Journal of Computer Applications
计算机应用,2018,38(3):650-654,660
ISSN 1001-9081
CODEN JYIIDU
2018-03-10
http://www.joca.cn

文章编号:1001-9081(2018)03-0650-05
DOI:10.11772/j.issn.1001-9081.2017092226

AdaBoost 的多样性分析及改进

王玲娣,徐 华*
(江南大学 物联网工程学院,江苏 无锡 214122)
(* 通信作者电子邮箱 joanxh2003@163.com)

摘 要:针对 AdaBoost 算法下弱分类器间的多样性如何度量问题以及 AdaBoost 的过适应问题,在分析并研究了4种多样性度量与 AdaBoost 算法的分类精度关系的基础上,提出一种基于双误度量改进的 AdaBoost 方法。首先,选择 Q 统计、相关系数、不一致度量、双误度量在 UCI 数据集上进行实验。然后,利用皮尔逊相关系数定量计算多样性与测试误差的相关性,发现在迭代后期阶段,它们都趋于一个稳定的值;其中双误度量在不同数据集上的变化模式固定,它在前期阶段不断增加,在迭代后期基本上不变,趋于稳定。最后,利用双误度量改进 AdaBoost 的弱分类器的选择策略。实验结果表明,与其他常用集成方法相比,改进后的 AdaBoost 算法的测试误差平均降低1.5个百分点,最高可降低4.8个百分点。因此,该算法可以进一步提高分类性能。

关键词:多样性;AdaBoost;集成学习;双误度量;弱分类器
中图分类号:TP181 **文献标志码**:A

Diversity analysis and improvement of AdaBoost

WANG Lingdi, XU Hua*
(*School of Internet of Things Engineering*, *Jiangnan University*, *Wuxi Jiangsu* 214122, *China*)

Abstract: To solve the problem of how to measure diversity among weak classifiers created by AdaBoost as well as the over-adaptation problem of AdaBoost, an improved AdaBoost method based on double-fault measure was proposed, which was based on the analysis and study of the relationship between four diversity measures and the classification accuracy of AdaBoost. Firstly, Q statistics, correlation coefficient, disagreement measure and double-fault measure were selected for experiment on the data sets from the UCI (University of CaliforniaIrvine Irvine) machine learning repository. Then, the relationship between diversity and ensemble classifier's accuracy was evaluated with Pearson correlation coefficient. The results show that each measure tends to a stable value in the later stage of iteration; especially double-fault measure changes similarly on different data sets, increasing in the early stage and tending to be stable in the later stage of iteration. Finally, a selection strategy of weak classifier based on double-fault measure was put forward. The experimental results show that compared with the other commonly used ensemble methods, the test error of the improved AdaBoost algorithm is reduced by 1.5 percentage points in average, and 4.8 percentage points maximally. Therefore, the proposed algorithm can improve classification performance.

Key words: diversity; AdaBoost; ensemble learning; double-fault measure; weak classifier

0 引言

集成学习是当前机器学习的热点研究方向之一,和传统单个分类器的构造目的不同,它并非力求得到单一最优分类器,而是按照一定策略集成一组个体分类器。在两种经典的集成算法:Boosting[1]和 Bagging[2]被提出之后,研究者又陆续提出了大量的集成学习算法。其中 Boosting 算法可将粗糙的、不太正确的、简单的初级预测方法,按照一定的规则构造出一个复杂的,精确度很高的预测方法,但是很难运用于实际中;AdaBoost[3]的出现有效地解决了这一问题,因此 AdaBoost 成为了 Boosting 家族的代表算法,受到极大的关注,成功应用于声音文件检索[4]、人脸识别[5]、癌症诊断[6]及目标检测[7-8]等实际问题中。

集成学习主要有两个阶段:一是基分类器的生成;二是组合策略的选择。将相同的基分类器进行集成是无意义的,因为组合而成的分类器与基分类器的分类结果必然相同。所以基分类器之间要存在差异,即分类器多样性。Krogh 等[9]证明,集成的泛化误差是由个体分类器的平均泛化误差和平均差异度决定的。虽然目前已存在多种多样性度量方式,但是关于它的严格定义并不统一[10-11],只是可以从大量研究资料中获知,多样性有益于集成方法的设计,如:2012 年,文献[12]使用遗传算法组合不同的多样性用于选择性集成;而文献[13]于2014 年通过向量空间模型形象地论证了多样性的有效性;2015 年文献[14]明确提到多样性是集成学习成功的重要条件;文献[15]在 2016 年研究了很可能接近正确的(Probably Approximately Correct, PAC)学习框架下多样性对基于投票组合策略的集成方法泛化能力的影响。多样性对于 AdaBoost 来说同样重要,文献[16]提出一种基于随机子空间和 AdaBoost 自适应集成方法,将随机子空间融合到 AdaBoost 的训练过程中,目的就是增加 AdaBoost 的多样性。文献[17]

收稿日期:2017-09-13;**修回日期**:2017-10-15。 **基金项目**:江苏省自然科学基金资助项目(BK20140165)。
作者简介:王玲娣(1991—),女,安徽宿州人,硕士研究生,主要研究方向:机器学习、数据挖掘; 徐华(1978—),女,江苏无锡人,副教授,博士,主要研究方向:计算机智能、车间调度、大数据。

详细总结了 AdaBoost 的发展,并指出它的进一步研究方向之一是其弱分类器的多样性研究,因为有关分类器多样性的研究,有效结论太少,有待深入与完善。也有文献[18]研究了多样性度量在 AdaBoost. M2 算法下的变化,得到一些规律,但如何使用这些规律以及最终能否提高集成性能,并没给出答案。

针对上述问题,本文研究了 4 种成对型多样性度量在 AdaBoost 算法下的变化;并利用皮尔逊相关系数定量分析多样性度量和分类性能之间的关系,发现双误度量(Double Fault, DF) 变化模式固定——先增加后平缓;进一步,提出了一种基于 DF 改进的 AdaBoost 算法。结果表明改进后的算法可以抑制 AdaBoost 的过适应现象,降低错误率。

1　成对型多样性度量

关于成对多样性的研究主要集中在以下 3 个方面: 1) 多样性的度量方法; 2) 多样性度量与集成学习精度的关系; 3) 如何利用多样性度量更好地选择分类器来构建集成系统,以提高集成学习的性能。本文按照上述思路,先介绍 4 个成对型多样性度量方法,然后研究这四种多样性度量与 AdaBoost 产生的分类器精度有怎样的关系,最后利用 DF 改进 AdaBoost 算法。

成对型多样性度量是定义在两个分类器上的,假设分类器集合 $H = \{h_1, h_2, \cdots, h_m\}$, h_i 和 $h_j (i \neq j)$ 为两个不同的分类器,它们对同一组样本分类情况组合如表 1 所示,其中样本总数为 n。表 1 中, n^{11} (n^{00}) 代表被 h_i 和 h_j 共同正确(错误) 分类的样本数, n^{10} 代表被 h_i 正确分类、h_j 错误分类的样本数, n^{01} 代表被 h_i 错误分类、h_j 正确分类的样本数,并且它们满足式(1) :

$$n^{11} + n^{00} + n^{10} + n^{01} = n \tag{1}$$

表 1　两个分类器的分类结果组合

Tab. 1　Result combination of two classifiers

分类器分类结果	h_i correct(1)	h_i incorrect(0)
h_j correct(1)	n^{11}	n^{10}
h_j incorrect(0)	n^{01}	n^{00}

下面将分别介绍 4 种成对型多样性度量。

1) Q 统计。

Q 统计(Q-statistics, Q) [19] 源于统计学,它的计算公式如下:

$$Q_{i,j} = \frac{n^{11}n^{00} - n^{10}n^{01}}{n^{11}n^{00} + n^{10}n^{01}} \tag{2}$$

由式(2) 可知 Q 的取值范围是 $[-1, 1]$。当两个分类器的分类结果趋于一致时, Q 值为正,否则为负,完全相同时为 1,完全不同时为 -1。

2) 相关系数。

相关系数(Correlation coefficient, ρ) [20] 源于统计学, ρ 的取值范围为 $[-1, 1]$,计算公式如下:

$$\rho_{i,j} = \frac{n^{11}n^{00} - n^{10}n^{01}}{\sqrt{(n^{11} + n^{00})(n^{01} + n^{00})(n^{11} + n^{01})(n^{10} + n^{00})}} \tag{3}$$

3) 不一致度量。

不一致度量(Disagreement Measure, DM) [21] 衡量的是两个分类器分类结果不一致的程度,它的值越大,表明两个分类器的多样性越大,取值范围为 $[0, 1]$,计算公式如下所示:

$$DM_{i,j} = (n^{10} + n^{01}) / n \tag{4}$$

4) 双误度量。

双误度量(DF) [22] 关注的是两个分类器在相同样本上出错的情况,取值范围 $[0, 1]$,最差的情况是两个分类器错误率都是 100%,此时 DF 的值为 1,分类器的正确性与多样性同时降到最低。计算公式如下:

$$DF_{i,j} = n^{00} / n \tag{5}$$

要评价一组分类器 $H = \{h_1, h_2, \cdots, h_m\}$ 的多样性,需要计算每对分类器之间多样性的平均值,见式(6) 。其中 $\overline{Div}$ 代表一组分类器的整体多样性, Div 代表两个分类器之间的多样性。

$$\overline{Div} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} Div_{i,j} \tag{6}$$

2　AdaBoost 算法

2.1　算法描述

能否使用多个弱分类器来构建一个强分类器? 这是一个有趣的理论问题。“弱”意味着分类器的性能仅仅比随机猜测略好,而“强”则表明分类器表现不错。AdaBoost 即脱胎于上述理论问题。AdaBoost 算法是一个迭代过程,原理是: 算法运行过程中会给训练样本赋予权重,一开始,初始化成相等值,然后根据弱分类器学习算法训练第一个弱分类器,接着根据该分类器的加权误差更新样本权重,降低被正确分类的样本权重,提高被错误分类的样本权重。基于新的样本权重分布,继续训练弱分类器。如此往复,便可得到一组弱分类器,每个弱分类器也有一个权重,代表它在最后集成中的重要性。

下面将具体介绍样本权重的更新过程。

对于二分类问题,令 $S = \{(\boldsymbol{x}_i, y_i) \mid i = 1, 2, \cdots, n\}$ 表示训练样本集,其中 $y_i \in \{-1, 1\}$ 代表样本标签。$\boldsymbol{D}_t$ 表示第 t 轮迭代的样本分布矢量,初始化时, $\boldsymbol{D}_1 = \{1/n, 1/n, \cdots, 1/n\}$。在 AdaBoost 算法中,基分类器 h_t 的重要性和它在样本权重分布上的错误率 ε_t 相关,也被称为加权误差,定义如下:

$$\varepsilon_t = \sum_{i=1}^{n} \boldsymbol{D}_t [\![h_t(\boldsymbol{x}_i) \neq y_i]\!] \tag{7}$$

式(7) 中, $[\![h(\boldsymbol{x}_i) \neq y_i]\!]$ 是指示函数,表达式 $h(\boldsymbol{x}_i) \neq y_i$ 为真时其值等于 1,否则等于 0。h_t 的重要性即权重,计算如下:

$$\alpha_t = \ln\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right) \tag{8}$$

接下来,根据 α_t 来更新样本($\boldsymbol{x}_i, y_i$) 权重,见式(9) :

$$\boldsymbol{D}_{t+1}(i) = (\boldsymbol{D}_t(i) \exp(-y_i h_t(\boldsymbol{x}_i) \alpha_t)) / Z_t \tag{9}$$

其中: $Z_t = \sum_{i=1}^{n} \boldsymbol{D}_t(i) \exp(-y_i h_t(\boldsymbol{x}_i) \alpha_t)$ 是归一化因子。最终 AdaBoost 将每个分类器的预测值基于 α_t 进行加权线性组合,这种方式使得 AdaBoost 可以惩罚那些错误率很高的分类器。强分类器的输出如下所示:

$$H(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{t=1}^{m} \alpha_t h_t(\boldsymbol{x})\right) \tag{10}$$

2.2　算法问题分析

由上述可知,在 AdaBoost 的训练过程中,分类器的重心

将被转移到比较难分类的样本上，这也是 AdaBoost 可以将“弱”变“强”的原因，但是如果训练样本中存在大量的噪声或者样本数据错误，就会出现过适应现象。因为这些噪声或错误点是难分类点，随着迭代的进行，它们的权值会呈指数增长，在这样的样本权重分布下，训练产生的弱分类器的错误率相对增大，从而它们在最后的加权组合中作用变得非常小。而且由于归一化，已经被正确分类的样本在过适应的情况下，权重变得非常小，可能会出现被忽视的情况，那些被迭代前期产生的弱分类器正确分类的样本，很有可能在最后组合分类器判断下的结果是错误的，最终导致退化，影响集成性能。所以，在迭代后期，要控制弱分类器对困难样本的关注，避免样本分布扭曲。

3 改进算法

针对上述问题，本文提出基于 DF 改进的 AdaBoost 算法，通过改进弱分类器选择策略，控制弱分类器之间的 DF 值，来避免对困难样本的过分关注。

3.1 改进弱分类器的选择策略

AdaBoost 算法使用单层决策树训练弱分类器，它的一个最基本理论上的性质是可以降低训练误差，Schapire 等[3]给出了 AdaBoost 训练误差的上界，见式(11)：

$$\frac{1}{n}\left[\!\left[H(x_i) \neq y_i \right]\!\right] \leqslant \prod_{t=1} Z_t \tag{11}$$

该公式提供了一个可以降低训练误差的重要方法，即最小化 $\prod_{t=1} Z_t$。然而，要实现它的精确全局最小化，是个非常复杂的优化问题，因为每加入一个新的子分类器，可能都需要修改已有子分类器的集成方式，因此 Schapire 等指出可以采用贪婪策略，在每一轮迭代中最小化 Z，最后证明了在每轮迭代中应选择加权误差最小的弱分类器加入集成。

假设 AdaBoost 第 t 轮迭代的候选弱分类器集合为 $C = \{h_t^1, h_t^2, \cdots, h_t^k\}$，则传统 AdaBoost 的弱分类器的选择策略是：

$$\arg\min_{h \in C}(\varepsilon_t(h)) \tag{12}$$

但是 AdaBoost 算法并没有考虑这种情况：候选的弱分类器集合中有两个或者多个弱分类器的加权误差相同(或者是相差很小)，但是这些弱分类器与已经加入集成的弱分类器间的差异性有所不同，而最终选择的弱分类器的多样性不是最好的，这样就会影响集成泛化能力；而且 AdaBoost 会出现过适应就是因为对于某些样本过于集中关注，当增加了分类器间的多样性，就可以适度分散这种集中关注度。因此，需要在选择弱分类器的时候，加入多样性的判断。首先分析相关系数 ρ，由式(3)可知，当两个基分类器的分类结果趋向不同时，ρ 值为负，即 $n^{10}n^{01} > n^{11}n^{00}$，当增大 $n^{10}n^{01}$ 时，$n^{11}n^{00}$ 相应地降低，但无法保证降低的是 n^{00}，从而无法保证平均分类精度，这意味着 ρ 与集成的分类性能关联并不紧密，同时它的计算公式相对于其他三个多样性度量公式最为复杂。Q 统计与 ρ 计算公式的分子相同，可以把 Q 统计看作是 ρ 的一种简化运算，因此 Q 存在着与 ρ 相同的问题。接下来分析不一致度量 DM，由式(4)可知，DM 越大，基分类器间的多样性越大，但同时平均精度也越低。增加多样性的目的是为了进一步提高集成算法的分类精度，所以这三种多样性度量从理论上分析都是不适合 AdaBoost 的。本文提出一种基于 DF 改进的弱分类器选择策略，如下所示：

$$\arg\min_{h \in C}(w_1\varepsilon_t(h) + w_2 DF_{t-1,t}) \tag{13}$$

其中：$w_1 + w_2 = 1$，分别代表加权误差与 DF 值在选择策略中的比重；$DF_{t-1,t}$ 表示候选弱分类器与上一轮迭代中已被选中弱分类器之间的 DF 度量值。由式(5)可知，DF 变小，表示 n^{00} 减少了，相对的 $n^{11} + n^{01} + n^{10}$ 就会增加。若增加的是 n^{11}，那么表明集成分类器的正确率提高了，若增加的是 $n^{10} + n^{01}$，则表明基分类器间的差异性增大，集成多样性提高了。对 AdaBoost 来说，DF 变小意味着两个弱分类器共同错分的样本数少了，它们各自有自己关注的困难样本，就不会对某些样本过于集中关注，避免某些样本的权值过大，进而抑制过适应。

关于 w_1、w_2 的取值，在 AdaBoost 过程中不是固定不变的，而是根据 AdaBoost 的训练情况动态调整。w_2 为已经加入集成的前 $t-1$ 个弱分类器间的平均 DF 值，根据式(6)可得：

$$w_2 = \frac{2}{(t-1)(t-2)} \sum_{i=1}^{t-2} \sum_{j=i+1}^{t-1} DF_{i,j} \tag{14}$$

$$w_1 = 1 - w_2 \tag{15}$$

根据式(13)和(14)可知，若是迭代中的整体平均 DF 值有增大的趋势，就会相应地增加 $DF_{t-1,t}$ 在选择标准中比重，控制对共同错分样本过分关注，从而达到抑制过适应的目的，否则，加权误差依然是选择标准中的重要因素。这样就能在弱分类器增加多样性的同时保证其准确性。式(11)已经说明了 AdaBoost 最终模型的训练集误差是有上界的，这表明该算法理论上可以收敛到误差边界；而修改后算法并没有破坏 AdaBoost 算法框架，依然按照原来贪心策略进行迭代，这一点保证了算法的可收敛性。

3.2 基于 DF 的弱分类器算法

根据单层决策树算法训练出的弱分类器的函数表达式如下：

$$h(\boldsymbol{x}_i) = \begin{cases} 1, & \text{sign}(\theta - x_{i,j})\, b < 0 \\ -1, & \text{其他} \end{cases} \tag{16}$$

其中：$b \in \{-1, 1\}$ 是一个指示不等号方向的参数，θ 是特征阈值。假设训练样本按照第 j 维特征值升序排列，使得 $x_{1,j} \leqslant x_{2,j} \leqslant \cdots \leqslant x_{m,j}$，则 θ 的取值范围如下：

$$\Theta_j = \{x_{1,j} - 1, x_{m,j} + 1\} \cup \left\{ \frac{x_{i,j} + x_{i+1,j}}{2} \mid i = 1, 2, \cdots, m-1 \right\} \tag{17}$$

则基于 DF 的弱分类器算法(Weak Learning algorithm based on Double Fault, WLDF)如下：

WLDF 算法。

输入：训练集 S，样本分布 D_t。

初始化：$EDF_{\min} = +\infty$，$h^* = \text{null}$

```
1)  根据式(14)和(15)计算出 w1 和 w2
2)  for 样本的每一特征 j:
3)    由式(17)计算 θ 取值范围 Θj
4)    for 每一个阈值 θ ∈ Θj:
5)      for 不等号 b ∈ {-1,1}:
6)        训练出一个弱分类器 ht
7)        计算 EDF = w1εt + w2DFt-1,t
8)        if EDF < EDFmin:
9)          EDFmin = EDF
10)         h* = ht
11)     end for
12)   end for
13) end for
```

输出：h^*。

4　实验

实验分为实验一和实验二。实验一研究 Q、ρ、DM、DF 四种多样性度量在 AdaBoost 算法迭代过程中的变化规律及其与集成泛化能力的相关性,实验二验证 WLDF 算法的有效性。实验机器配置为: Windows 10,内存 4 GB,CPU 3.2 GHz,算法基于 Python 2.7 实现。实验数据来自 UCI(University of CaliforniaIrvine Irvine) 数据库(http: //archive. ics. uci. edu/ml/datas-ets. html) ,具体信息见表 2。

表 2　实验数据集信息

Tab. 2　Information of data sets

数据集序号	数据集	实例数	维数
1	Balance	625	4
2	Chess	3 196	36
3	German	1 000	24
4	Heart	270	14
5	Pima	768	9
6	Sonar	208	61

4.1　实验一的结果及分析

为充分使用数据,实验一采用 10 折交叉验证,实验结果如图 1 所示。图 1 分别呈现了 6 个数据集的多样性度量变化与测试误差变化。图 1(a) ~(d) 4 个子图分别呈现了 ρ、Q、DM、DF 的变化情况,其中纵坐标是多样性度量值,横坐标是迭代次数(也是基分类器数目) ,10 次实验的每一次结果画一条实线表示,以此观察 10 次结果的变化规律是否相同。子图(e) 中纵坐标是 10 次实验结果的平均测试误差。

首先,整体观察图 1,可以看到四种多样性度量都在弱分类器数目增加到一定程度时,趋近一个值。观察图 1 中 German、Heart、Pima 以及 Sonar 数据集的实验结果,子图(a) ~(d) 中前阶段的线条很乱,这表明 10 次实验结果差别大,这时观察相应的子图(e) ,测试误差的变化很激烈,虽然总体方向是下降,但是曲线波动很大。而当多样性度量平稳变化时,见图 1 中 Balance 和 Chess 数据集的实验结果,四种多样性度量的 10 次结果几乎在一条线上,而再看测试误差变化,几乎没有波动,持续下降。这样定性看来,多样性与组合分类器精度之间有一定的关联。

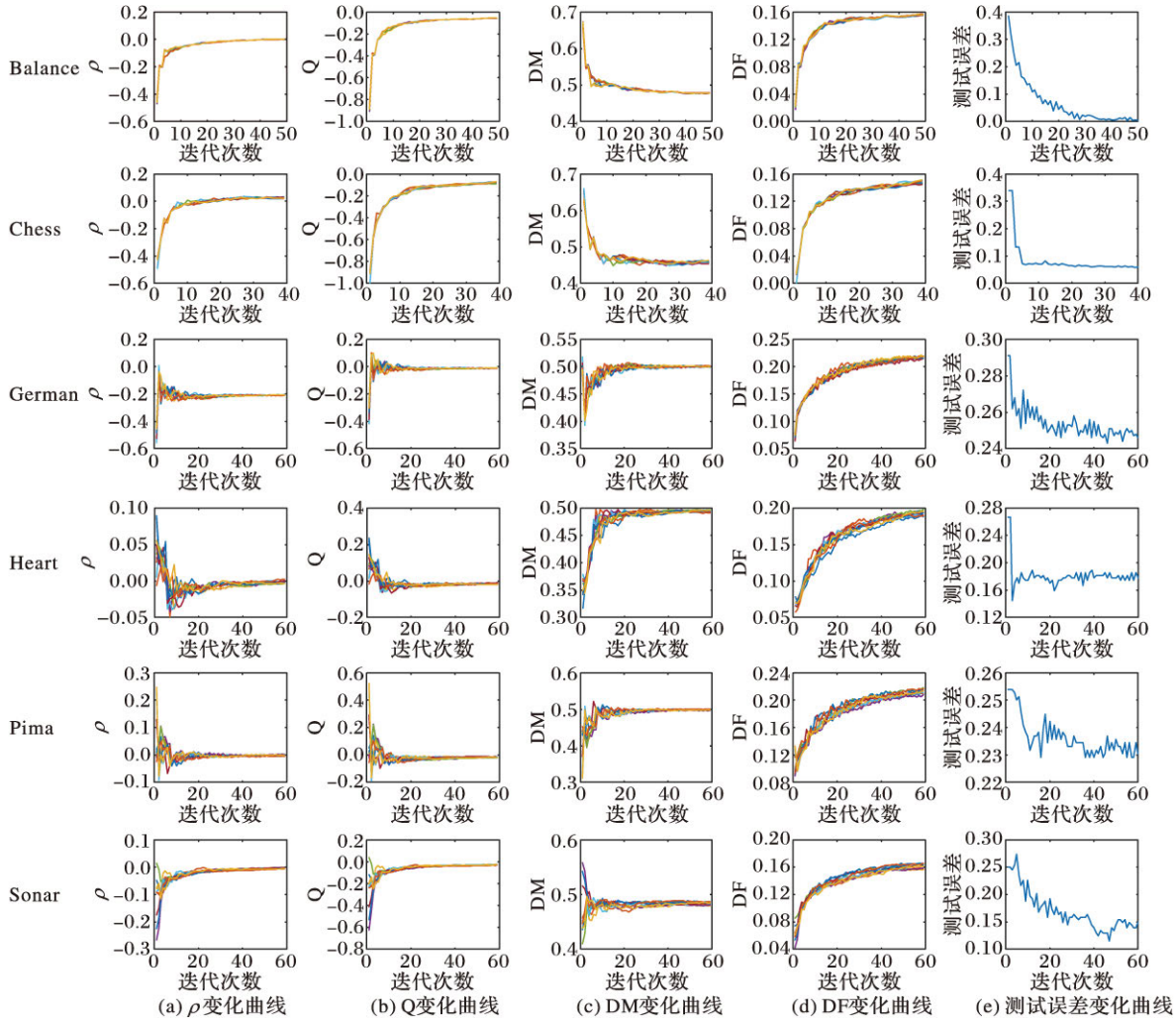

图 1　6 个数据集上的实验结果

Fig. 1　Experimental results on six data sets

然后,单独看图 1 的子图(d) ,这是 DF 的变化曲线,每条曲线的变化都是相同的模式,先单调递增后不变,而 ρ、Q、DM 在不同的数据集上变化有所区别。根据 DF 的计算公式,可以知道,它统计的是共同错分的样本占总数的比例,而 AdaBoost 算法特点是关注难分的样本,随着迭代的进行,可以看到 DF 的值基本保持不变,说明 AdaBoost 算法的关注点确

实集中到了这些共同错分的样本上。DF 也能对组合分类器的精确度有所反映,它最后趋近的值越大,组合分类器的精度就相对越差。

通过观察图 1,已经对多样性与分类精度之间的关系有了初步的直观认识,为了进行更客观地比较,采用定量分析的方法,利用皮尔逊相关系数公式如(18) 所示,计算多样性度量与测试误差的相关性,结果见表3。式(18) 中,x、y 表示两个变量,$E(x)$ 表示 x 的数学期望。

$$corr(x,y) = \frac{E(xy) - E(x)E(y)}{\sqrt{E(x^2) - E^2(x)}\sqrt{E(y^2) - E^2(y)}} \quad (18)$$

表 3 多样性度量与测试误差之间的皮尔逊相关系数值

Tab. 3 Pearson correlation coefficient between diversity measurement and test error

数据集序号	DF	ρ	Q	DM
1	-0.913	-0.859	-0.845	0.824
2	-0.868	-0.862	-0.869	0.866
3	-0.283	-0.128	-0.188	-0.112
4	-0.053	0.114	0.089	-0.092
5	-0.333	0.199	0.209	-0.315
6	-0.443	-0.377	-0.378	-0.299

分析表 3 可知,在 Balance、Chess、German、Sonar、Pima 数据集上,DF 与测试误差之间的相关性均高于其他三种多样性度量;在 Heart 上四种多样性度量与测试误差之间都是极弱相关。总的来说,DF 与测试误差之间的相关性最高。

4.2 实验二的结果及分析

实验二中,使用 WLDF 作为 AdaBoost 的弱学习算法记为 WLDF-Ada。为验证 WLDF-Ada 的有效性,实验采用 10 折交叉验证法,比较 WLDF-Ada 与 AdaBoost、Bagging、随机森林(Random Forest, RF) 以及文献[16]提出的 R-Ada 方法的 10 次平均测试误差,基分类器数目均为 50。其中,Ada.、Bag、RF 来自 python 机器学习工具箱 sicikt-learn(http://scikit-learn.org/stable/index.html) 。

表 4 四种算法测试误差对比

Tab. 4 Comparison of test errors of four algorithms

数据集序号	测试误差/%				
	Bagging	RF	AdaBoost	R-Ada	WLDF-Ada
1	7.86	5.97	0.35	0.15	0.35
2	0.35	0.97	5.41	5.34	5.17
3	24.30	23.69	25.80	23.30	23.25
4	16.66	17.04	21.11	19.44	16.30
5	24.09	23.72	24.02	23.66	22.58
6	19.21	17.56	18.86	16.93	16.64

分析表 4 可知: 在 Balance 数据集上,R_Ada 取得最小测试误差,WLDF_Ada 与 AdaBoost 次之,三者表现优于 Bagging、RF; 在 Chess 数据集上 Bagging 和 RF 优于其他三种 AdaBoost 算法。分析发现这是因为 Chess 数据属性之间存在强烈的相互影响,需要增加决策树的深度来改善分类性能,而本文实验中 AdaBoost 算法是以单层决策树作为弱分类器,Bagging 和 RF 则对基分类器决策树的深度没有限制。在 German 数据集上,WLDF_Ada 的测试误差比 Bagging、RF、AdaBoost、R_Ada 分别低 1.05%、0.44%、2.55%、0.05%。类似地在 Heart、Pima 以及 Sonar 数据集上 WLDF_Ada 的测试误差比 Bagging、RF、AdaBoost、R_Ada 分别低了 0.3%、0.74%、4.81%、3.14%; 1.51%、1.14%、1.44%、1.08% 以及 2.57%、0.92%、2.22%、0.29%。除了在 Chess 和 Balance 数据集上,WLDF_Ada 算法的表现均优于其他四种算法。单独比较 WLDF_Ada 与 AdaBoost,除了在 Balance 数据集上,WLDF_Ada 均比 AdaBoost 有不同程度上的性能提升。

5 结语

多样性是影响集成学习的重要因素,合适的多样性度量可以指导基分类器的选择以及组合。本文研究了 4 种成对型多样性度量与 AdaBoost 算法表现之间的关系,实验一的结果表明随着迭代的进行,4 种多样性度量值都趋于一个稳定的值,其中 DF 的变化模式固定。另外针对 AdaBoost 的过适应问题,本文改进了传统 AdaBoost 弱分类器的选择策略,提出了弱分类器学习算法 WLDF,实验二结果表明 WLDF 算法可以抑制对困难样本的过分关注,增加分类器间的多样性,改善 AdaBoost 的分类性能。DF 与 AdaBoost 算法的分类精度在一些数据集上关联并不紧密,下一步可以尝试根据样本权值以及弱分类器的权重,设计一个更合适 AdaBoost 算法的多样性度量方法。

参考文献 (References)

[1] SCHAPIRE R E. The strength of weak learnability [J]. Machine Learning, 1990, 5(2): 197 - 227.

[2] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123 - 140.

[3] SCHAPIRE R E, SINGER Y. Improved boosting algorithms using confidence-rated predictions [J]. Machine Learning, 1999, 37(3): 297 - 336.

[4] MORENO P J, LOGAN B, RAJ B. A boosting approach for confidence scoring [EB/OL]. [2017-03-06]. http://www.mirrorservice.org/sites/www.bitsavers.org/pdf/dec/tech-reports/CRL-2001-8.pdf.

[5] 廖广军,李致富,刘屿,等. 基于深度信息的弱光条件下人脸检测[J]. 控制与决策,2014,29(10):1866 - 1870. (LIAO G J, LI Z F, LIU Y, et al. Human face detection under weak light based on depth information [J]. Control and Decision, 2014, 29(10): 1866 - 1870.)

[6] PIAO Y, PIAO M, RYU K H. Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles [J]. Computers in Biology & Medicine, 2017, 80: 39 - 44.

[7] KIM B, YU S C. Imaging sonar based real-time underwater object detection utilizing AdaBoost method [C]// UT 2017: Proceedings of the 2017 IEEE Underwater Technology. Piscataway, NJ: IEEE, 2017: 1 - 5.

[8] 李文辉,倪洪印. 一种改进的 AdaBoost 训练算法[J]. 吉林大学学报(理学版),2011,49(3):498 - 504. (LI W H, NI H Y. An improved AdaBoost training algorithm [J]. Journal of Jilin University (Science Edition), 2011, 49(3): 498 - 504.)

[9] KROGH B A, VEDELSBY J. Neural network ensembles, cross validation, and active learning [J]. Advances in Neural Information Processing Systems, 1994, 7(10): 231 - 238.

[10] KUNCHEVA L I. That elusive diversity in classifier ensembles [C]// Proceedings of the 1st Iberian Conference on Pattern Recognition and Image Analysis, LNCS 2652. Berlin: Springer, 2003: 1126 - 1138.

[11] 孙博,王建东,陈海燕,等. 集成学习中的多样性度量[J]. 控制与决策,2014,29(3):385 - 395. (SUN B, WANG J D, CHEN H Y, et al. Diversity measures in ensemble learning [J]. Control and Decision, 2014, 29(3): 385 - 395.)

(下转第 660 页)

and Forecasting, 2013, 28(3): 570 - 585.
[9] ZHENG J F, ZHANG J, ZHU K Y, et al. Gust front statistical characteristics and automatic identification algorithm for CINRAD [J]. Acta Meteorologica Sinica, 2014, 28(4): 607 - 623.
[10] TAGLIAFERRI F, VIOLA I M, FLAY R G J. Wind direction forecasting with artificial neural networks and support vector machines [J]. Ocean Engineering, 2015, 97(15): 65 - 73.
[11] 赵丽艳. 基于多普勒天气雷达的风切变预警算法研究[D]. 天津: 中国民航大学, 2016: 37 - 38. (ZHAO L Y. Wind shear forecasting algorithm based on Doppler weather radar [D]. Tianjin: Civil Aviation University of China, 2016: 37 - 38.)
[12] HWANG Y, YU T Y, LAKSHMANAN V, et al. Neuro-fuzzy gust front detection algorithm with S-band polarimetric radar [J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(3): 1618 - 1628.
[13] DENG H Y, ZHU Q X, SONG X L. Decision-based marginal total variation diffusion for impulsive noise removal in color images [J]. Journal of Sensors, 2017: Article ID 7635641.
[14] RUDIN L I, OSHER S. Total variation based image restoration with free local constraints [C]// Proceedings of the 1994 IEEE International Conference Image Processing. Piscataway, NJ: IEEE, 1994, 1: 31 - 35.
[15] TABARY P, GUIBERT F, PERIER L, et al. An operational triple-PRT Doppler scheme for the French radar network [J]. Journal of Atmospheric and Oceanic Technology, 2006, 23(12): 1645 - 1656.
[16] LIM E, SUN J. A velocity dealiasing technique using rapidly updated analysis from a four-dimensional variational Doppler radar data assimilation system [J]. Journal of Atmospheric and Oceanic Technology, 2010, 27(7): 1140 - 1152.
[17] JAIN A K. Data clustering: 50 years beyond K-means [J]. Pattern Recognition Letters, 2010, 31(8): 651 - 666.
[18] 周双, 冯勇, 吴文渊, 等. 一种基于模糊 C 均值聚类小数量计算最大 Lyapunov 指数的新方法[J]. 物理学报, 2016, 65(2): 020502. (ZHOU S, FENG Y, WU W Y, et al. A novel method based on the fuzzy C-means clustering to calculate the maximal Lyapunov exponent from small data [J]. Acta Physica Sinica, 2016, 65(2): 42 - 48.)
[19] 黄仪方, 朱志愚. 航空气象[M]. 成都: 西南交通大学出版社, 2002: 131 - 134. (HUANG Y F, ZHU Z Y. Aviation Weather [M]. Chengdu: Southwest Jiaotong University Press, 2002: 131 - 134.)

This work is partially supported by the National Natural Science Foundation of China (U1533113, U1433202), the Fundamental Research Funds for the Central Universities (3122016B001).

XIONG Xinglong, born in 1962, M. S., professor. His research interests include signal and information processing, LIDAR (Laser Intensity Direction And Ranging) weather detection.

YANG Lixiang, born in 1990, M. S. candidate. Her research interests include weather radar weather detection, image processing.

MA Yuzhao, born in 1978, Ph. D., associate professor. Her research interests include fiber optics, aviation weather detection, electromagnetic computation.

ZHUANG Zibo, born in 1980, M. S., lecturer. His research interests include aviation weather detection, data processing.

(上接第 654 页)

[12] CAVALCANTI G D C, OLIVEIRA L S, MOURA T J M, et al. Combining diversity measures for ensemble pruning [J]. Pattern Recognition Letters, 2016, 74(C): 38 - 45.
[13] 杨春, 殷绪成, 郝红卫, 等. 基于差异性的分类器集成: 有效性分析及优化集成[J]. 自动化学报, 2014, 40(4): 660 - 674. (YANG C, YIN X C, HAO H W, et al. Classifier ensemble with diversity: effectiveness analysis and ensemble optimization [J]. Acta Automatica Sinica, 2014, 40(4): 660 - 674.)
[14] PARVIN H, MIRNABIBABOLI M, ALINEJAD-ROKNY H. Proposing a classifier ensemble framework based on classifier selection and decision tree [J]. Engineering Applications of Artificial Intelligence, 2015, 37: 34 - 42.
[15] LI N, YU Y, ZHOU Z H. Diversity regularized ensemble pruning [C]// Proceedings of the 2012 Joint European Conference on Machine Learning and Knowledge Discovery in Databases, LNCS 7523. Berlin: Springer, 2012: 330 - 345.
[16] 姚旭, 王晓丹, 张玉玺, 等. 基于随机子空间和 AdaBoost 的自适应集成方法[J]. 电子学报, 2013, 41(4): 810 - 814. (YAO X, WANG X D, ZHANG Y X, et al. A self-adaption ensemble algorithm based on random subspace and AdaBoost [J]. Acta Electronica Sinica, 2013, 41(4): 810 - 814.)
[17] 曹莹, 苗启广, 刘家辰, 等. AdaBoost 算法研究进展与展望[J]. 自动化学报, 2013, 39(6): 745 - 758. (CAO Y, MIAO Q G, LIU J C, et al. Advance and prospects of AdaBoost algorithm [J]. Acta Automatica Sinica, 2013, 39(6): 745 - 758.)
[18] MEDDOURI N, KHOUFI H, MADDOURI M S. Diversity analysis on boosting nominal concepts [C]// Proceedings of the 2012 Pacific-Asia Conference on Knowledge Discovery and Data Mining, LNCS 7301. Berlin: Springer, 2012: 306 - 317.
[19] YULE G U. On the association of attributes in statistics: with illustrations from the material of the childhood society, &c [J]. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 1900, 194 (252/253/254/255/256/257/258/259/260/261): 257 - 319.
[20] KUNCHEVA L I, WHITAKER C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy [J]. Machine Learning, 2003, 51(2): 181 - 207.
[21] SKALAK D B. The sources of increased accuracy for two proposed boosting algorithms [C]// AAAI '96: Proceedings of the Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms. Menlo Park, CA: AAAI Press, 1996: 120 - 125.
[22] GIACINTO G, ROLI F. Design of effective neural network ensembles for image classification purposes [J]. Image and Vision Computing, 2001, 19(9/10): 699 - 707.

This work is partially supported by the National Natural Science Foundation of Jiangsu Province (BK20140165).

WANG Lingdi, born in 1991, M. S. candidate. Her research interests include machine learning, data mining.

XU Hua, born in 1978, Ph. D., associate professor. Her research interests include computer intelligence, workshop scheduling, large data.