一种新的改进 AdaBoost 弱分类器训练算法

谢红跃 方昱春 蔡起运
(上海大学计算机工程与科学学院, 上海 200072)

摘 要 AdaBoost 是机器学习中比较流行的分类算法。通过研究弱分类器的特性, 提出了两种新的弱分类器的阈值和偏置计算方法, 二者可以使弱分类器识别率大于 50%, 从而保证在弱分类器达到一定数目的情况下, AdaBoost 训练收敛。对两种阈值和偏置计算方法的仿真实验结果表明, 在错分率降可接受的范围内, 二者均使用较少的弱分类器便可获得高识别率的强分类器。

关键词 弱分类器 AdaBoost 算法 强分类器 错分率
中图法分类号: TP301.6 **文献标识码**: A **文章编号**: 1006-8961(2009)11-2411-05

A New Weak Classifier Training Method for AdaBoost Algorithm

XIE Hong-yue, FANG Yu-chun, CAI Qi-yun
(*Department of Computer Engineering and Science, Shanghai University, Shanghai* 200072)

Abstract AdaBoost is a very popular classification algorithm on machine leaning. By studying the characteristics of the weak classifier, this paper proposes two new methods to calculate the threshold and bias of the weak classifier. The two methods make the correct rate of weak classifier larger than 50%, assure the convergence of AdaBoost training when the weak classifier reach a certain number. Simulation experiments show when the error rate is in an acceptable range, the algorithms using fewer weak classifiers will be able to guarantee the strong classifier to maintain a high correct rate.

Keywords weak classifier, AdaBoost algorithm, stronger classifier, error rate

1 引 言

AdaBoost 算法是机器学习中一种非常重要的特征分类算法, 被广泛应用于图像检索[1]、人脸表情识别[2]等应用中。Leshem 将 AdaBoost 算法应用到交通管理信息系统中[3], 利用弱学习器来训练道路交通数据, 并且预测道路交通流量情况, 取得良好的效果。Lin 将 Real AdaBoost 算法应用到基于内容的图像检索系统中[4], 通过对图像的类别短语进行训练, 达到降低噪声的效果, 实验显示较 K-NN(K-nearest neighbor)分类算法准确性有所提高。Dai 等人将 AdaBoost 算法应用到区域图像检索中[1], 通过使用 AdaBoost 弱分类器对特征反复训练, 得到具有较小错分率的强分类器, 从而返回更加精确的查询结果。为了解决不同的特征融合分类问题, Yin 等人提出了一种改进的 boosting 算法[5], 使用一个弱分类器仅对某一个特征集进行训练, 最终根据权重将这些弱分类器组合成一个强分类器, 该方法在手写数字识别中取得了较好的效果。Viola 等人提出了的样本权重更新方法[6], 被正确分类样本权重减小, 而错误分类样本权重不变。AdaBoost 在学习训练的过程中要解决的是每一轮样本训练集的样本分布问题, 其中正负样本的权重更新及错分率的处理

基金项目: 上海市自然科学基金项目(08ZR1408200); 上海市重点学科建设项目(J50103); 中国科学院模式识别国家重点实验室开放课题基金(08-2-16)
收稿日期: 2009-06-20; **改回日期**: 2009-09-01
第一作者简介: 谢红跃(1982 ~), 男。上海大学计算机工程与科学学院硕士研究生。主要研究方向为图像处理、模式识别。E-mail: hongyue. xie@ gmail. com

至关重要。对样本进行两类划分,以便保证弱分类器分得的样本准确率大于随机猜测的准确率。李闯等人提出了针对目标检测问题的改进 AdaBoost 算法[7],采用了新的参数求解方法,弱分类器的加权参数不但与错分率有关,还与其对正样本的识别能力有关。Kim 等人提出了特征值是基于高斯概率分布的 AdaBoost 算法[8],通过特征值的分布与高斯概率分布的均值距离来判别正负样本。

AdaBoost 在学习训练的过程中弱分类器使用某种策略对样本进行两类划分,以便保证弱分类器分得的样本准确率大于 50%,据此本文提出了新的 AdaBoost 样本阈值和偏置计算方法,这二种方法依据样本权重的大小计算出对应的样本阈值,并来区分正确分类和错误分类的样本,从而使得弱分类器划分准确性大于 50%。

2 AdaBoost 分类训练算法

已知有 N 个训练样本 $\{x_1,y_1\},\cdots,\{x_i,y_i\},\cdots,\{x_N,y_N\}$ 其中 x_i 为样本特征,y_i 为类别标识;$y_i\in\{-1,+1\}$;-1 表示反例,$+1$ 表示正例。训练样本有 K 维特征,表示为 v_k,其中 $1\leqslant k\leqslant K$;对第 i 个样本 x_i,特征:$v_1(x_i),v_2(x_i),\cdots,v_k(x_i)$,输入特征 $v_k(x)$ 对应一个简单的二值判断公式

$$h_{t,k}(x)=\begin{cases}1 & p_kv_k(x)<p_k\theta_k,\\ 0 & \text{其他}\end{cases}$$

其中,$t\in\{t\mid 1\leqslant t\leqslant T\}$。

弱分类器中由阈值 θ_k 和偏置 p_k 决定样本的属于正例还是反例,偏置 p_k 决定不等式方向,$p_k\in\{-1,+1\}$。训练使用 T 个弱分类器 $h_t(x_i)$,并且弱分类器满足条件

$$\min_{h\in\boldsymbol{H}}\frac{1}{N}\sum_{i=1}^{N}w_t(x_i)\varphi(y_i,h_t(x_i)),$$

$$\varphi(y_i,h(x_i))=\begin{cases}1 & y_i=h(x_i)\\ -1 & y_i\neq h(x_i)\end{cases}$$

训练算法步骤如下:

(1)设 $\boldsymbol{A}$ 代表正例,$\overline{\boldsymbol{A}}$ 代表反例,初始权值为

$$w_0(x_i)=\begin{cases}\dfrac{1}{2n} & x_i\in\boldsymbol{A}\\ \dfrac{1}{2m} & x_i\in\overline{\boldsymbol{A}}\end{cases},\ \forall x_i\in\boldsymbol{\Omega}=\boldsymbol{A}+\overline{\boldsymbol{A}}=\{x_1,\cdots,x_N\}$$

其中,$t=1,n+m=N$。

(2)对于弱分类器 h,每个特征 k,确定阈值 θ_k 和偏置 p_k,计算 $h_{t,k}(x)=\begin{cases}1 & \{p_kv_k(x)<p_k\theta_k\}\\ 0 & \text{其他}\end{cases}$。

(3)遍历每个特征 k,使得以权重计算的错误率 $\varepsilon_{t,k}(h)=\sum_i^N w_i(x_i)\varphi(y_i,h_{t,k}(x_i))$ 最小;找出具有最小的错误 $\varepsilon_{t,k}$ 的简单分类器 $h_{t,k}$,$\varepsilon_t=\min(\varepsilon_{t,k})$。

(4)对样本的权重进行更新

$$w_{t+1}(x_i)=\frac{1}{Z_t(a_t)}w_t(x_i)\exp(-y_ia_th_t(x_i))$$

其中 $a_t=\frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}$,$Z_t(a_t)=\sum_{i=1}^{N}w_t(x_i)$。

若 $t=T$,则算法循环终止,否则 $t=t+1$,重复步骤 2 ~ 步骤 4。

(5)最后的强分类器是

$$h(x)=\sum_{t=1}^{T}a_th_t(x)$$

3 弱分类器的阈值和偏置选取

3.1 弱分类器

弱分类器对于区分候选正样本和负样本至关重要,为了保证最终的强分类器能够收敛,弱分类器对正负样本分类的准确率必须大于 0.5,这样训练算法最终才会收敛。

弱分类器的过程描述为

(1)计算弱分类器的阈值 θ 和偏置 p。

(2)计算样本累加

$$\varepsilon_{t,k}(h)=\sum_i^N w_i(x_i)\varphi(y_i,h(x_i))$$

(3)计算弱分类器 t 最小

$$\varepsilon_t=\min(\varepsilon_{t,k}),1\leqslant k\leqslant K$$

若弱分类器 $\varepsilon_t(h_t)<\frac{1}{2}$,那么 $a_t=\frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}>0$,从而由 $\varepsilon(h_A)<\varepsilon(h_B)$ 推出 $a(h_A)>a(h_B)$。表明弱分类器错分率越小,在强分类器中的权重越大。本文所提出的阈值和偏置选取方法正是基于此选取的。

3.2 阈值和偏置

弱分类器划分样本的核心是选取相应的阈值和偏置,阈值和偏置的选取是保证弱分类器对某一样本特征分类准确性大于 0.5 的关键。如果要求分类器在尽可能少的训练次数内达到对正样本较高的分类识别率,同时又要权衡负样本的错分率,不至于让负样本错分率不可接受。这就需要同时考虑正负样

本的权重。方法一、二同时兼顾正负样本的错分率,保证了最终的强分类器错分率能够趋于0。

令
$$w_p(x_i) = y_i w(x_i)$$
$$w_n(x_i) = (1-y_i) w(x_i)$$

分别为正、负样本的权重分布,根据弱分类器中样本权重和样本的关系,提出弱分类器中选取阈值和偏置的两种方法如下:

方法 1　正、负样本权重阈值分别为

$$\begin{aligned}\zeta_p(w(x)) &= \max(w_p(x)) - w_p(x) + w_n(x)\\ \zeta_n(w(x)) &= \max(w_n(x)) - w_n(x) + w_p(x)\end{aligned} \tag{1}$$

方法 2　正、负样本权重阈值分别为

$$\begin{aligned}\zeta_p(w(x)) &= \max(w_n(x)) - w_p(x) + w_n(x)\\ \zeta_n(w(x)) &= \max(w_p(x)) - w_n(x) + w_p(x)\end{aligned} \tag{2}$$

两者的特点在于根据正负样本权重阈值计算样本阈值,再对样本进行划分。设正、负样本阈值为

$$\theta_p = f(\min(\zeta_p(w(x))))$$
$$\theta_n = f(\min(\zeta_n(w(x))))$$

其中函数 f 是样本权重关于所对应的样本的函数。样本阈值计算公式则为

$$\theta = \begin{cases}\theta_p & \theta_p \leqslant \theta_n\\ \theta_n & \theta_p > \theta_n\end{cases} \tag{3}$$

偏置

$$p = 2((\zeta_p(w(x)) > \zeta_n(w(x))\ ?\ 1:0) - 0.5) \tag{4}$$

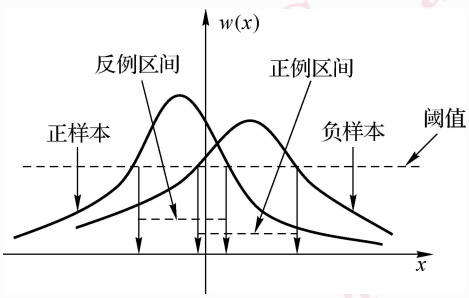


图 1　正负样本权重分布与阈值示意图

Fig. 1　Diagram of threshold and weight distribution of positive & negative

从图 1 可以看出,如果正样本大于阈值,那么标记该样本为 0,正样本小于阈值,那么标记样本为 1,如果负样本大于阈值,那么标记该样本为 1,负样本小于阈值标记为 0。偏置取值为 -1 和 +1,表示是将当前正样本或负样本标识为正例或反例,这取决于正负样本阈值的大小,算法能对上次错分的样本再次划分,这保证这种划分的准确率要大于 0.5。

3.3　收敛性分析

对于第 t 个弱分类器 h_t,Z 是权重归一化的因子,可得

$$\begin{aligned}Z_t(a_t) &= \sum_{x_i \in \Omega} w_t(x_i)\exp(-y_i a_t h_t(x_i))\\ &= \sum_{x_i \in A} w_t(x_i) e^{-a_t} + \sum_{x_i \in \overline{A}} w_t(x_i) e^{a_t}\end{aligned} \tag{5}$$

对等式两边求导数得到

$$\frac{dZ_t(a_t)}{da_t} = \sum_{x_i \in A} -w_t(x_i) e^{-a_t} + \sum_{x_i \in \overline{A}} w_t(x_i) e^{a_t} = 0$$

即 $\sum_{x_i \in A} w_t(x_i) = \sum_{x_i \in \overline{A}} w_t(x_i) e^{2a_t}$,又 $\varepsilon_t(h) = \sum_{x_i \in \overline{A}} w_t(x_i)$,可以推导出

$$\begin{aligned}Z_t(a_t) &= \sum_{x_i \in A} w_t(x_i) e^{-a_t} + \sum_{x_i \in \overline{A}} w_t(x_i) e^{a_t}\\ &= 2\sqrt{\varepsilon_t(h_t)(1-\varepsilon_t(h_t))}\end{aligned} \tag{6}$$

弱分类器满足假设,$\gamma_t = \frac{1}{2} - \varepsilon_t(h_t)$,$\gamma_t \in \left(0, \frac{1}{2}\right]$结合式(2)可以得出

$$Z_t(a_t) = \sqrt{(1-4\gamma_t^2)} = \exp\{-2\gamma_t^2\} \tag{7}$$

因为 ε 的上界为 Z,可以得到 $\varepsilon \leqslant Z \leqslant \exp\{-2\gamma_t^2\}$,这表明弱分类器错分率小于 0.5,那么强分类器的错分率将呈指数线下降,算法最终是收敛的。

4　实验结果及分析

训练样本是由随机函数生成,数据集包含 400 个样本,8 维特征,如图 2 所示。数据集中样本的类别标签符合二值分类的类别即属于(-1,+1),弱分类器效果如图 3 所示,训练样本使用改进弱分类

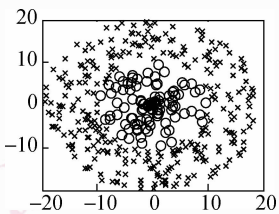


图 2　样本集合

Fig. 2　Examples set

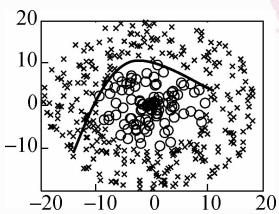
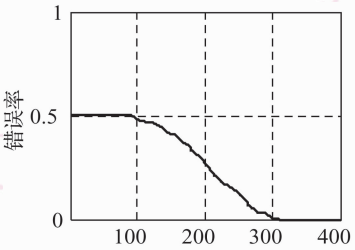


图 3　弱分类器划分的示意图

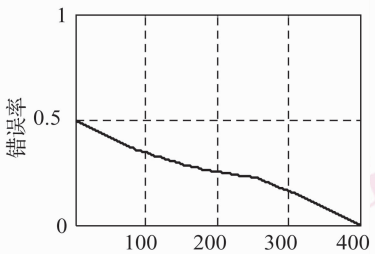
Fig. 3　Weak classifier's division

器阈值和偏置的 AdaBoost 算法进行分类实验。

实验得到的正负样本累积错分率曲线如图 4 所示,可见样本数目相同的情况下正负样本的累积错分率曲线并不相同,反映了正负样本的分布是不同的,因此权衡正负样本错分率尤为重要。



(a) 正样本



(b) 负样本

图 4 正负样本累积错误率曲线

Fig. 4 The cumulative error rate curve of positive & negative

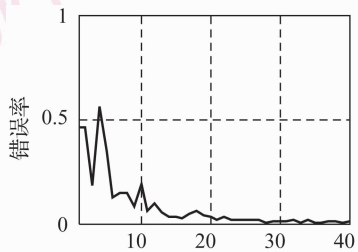
表 1 是不同弱分类器数目下两种方法的错分率,显示出弱分类器个数的增加,错分率不一定就降低。分类器个数相差不大的情况下错分率有反复。

表 1 不同弱分类器数目下两种方法错分率比较

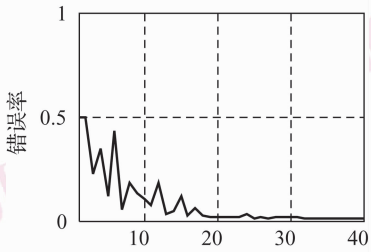
Tab. 1 Error rate comparison of the two methods with different weak classifier number

弱分类器个数	方法 1	方法 2
5	0.347 5	0.115 0
10	0.190 0	0.105 0
15	0.035 0	0.115 0
20	0.030 0	0.017 5
25	0.015 0	0.010 0
30	0.012 5	0.015 0
35	0.005 0	0.010 0
40	0.012 5	0.012 5

从图 5 可以更加直观地看出在分类器个数相差不大的情况下错分率可能有震荡,但是随着弱分类器个数大幅增加,错分率是呈递减趋势的,表明弱分类器具有学习功能,算法能够最终收敛。



(a) 方法1



(b) 方法2

图 5 不同弱分类器数目下强分类器错分率曲线

Fig. 5 Error rate curve of strong classifier with different weak classifier number

方法 1 和方法 2 都采用均衡正负样本的错分率而选择界于正样本最大权重和负样本最大权重之间作为阈值条件,因为样本连续被错分,该样本权重越大,阈值条件就是将样本那些错分的样本与那些未被错分的样本进行区分,实验表明这种选取方法最终是收敛的。

阈值变化是因为权重的变化引起来的,未正确分类的样本权重提高了,样本权重阈值对应也发生了变化,这种变化带来的效果是对前一次迭代过程中未进行正确分类的样本在下一次迭代中能够有可能正确分类。

由于算法在提升过程中,不断地对训练集中那些样本因权值更新后权重小于阈值的样本再次划分,关注正负样本中难以划分的样本,并且兼顾正负样本错分率,获得了较快的收敛速度。

5 结 论

本文提出的 AdaBoost 中弱分类器划分样本阈值和偏置新的计算方法,是一种普适的算法,可以应用于图像特征的训练和分类上。具有广阔的应用前

景。该方法使用较少的弱分类器能保证强分类器在保持较高的识别率的同时算法能够有比较快的收敛速度。

参考文献(References)

1 Dai Sheng-Yang, Zhang Yu-Jin. AdaBoost in region-based image retrieval[A]. In:Processings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)[C], Montreal, Canada,2004,**3**: 429-432.

2 Shan S, Yang P, Chen X. AdaBoost gabor fisher classifier for face recognition[A]. In: Proceeding of the International Workshop on Analysis and Modeling of Faces and Gestures, AMFG 2005[C], Beijing, China, 2005,**3723**:278-291.

3 Leshem G, Ritov Y. Traffic flow prediction using AdaBoost algorithm with random forests as a weak learner[A]. In: Proceedings of the International Conference on Computer, Information, and Systems Science, and Engineering [C], Bangkok, Thailand, 2007: 193-198.

4 Lin Wei-Chao, Oakes M, Tait J. Real AdaBoost for large vocabulary image classification [A]. In: Proceedings of the International Workshop on Content-based Multimedia Indexing, 2008 [C], London, Unit Kingdom,2008:192-199.

5 Yin X C, Liu C P, Han Z. Feature combination using boosting[J]. Pattern Recognition Letters, 2005,**26**(14): 2195-2205.

6 Viola P, Jones M. Fast and robust classification using asymmetric AdaBoost and a detector cascade [A]. In Advances in Neural Information Processing Systems 14[C], Cambridge, MA, USA: MIT Press,2002: 1311-1318.

7 Li Chuang, Ding Xiao-qing, Wu You-shou. A revised AdaBoost algorithm—AD AdaBoost[J]. Chinese Journal of Computers, 2007, **30**(1):103-109.[李阅,丁晓青,吴佑寿. 一种改进的 AdaBoost 算法——AD AdaBoost[J]. 计算机学报. 2007, **30**(1): 103-109.]

8 Kim Jeong-hyun, Park Jong-hyun, Kang Dong-joong. Method to improve the performance of the AdaBoost algorithm using gaussian probability distribution. Control [A]. In: Processings of the International Conference on Control, Automation and Systems[C], Seoul, Korea (South), 2008:1749-1752.