、模式识别; 付忠良(1967-),男,重庆合川人,研究员,博士生导师,主要研究方向:机器学习、模式识别; 李昕(1985-),男,陕西汉中,人,博士研究生,主要研究方向:图形图像处理、模式识别。

育等。这种一个对象具有两个或者多个标签集的多标签学习问题的研究尚且不多,该研究对解决大量数据快速准确且全面的分类问题具有十分重要的研究意义和应用价值。

本文针对一个对象具有两个标签集的多标签学习问题进行了深入的研究,基于算法适应的方法,在连续 AdaBoost(real Adaptive Boosting)算法^[7]的基础上提出了基于双标签集的标签匹配集成学习算法,即将一个对象用一个示例来描述,该示例具有分别来自于不同标签集的两个类别标记并形成标签对,学习的目的是同时准确地从两个标签集中学习到合适的类别标记赋予未见示例,本文算法未考虑样本标签间的相关性^[9-11]。该算法与现有多标签学习算法相比,缩小了标签空间的范围,并从概率论的角度分析,除了正确选择单个标签集的正确标记之外,又增加了两个标签集之间标记组合的可能性,因此,在一定程度上增加了学习的难度。对 Scene、Yeast^[12]和 Image^[13]三个多标签数据集的实验结果说明,本文算法取得了较好的分类效果。

1 基于双标签集的标签匹配算法的构造

设 X 为示例空间,训练样本集 $S = \{(x_1, (y_1, z_1)), (x_2, (y_2, z_2)), \dots, (x_m, (y_m, z_m))\}$, 其中 $y_i \in L_1 = \{a_1, a_2, \dots, a_K\}$, $z_i \in L_2 = \{b_1, b_2, \dots, b_W\}$, 即 $x_i \in X$ 的类别为来自两个不同的标签集的标签对 $\{y_i, z_i\}$ 。假设样本及样本标签都是相互独立的,把标签集转换为 $L_1 = \{1, 2, \dots, K\}$, $L_2 = \{K+1, K+2, \dots, N\}$, $N = K+W$, $x_i \in X$ 的标签集记为 $Y_i = \{y_i, z_i\}$, 标签空间为 $K \times W$ 。训练得到弱分类器 $h_i(x)$ 后得到组合分类器仍然为 $f(x)$, 其对应标签 l 的输出值 $f(x, l) = \sum_{i=1}^T h_i(x, l)$, 但是,其输出标签对为 $(\arg \max_{k \in \{1, 2, \dots, N\}} \{f(x, k)\}, \arg \min_{k \in \{1, 2, \dots, N\}} \{f(x, k)\})$, 注意,此处在整个标签集 $L = \{1, 2, \dots, N\}$ 上,目标结果即为组合分类器输出的最大值和最小值所对应的标签。学习算法的目标函数(学习错误率)定义^[14-15]为:

$$\mathcal{E} = E_{x \in X} \left[c_1 \times \left[\arg \max_{k \in \{1, 2, \dots, N\}} \{f(x, k)\} \neq y \right] + c_2 \times \left[\arg \min_{k \in \{1, 2, \dots, N\}} \{f(x, k)\} \neq z \right] \right] \quad (1)$$

其中:对于任意的谓词 π , 当 π 成立时, $[\pi]$ 取值为 1, 否则取值为 0; $\{y, z\}$ 是 $x \in X$ 的标签对; c_1 和 c_2 为加权系数,可通过调整该值来调整标签对中单个标签出现分类错误的关注程度,一般 $c_1 = c_2 = 0.5$ 。

定义 $C(i, l) = [l \notin Y_i]$, 即当 $l \in Y_i$ 时 $C(i, l) = 0$, 其余 $C(i, l) = 1 (i = 1, 2, \dots, m, l = 1, 2, \dots, N)$ 。等价表达方式如下:

定义 $C_1(i, y_i) = 0$, 即当 $l = y_i$ 时 $C_1(i, l) = 0$, 其余 $C_1(i, l) = 1 (i = 1, 2, \dots, m, l = 1, 2, \dots, N)$;

定义 $C_2(i, z_i) = 0$, 即当 $l = z_i$ 时 $C_2(i, l) = 0$, 其余 $C_2(i, l) = 1 (i = 1, 2, \dots, m, l = 1, 2, \dots, N)$ 。

对应到样本空间的训练错误率为

$$\begin{aligned} \mathcal{E} &\leq \sum_{i=1}^m \sum_{l=1}^N (w_i (C_1(i, l) \exp(f(x_i, l) - \bar{f}(x_i)) + \\ &\quad C_2(i, l) \exp(-f(x_i, l) + \bar{f}(x_i))) = \\ &Z_0 \sum_{i=1}^m \sum_{l=1}^N (w_{i,l}^{1,1} \prod_{t=1}^T \exp(h_t(x_i, l) - \bar{h}_t(x_i)) + \\ &\quad w_{i,l}^{1,2} \prod_{t=1}^T \exp(-h_t(x_i, l) + \bar{h}_t(x_i))) \quad (2) \end{aligned}$$

其中: $w_{i,l}^{1,1} = w_i C_1(i, l) / Z_0$, $w_{i,l}^{1,2} = w_i C_2(i, l) / Z_0$, $\bar{h}_t(x) = \frac{1}{N} \sum_{k=1}^N h_t(x, k)$, $\bar{f}(x) = \frac{1}{N} \sum_{k=1}^N f(x, k) = \sum_{i=1}^T \bar{h}_i(x)$, 而 Z_0 为 $w_{i,l}^{1,1} + w_{i,l}^{1,2}$ 的归一化因子, 且 $Z_0 = \sum_{i=1}^m \sum_{l=1}^N (w_i C_1(i, l) + w_i C_2(i, l))$ 。

从式(2)中提取包含 $h_1(x)$ 的项, 令

$$Z_1 = \sum_{i=1}^m \sum_{l=1}^N (w_{i,l}^{1,1} \exp(h_1(x_i, l) - \bar{h}_1(x_i)) + w_{i,l}^{1,2} \exp(-h_1(x_i, l) + \bar{h}_1(x_i))) \quad (3)$$

$$w_{i,l}^{2,1} = w_{i,l}^{1,1} \exp(h_1(x_i, l) - \bar{h}_1(x_i)) / Z_1 \quad (4)$$

$$w_{i,l}^{2,2} = w_{i,l}^{1,2} \exp(-h_1(x_i, l) + \bar{h}_1(x_i)) / Z_1 \quad (5)$$

式(2)将变为

$$Z_0 Z_1 \sum_{i=1}^m \sum_{l=1}^N (w_{i,l}^{2,1} \prod_{t=2}^T \exp(h_t(x_i, l) - \bar{h}_t(x_i)) + w_{i,l}^{2,2} \prod_{t=2}^T \exp(-h_t(x_i, l) + \bar{h}_t(x_i))) \quad (6)$$

式(6)形似式(2), 可形成递推公式。先给予最小化 Z_1 来构造 $h_1(x)$, 然后由式(4)、式(5)计算 $w_{i,l}^{2,1}$ 和 $w_{i,l}^{2,2}$, 类似的方法构造 $h_2(x)$ 等, 一般地, 可基于最小化 Z_t 来构造 $h_t(x)$,

$$Z_t = \sum_{i=1}^m \sum_{l=1}^N (w_{i,l}^{t,1} \exp(h_t(x_i, l) - \bar{h}_t(x_i)) + w_{i,l}^{t,2} \exp(-h_t(x_i, l) + \bar{h}_t(x_i))) \quad (7)$$

$$w_{i,l}^{t+1,1} = w_{i,l}^{t,1} \exp(h_t(x_i, l) - \bar{h}_t(x_i)) / Z_t \quad (8)$$

$$w_{i,l}^{t+1,2} = w_{i,l}^{t,2} \exp(-h_t(x_i, l) + \bar{h}_t(x_i)) / Z_t \quad (9)$$

设 $h_t(x)$ 把 X 划分为 n_t 段, 并对同一划分段内的目标输出值相同, 该划分对样本集也会有一个划分, 设为 $S = S_1^t \cup S_2^t \cup \dots \cup S_{n_t}^t$, $S_i^t \cap S_j^t = \emptyset (i \neq j)$, 当 $x_i \in S_j^t$ 时, $h_t(x_i)$ 对应标签的输出值 $h_t(x_i, l)$ 与 j 有关, 记为 $\alpha_t^{j,l} (j = 1, 2, \dots, n_t)$ 。合并 Z_t 中相同项, 并记 $p_t^{j,l} = \sum_{i: (x_i \in S_j^t)} w_{i,l}^{t,1}$, $q_t^{j,l} = \sum_{i: (x_i \in S_j^t)} w_{i,l}^{t,2}$, 此时,

由式(7)可得:

$$Z_t = \sum_{j=1}^{n_t} \sum_{l=1}^N (p_t^{j,l} \exp(\alpha_t^{j,l} - \frac{1}{N} \sum_{k=1}^N \alpha_t^{j,k}) + q_t^{j,l} \exp(-\alpha_t^{j,l} + \frac{1}{N} \sum_{k=1}^N \alpha_t^{j,k})) \quad (10)$$

该式求极值点很困难^[12], 由式(10)可知以下两式:

$$\begin{aligned} &\sum_{j=1}^{n_t} \sum_{l=1}^N (p_t^{j,l} \exp(\alpha_t^{j,l} - \frac{1}{N} \sum_{k=1}^N \alpha_t^{j,k})) \\ &\sum_{j=1}^{n_t} \sum_{l=1}^N (q_t^{j,l} \exp(-\alpha_t^{j,l} + \frac{1}{N} \sum_{k=1}^N \alpha_t^{j,k})) \end{aligned}$$

分别在 $\alpha_t^{j,l} = -\ln(p_t^{j,l})$ 和 $\alpha_t^{j,l} = \ln(q_t^{j,l})$ 处取得极值, 将此二极值点加权平均 $\alpha_t^{j,l} = -0.5 \ln(p_t^{j,l}) + 0.5 \ln(q_t^{j,l})$, 不妨令 Z_t 在 $\alpha_t^{j,l} = 0.5 \ln(q_t^{j,l}/p_t^{j,l})$ 处取值, 将 $\alpha_t^{j,l} = 0.5 \ln(q_t^{j,l}/p_t^{j,l})$ 代入式(10)可得

$$Z_t \geq \sum_{j=1}^{n_t} \sum_{l=1}^N (2 \sqrt{p_t^{j,l} q_t^{j,l}}) \quad (11)$$

即当 $\alpha_t^{j,l} = 0.5 \ln(q_t^{j,l}/p_t^{j,l})$ 时, Z_t 取到极小值 $Z_t = \sum_{j=1}^{n_t} \sum_{l=1}^N (2 \sqrt{p_t^{j,l} q_t^{j,l}})$, 由上述分析可知, 因为

$$Z_t = \sum_{j=1}^{n_t} \sum_{l=1}^N (2 \sqrt{p_t^{j,l} q_t^{j,l}}) \leq \sum_{j=1}^{n_t} \sum_{l=1}^N (p_t^{j,l} + q_t^{j,l}) =$$

$$\sum_{i=1}^m \sum_{l=1}^N (\omega_{i,l}^{1,1} + \omega_{i,l}^{1,2}) = 1$$

当且仅当 $p_t^{j,l} = q_t^{j,l}$ 对所有 $j \in \{1, 2, \dots, n_t\}$ 和 $l \in \{1, 2, \dots, N\}$ 都成立时, $Z_t = 1$ 。因此,一般情况下有 $Z_t < 1$, 即算法的训练错误率可随着弱分类器个数逐渐增加而逐渐降低, 并且

有 $\varepsilon_T = \prod_{t=0}^T Z_t$, 因此, 在 $\alpha_t^{j,l} = 0.5 \ln(q_t^{j,l}/p_t^{j,l})$ 处取 $Z_t =$

$\sum_{j=1}^{n_t} \sum_{l=1}^N (2 \sqrt{p_t^{j,l} q_t^{j,l}})$ 是合理的。

通过以上详细的推导, 从理论上证明了构造类标签分属两个标签集的双标签分类问题的标签匹配集成学习算法的正确性, 于是, 可得如下基于双标签集的标签匹配集成学习算法:

步骤1 初始化权值: $\omega_{i,l}^{1,1} = C_1(i, l)/Z_0, \omega_{i,l}^{1,2} = C_2(i,$

$l)/Z_0, \omega_{i,l}^{1,1} + \omega_{i,l}^{1,2}$ 的归一化因子 $Z_0 = \sum_{i=1}^m \sum_{l=1}^N (C_1(i, l) + C_2(i, l))$ 。

步骤2 DO FOR $t = 1, 2, \dots, T$

①基于有权值 $\omega_{i,l}^{1,1}, \omega_{i,l}^{1,2}$ 的 S 训练 t 个弱分类器:

a) 对划分 $S = S_1^t \cup S_2^t \cup \dots \cup S_{n_t}^t$, 计算 $p_t^{j,l} = \sum_{i: (x_i \in S_j^t)} \omega_{i,l}^{t,1}$,

计算 $q_t^{j,l} = \sum_{i: (x_i \in S_j^t)} \omega_{i,l}^{t,2} (l = 1, 2, \dots, N)$ 。

b) 定义 $h_t(x): x \in S_j^t$ 时, $h_t(x, l) = 0.5 \ln(q_t^{j,l}/p_t^{j,l}) (j = 1, 2, \dots, n_t)$ 。

c) 选取 $h_t(x)$: 最小化 $Z_t = \sum_{j=1}^{n_t} \sum_{l=1}^N (2 \sqrt{p_t^{j,l} q_t^{j,l}})$ 来选 $h_t(x)$ 。

②调整权值:

$\omega_{i,l}^{t+1,1} = (\omega_{i,l}^{t,1}/Z_t) \exp(h_t(x_i, l) - \bar{h}_t(x_i))$; $l = 1, 2, \dots, N$

$\omega_{i,l}^{t+1,2} = (\omega_{i,l}^{t,2}/Z_t) \exp(-h_t(x_i, l) + \bar{h}_t(x_i))$; $l = 1, 2, \dots, N$

步骤3 IF $t = T$, 退出循环;

组合 T 个弱分类器为强分类器: $f(x, l) = \sum_{t=1}^T h_t(x, l)$,

输出标签对 $(\arg \max_{k \in \{1, 2, \dots, N\}} \{f(x, k)\}, \arg \min_{k \in \{1, 2, \dots, N\}} \{f(x, k)\})$ 。

该算法的时间复杂度与弱分类器的构造方法有关, 本文算法是一种较快的算法, 基于单个特征采用阈值法来构造弱分类器时, 算法时间复杂度为 $O(mdT)$, 其中: m 为训练样本数, d 为样本特征个数, T 为弱分类器的个数。该算法的空间复杂度则为 $O(mN)$, 其中: m 为训练样本数, $N = K + W$, K 为第一个标签集的标签个数, W 为第二个标签集的标签个数。

2 标签匹配算法的度量方法

针对多标签学习系统的性能评价, 研究者们相继提出了一系列多标签评价指标, 总体来看分为基于样本^[8]和基于标签^[4]两种类型, 基于样本的多标签评价指标首先衡量分类器在单个测试样本上的分类效果, 然后返回其在整个测试集上的“均值(mean value)”作为最终的测试结果。基于标签的方法则是通过对每个单独的标签的预测结果进行度量, 然后再对所有标签的结果取平均。

本文针对标签匹配学习算法数据结构的复杂性及标签集的特殊性, 选择基于样本的 Subset accuracy 评价指标。给定多标签分类器 $h(x)$, 组合分类器仍然为 $f(x)$, 以及多标签测试集 $S = \{(x_i, (y_i, z_i)) | 1 \leq i \leq P\}$, 其中 $y_i \in L_1 = \{a_1, a_2,$

$\dots, a_K\}$, $z_i \in L_2 = \{b_1, b_2, \dots, b_W\}$, 即 $x_i \in X$ 的类别为来自两个不同的标签集的标签对 $\{y_i, z_i\}$ 。

$$\text{subset_accuracy} = \frac{1}{P} \sum_{i=1}^P [\max(f(x_i)) = y_i \& \min(f(x_i)) = z_i]$$

该评价指标用于考察预测的标签集合与真实标签集合完全吻合的样本占测试集的比例情况。该指标取值越大则系统性能越优, 其最优取值为1。当标记空间中单标签集合包含大量类别标记或多标签集合出现大量标签组合时, 学习系统往往难以给出与真实的标记集合完全吻合的预测, 此时该评价指标的取值将会很低^[16]。

3 实验与分析

3.1 实验方法和数据

实验使用 Matlab R2013a 开发平台, 实验数据集为 Scene、Yeast^[12]和 Image^[13]的三个常用的多标签数据集, 具体描述见表1。

表1 原多标签数据集

数据集名称	样本容量	属性数	类别数	标签基数
Scene	2407	294	6	1.074
Yeast	2417	103	14	4.237
Image	2000	294	5	1.236

首先, 对样本数据进行预处理。由于尚无公认的常用的多标签数据包含两个或多个标签集, 本文针对以上三个常用多标签数据集进行人为改进, 具体方法如下: 1) 将一组多标签集人为划分为两个标签集; 2) 对标签集进行重组, 必须满足每个标签集中有且只有一个原有标签; 3) 删除冗余样本示例, 保证一个原有样本示例得到一个新样本。最终使用数据集为: Scene 数据集的原标签集含有6个标签, 划分为2个分别含4个标签和2个标签的子标签集; Yeast 数据集的原标签集含有14个标签, 划分为2个分别含7个标签的子标签集; Image 数据集的原标签集含有5个标签, 划分为2个分别含3个标签和2个标签的子标签集(具体见表2)。

表2 实验数据集

数据集名称	样本容量	属性数	标签集1	标签集2
Scene	152	294	4	2
Yeast	1858	103	7	7
Image	367	294	3	2

其次, 确定训练方法。基于单个属性(特征)来构造弱分类器, 实验中采取4段划分, 其3个分段阈值为: 含有相同标签的样本均值、最大值与该均值的平均, 以及最小值与该均值的平均。实验时对数据进行6:4比例随机划分得到训练数据集和测试数据集, 训练40个弱分类器(即 $M = 40$), 随机实验10次统计平均测试错误。下文用 AdaBoost.LM 代表基于双标签集的标签匹配集成学习算法。

3.2 实验结果与分析

本文采用多标签学习的评价指标为基于样本的 Subset accuracy。基于评价指标的定义及 AdaBoost 算法固有的性质, 实验结果的值越大越好, 并随着弱分类器个数的增加而呈逐渐增大的趋势。具体实验结果如下:

首先, 对指标 Subset accuracy 在数据集 Scene、Yeast 和

Image 的训练结果的趋势图进行分析,并将 AdaBoost.LM 的趋势图与 AdaBoost.MH 的趋势图进行比较。图 1~3 显示,两个算法都具有一定的分类效果,并且正确率随着弱分类器个数的增加而增大,整体上 AdaBoost.LM 算法相比 AdaBoost.MH 算法的趋势图略显平滑,但波动比较大,波动较大的原因在于数据集的选取可能存在一定的不合理之处,但从趋势图的走向上能够说明本文算法的正确性。

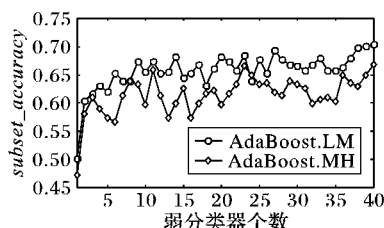


图1 在 Scene 数据集上的实验结果

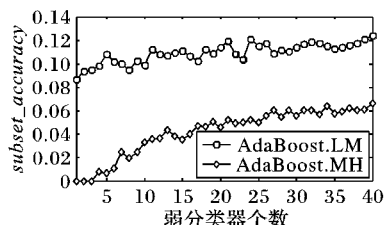


图2 在 Yeast 数据集上的实验结果

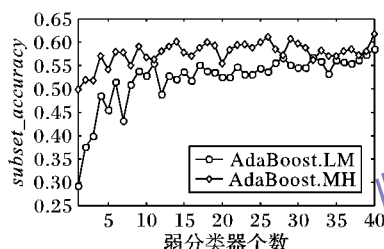


图3 在 Image 数据集上的实验结果

其次,将本文算法与现有多标签学习算法在指标 Subset accuracy 上的实验结果进行比较和分析。由表 3 可知,在 Scene 和 Yeast 数据集上,AdaBoost.LM 算法的正确率都要高于 AdaBoost.MH 算法、ML- k NN 和 RankSVM 算法,并且在 Scene 数据集上分别提升了 4.63%、2.83%、4.31%,在 Yeast 数据集上分别提升了 5.78%、12.29%、8.23%;在 Image 数据集上的正确率略差于 RankSVM,优于 ML- k NN,AdaBoost.MH 分类效果最好。四种算法在 Yeast 数据集上结果差的原因在于 AdaBoost.LM 算法是在 7×7 的标签空间作预测,其余三种算法则是在 2^{14} 标签空间作预测,所以后三种算法的分类准确率要低得多。实验结果表明,本文算法是相对稳定且有效的。

表3 四个数据集上 Subset accuracy 指标性能比较

测试数据集	AdaBoost.LM ($M=40$)	AdaBoost.MH ($M=40$)	ML- k NN ($k=7$)	Rank-SVM (RBF)
Scene	0.7152	0.6689	0.6869	0.6721
Yeast	0.1240	0.0662	0.0011	0.0417
Image	0.5850	0.6177	0.5558	0.5986

4 结语

针对双标签集学习问题的特殊情况,本文提出了一种基于双标签集的标签匹配集成学习算法。该算法经多次迭代更新权重,不断增加第一个标签集中正确标签的权重,同时缩小

第二个标签集中正确标签的权重,最终组合分类器并输出最大值和最小值所对应的标签即为正确的标签对,算法的优势在于降低了算法的时间复杂度和空间复杂度,不足之处在于算法忽略了标签间的相关性,并且对数据集的要求较特殊,在实际应用中仍有很大的改进空间,今后的工作将侧重于对标签相关性作进一步研究。

参考文献:

- [1] TSOU MAKAS G, KATAKIS I. Multi-label classification: an overview [J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1-13.
- [2] BOUTELL M R, LUO J, SHEN X, et al. Learning multi-label scene classification [J]. Pattern Recognition, 2004, 37(9): 1757-1771.
- [3] FÜRNKRANZ J, HÜLLERMEIER E, LOZA MENCÍA E, et al. Multilabel classification via calibrated label ranking [J]. Machine Learning, 2008, 73(2): 133-153.
- [4] TSOU MAKAS G, VLAHAVAS I. Random k -labelsets: an ensemble method for multilabel classification[C]// ECML 2007: Proceedings of the 18th European Conference on Machine Learning, LNCS 4701. Berlin: Springer-Verlag, 2007: 406-417.
- [5] ZHANG M L, ZHOU Z H. A k -nearest neighbor based algorithm for multi-label classification [C]// Proceedings of the 2005 IEEE International Conference on Granular Computing. Piscataway: IEEE, 2004, 2: 718-721.
- [6] ELISSEEFF A, WESTON J. A kernel method for multi-labeled classification [C]// NIPS 02: Proceedings of Advances in Neural Processing Systems. Cambridge: MIT Press, 2002: 681-687.
- [7] SCHAPIRE R E, SINGER Y. Improved boosting algorithms using confidence-rated predictions [J]. Machine Learning, 1999, 37(3): 297-336.
- [8] SCHAPIRE R E, SINGER Y. Boostexter: a boosting based system for text categorization [J]. Machine Learning, 2000, 39(2/3): 135-168.
- [9] ZHANG M-L, ZHANG K. Multi-label learning by exploiting label dependency[C]// KDD'10: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2010: 999-1007.
- [10] DEMBCZYNSKI K, WAEGEMAN W, CHENG W, et al. On label dependence in multi-label classification[C]// Proceedings of the 2nd International Workshop on Learning from Multi-Label Data. Haifa: [s. n.], 2010: 5-12.
- [11] HUANG S-J, ZHOU Z-H. Multi-label learning by exploiting label correlations locally[C]// Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2012: 949-955.
- [12] Mulan: a Java library for multi-label learning [EB/OL]. [2014-01-16]. <http://mulan.sourceforge.net/datasets.html>.
- [13] ZHANG M-L. Data sets used in multi-instance learning [EB/OL]. [2013-12-05]. <http://cse.seu.edu.cn/people/zhangml/Resources.htm#data>.
- [14] FU Z. Cost-sensitive AdaBoost algorithm for multi-class classification problems [J]. Acta Automatica Sinica, 2011, 37(8): 973-983. (付忠良. 多分类问题代价敏感 AdaBoost 算法[J]. 自动化学报, 2011, 37(8): 973-983.)
- [15] FU Z. An ensemble learning algorithm for direction prediction [J]. Shanghai Jiaotong University, 2012, 46(2): 250-258. (付忠良. 一种用于方向预测的集成学习算法[J]. 上海交通大学学报, 2012, 46(2): 250-258.)
- [16] TSOU MAKAS G, KATAKIS I, VLAHAVAS I. Mining multi-label data[M]// Data Mining and Knowledge Discovery Handbook. Berlin: Springer-Verlag, 2010: 667-686.