

一种改进 AdaBoost 算法的方法

方义秋¹, 卢道兵², 葛君伟³

(1 重庆邮电大学 计算机科学与技术学院, 重庆 400065;

2 重庆邮电大学 GIS 研究所, 重庆 400065;

3 重庆邮电大学 软件学院, 重庆 400065)

摘 要: 针对传统 AdaBoost 算法的不足, 分析了分类器训练时间长和训练过程中容易出现训练结果对训练样本严重收敛的问题(过训练), 并提出了解决这一问题的有效方法. 新方法主要将特征值和排序结果进行缓存以及对训练样本及时更新. 使用该方法训练级联车牌检测器, 实验结果表明, 新方法较好地解决了传统 AdaBoost 算法中所出现的过训练和训练耗时的问题, 在提高检测率的同时降低了误检率, 并且训练时间缩短了 50% 左右.

关键词: AdaBoost 算法; 样本更新; 特征值; 过训练; 车牌检测

中图分类号: TP391.41

文献标识码: A

文章编号: 1000-7180(2010)07-0040-04

One Improved Method for AdaBoost Algorithm

FANG Yi-qiu¹, LU Dao-bing², GE Jun-wei³

(1 College of Computer Science and Technology, Chongqing University of Posts & Telecommunications, Chongqing 400065, China; 2 GIS Research Centre, Chongqing University of Posts & Telecommunications, Chongqing 400065, China;

3 College of Software, Chongqing University of Posts & Telecommunications, Chongqing 400065, China)

Abstract: Focusing on the disadvantages of classical AdaBoost algorithm, this paper mainly analyzes the issue that the training time for classifiers is time-consuming and in training process the training results severely constringe the training samples(excessive training)and a new method is advanced to avoid the problems. The new method is to buffer the computational results of sorted feature values and update the training samples in time. As a result, using the method to train a cascade license plate, the experimental results show that the new method does not lead to the issue of excessive training and time consuming like classical AdaBoost often does, and moreover, the training time is shortened to 50 percent with a high detection rate and a low false alarm rate.

Key words: AdaBoost; sample update; feature value; excessive training; license plate detection

1 引言

车牌检测技术在智能交通管理系统中发挥着重要作用,它是车牌识别系统的一个关键技术环节,并已成为现代交通工程领域中研究的热点问题之一^[1].近年来出现了大量车牌检测方法,在众多的检测方法中,AdaBoost 算法^[2]能够从根本上解决实时运动目标检测的速度问题,同时有较好的识别效果.该算法是一种自适应的 Boosting 算法^[3],利用该

算法可以将一组弱学习算法提升为一个强学习算法,最终得到一个识别准确率理想的分类器.

尽管使用 AdaBoost 算法构建分类器的检测速度非常卓越,但它的训练过程却十分漫长. P. Viola^[4]等曾经花费数周时间在高级工作站上完成他们的训练;其后, R. Xiao^[5]等提出了 Boosting Chain 训练算法,减少了强分类节点需要的弱分类器个数,在一定程度上减少了训练时间; Jianxin Wu^[6]等提出训练中直接选择特征来加快训练速度.然而,这些算

法并未显著地减少分类器的训练时间.为此,本文提出了通过缓存特征值和排序结果来大幅提高训练速度,从而大大缩短了分类器的训练时间.

2 传统的 AdaBoost 算法

AdaBoost 算法是一种分类器算法^[7],其基本思想是利用大量分类能力一般的简单分类器通过一定方法叠加起来,构成一个分类能力很强的强分类器^[8].理论证明,只要每个简单分类器分类能力比随机猜测要好,当简单分类器个数趋向于无穷时,强分类器的错误率将趋于零.

AdaBoost 算法用于车牌检测时,从很大的车牌特征集中选择很小的一部分关键特征从而产生一个极其有效的分类器^[9].这些简单特征都有一定的车牌和非车牌区分性.最终系统使用数千个一维简单分类器,组合起来达到很好的分类效果.算法描述如下:

已知有 n 个训练样本 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 构成训练样本集,其中, $y_i = \{0, 1\}$ ($i = 1, 2, \dots, n$) 对应样本的假和真.在训练样本中共有 m 个假样本, l 个真样本,待分类物体有 r 个简单特征,表示为 $f_j(*)$,其中 $1 \leq j \leq r$.对于第 i 个样本 x_i ,它的特征为 $\{f_1(x_i), f_2(x_i), \dots, f_j(x_i), \dots, f_r(x_i)\}$,对每一个输入特征 f_i 有一个简单二值分类器.第 j 个特征的弱分类器由一个阈值 θ_j 、一个特征 f_j 和一个指示不等式方向的偏置 P_j (只有 ± 1 两种情况) 构成.当 $P_j f_j < P_j \theta_j$ 时, $h_j(x) = 1$; 否则, $h_j(x) = 0$.

训练目标是通过对比正负样本进行分析,选择分类错误率最低的 t 个弱分类器,最终优化组合成一个强分类器.

训练方法如下:

(1) 设 $w_{t,i}$ 为第 t 次循环中第 i 个样本的误差权重,对训练样本的误差权重按如下公式初始化:对于 $y_i = 0$ 的样本, $w_{1,i} = 1/2m$; 对于 $y_i = 1$ 的样本, $w_{1,i} = 1/2l$;

(2) for $t = 1, 2, \dots, T$ (迭代运算, T 为训练的次数)

① 权重归一化,使得 $w_{t,i}$ 为 $w_{t,i} = w_{t,i} / \sum_{i=1}^n w_{t,i}$;

② 对于每个特征 j ,训练出其弱分类器 h_j ,也就确定了阈值 θ_j 和偏置 P_j ,使其目标函数 $\epsilon_j =$

$\sum_{i=1}^n w_{t,i} |h_j(x_i) - y_i|$ 达到最小;

③ 从上述步骤 ② 确定的弱分类器中找出一个具有最小错误 ϵ_t 的弱分类器 h_t ;

④ 更新每个样本所对应的权重 $w_{t+1,i} = w_{t,i} \beta_i^{1-\epsilon_i}$, ϵ_i 确定的方法为:若第 i 个样本 x_i 被正确分类,则 $\epsilon_i = 0$; 反之 $\epsilon_i = 1$, $\beta_i = \epsilon_i / (1 - \epsilon_i)$;

(3) 最后形成的强分类器为

$$h(x) = \begin{cases} 1 & \sum_{t=1}^T \alpha_t h_t(x) \geq 0.5 \sum_{t=1}^T \alpha_t \\ 0 & \text{otherwise} \end{cases}$$

其中, $\alpha_t = \ln(1/\beta_t)$.

对上述算法的时间复杂度进行分析:循环内样本权重归一化时间复杂度为 $O(n)$; 步骤 ② 中需对特征值进行排序,其时间复杂度为 $O(n \log_2 n)$; 更新样本权重的时间复杂度为 $O(n)$. 假设所有特征数为 K , 则训练弱分类器的时间复杂度为 $K * O(n \log_2 n + 2n)$. 由于实际训练中 K 的数目非常大.因此,训练整个强分类器的时间复杂度为 $T * K * O(n \log_2 n + 2n)$.

3 改进的 AdaBoost 算法

从传统 AdaBoost 算法分析中可以看出,训练时间主要消耗在每次迭代训练都要进行特征值的计算、排序和阈值选择上;同时,当过训练现象发生后,迭代运算仍在进行,此时的训练已经不具有实用性.为了有效地解决上述两个问题,现对 AdaBoost 算法进行如下改进:

(1) 确定共挑选 S 个强分类器,设定每个强分类的平均误差为 T_s ;

(2) 设 $w_{t,i}$ 为第 t 次循环中第 i 个样本的误差权重,对训练样本的误差权重按如下公式初始化:对于 $y_i = 0$ 的样本, $w_{1,i} = 1/2m$; 对于 $y_i = 1$ 的样本, $w_{1,i} = 1/2l$;

(3) 计算所有样本对所有特征的特征值,并对不同特征的特征向量进行排序,排序结果中每个特征值保存样本的序号;

(4) for $t = 1, 2, \dots, T$ (迭代运算, T 为训练的次数)

① 权重归一化,使得 $w_{t,i}$ 为 $w_{t,i} = w_{t,i} / \sum_{i=1}^n w_{t,i}$;

② 对于每个特征 j ,将排序好的特征值结果取

出,并与更新后的样本权值一起,使用改进后的阈值选择算法训练一个弱分类器 h_j ,分类器的错误率: ϵ_j

$$= \sum_{i=1}^n w_{i,j} |h_j(x_i) - y_i|;$$

③ 从上述步骤(2)确定的弱分类器中找出一个具有最小错误 ϵ_t 的弱分类器 h_t ;

④ 更新每个样本所对应的权重 $w_{t+1,i} = w_{t,i} \beta_t^{1-e_i}$, e_i 确定的方法为:若第 i 个样本 x_i 被正确分类,则 $e_i = 0$;反之 $e_i = 1$, $\beta_t = \epsilon_t / (1 - \epsilon_t)$;

⑤ $t = t + 1$;

(5) 计算当前强分类器的平均误差

① 将当前训练得到的若干弱分类器线性组合在一起,得到一个强分类器:

$$h(x) = \begin{cases} 1 & \sum_{i=1}^T \alpha_i h_i(x) \geq 0.5 \sum_{i=1}^T \alpha_i \\ 0 & \text{otherwise} \end{cases}$$

其中 $\alpha_t = \ln(1/\beta_t)$.

② 计算当前强分类器的平均误差:

$$\epsilon = \frac{1}{n} \sum_{i=1}^n |h(x_i) - y_i|$$

③ 如果 $\epsilon \geq T_s$,转步骤(4)中的①.否则,转步骤(6).

(6) 输出当前强分类器,更新训练样本,转步骤(2).

本文改进的 AdaBoost 算法主要体现在两个方面.其一是步骤(3)和(4)中的②,由于只需要一次对所有特征进行特征值计算和排序,其后每次迭代只进行阈值选择计算,这样整个强分类器的时间复杂度变为 $K * O(n \log_2 n + n) + T * K * O(n)$.而在训练级联分类器中, T 越靠后其值越大,同时强分类器层数很少,因此整个训练时间主要有 $T * K * O(n)$ 决定.实验表明,这样的改进可以使得整个级联分类器的训练速度随着样本数目的变化得到大幅度提高;其二是步骤(1)、(5)中的②、③及(6),当强分类器的平均误差大于设定的误差 T_s 时,程序进行迭代训练;反之,则输出当前强分类器并更新训练样本,重新对样本权重进行初始化.由于每次输出强分类器时都有更新后的样本补充到样本训练集中,这样保证了训练出来的分类器具有很强的分类能力,从而避免了过训练问题的发生.

4 更新训练样本

更新样本是指在现有训练样本的基础上对训练样本进行变化处理后重新使用的过程.理论上训练

样本越多,训练结果越好.但是由于训练样本始终是有限的,特别是车牌样本,不仅要数量多而且要有不同角度的旋转,这样训练出来的分类器才有较高的实用性.

针对车牌和非车牌样本,需要采取不同的方式进行更新.其中车牌样本的处理方式由于它的特殊性不可以对其镜像变换,而只有取其小角度旋转(5度以内).图1是车牌图像使用上述方法处理后的结果.

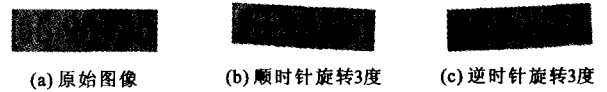


图1 车牌样本更新方式

相对于车牌样本的处理方法,非车牌样本的更新方法就比较多.常用方法可以是镜像、旋转、加噪操作等.通过各种方式都可以达到更新非车牌样本的目的.图2给出了一个非车牌样本进行更新的例子.

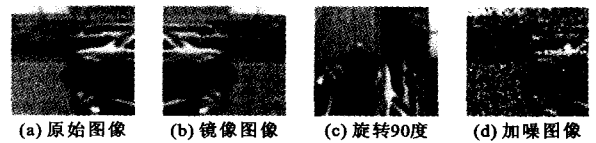


图2 非车牌样本更新方式

5 实验结果与分析

5.1 训练过程分析

将改进的 AdaBoost 算法应用于训练车牌分类器.本文的训练集和测试集都是从一段公路的车流录像中截取出来的.录像流的图片大小是 $352 * 288$ 像素,每秒钟 25 帧图片.由于经过公路某处的车辆不是很连续,因此采用了动态侦测法,当有运动物体经过路口的时候才把此帧图片截取出来用于试验.从此录像流中总共截取了 8 000 幅图片(含车牌和非车牌).对这些图片经过一个样本管理程序的处理从中得到了用于训练的车牌图片 2 500 幅,伪车牌图片 5 000 幅,用于测试的图片 500 幅.最后将这些样本图片进行预处理归一化为 20(20 像素大小,实验中选用了 36 648 个 Haar-like 特征.测试环境为 P4 3.0GHz CPU, 2GB 内存,操作系统为 Windows XP,在 OpenCV1.0 平台上编程实现,得到一个弱分类器的时间约为 41s.训练过程中要求所得强分类器平均误差从 1% 开始每次下降 0.1%,训练总共得到 20 个强分类器,共 1 835 个弱分类器,训练时间约 18.3 个小时.图3为训练得到的弱分类器个数及其对应的检测率关系图.

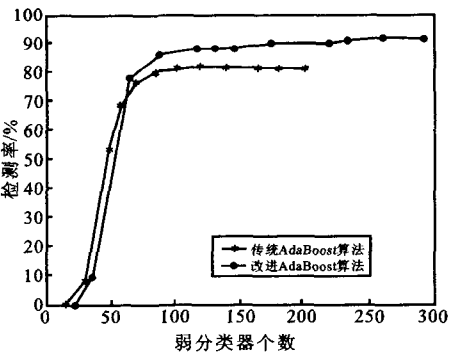


图 3 弱分类器个数与检测率关系

从图 3 中两条曲线可以看出,随着弱分类器个数的增加,传统 AdaBoost 算法的检测率并不是一直增大,而是趋于稳定,这是由于出现了过训练现象.分类器达到一定检测精度后迭代运算仍在进行,此时又没有新的样本补充到样本训练集中,这样训练出来的分类器对样本分类能力基本不变.因此,传统算法的检测率相对较低.本文改进的算法能够有效地避免过训练问题的发生,从而能够保持稳定上升的检测率.

5.2 改进前后训练时间比较

实验中选取 6 000 个样本(2 000 个车牌样本,4 000 个非车牌样本)和 7 500 个样本(2 500 个车牌样本,5 000 个非车牌样本)分别对传统算法和改进的算法进行训练,训练结果见表 1.

表 1 改进前后训练时间比较

训练参数	训练样本 (正+负)	Haar 特征 数量	单个弱分类器 训练时间	训练 总时间	使用弱分 类器总数
改进训练	7 500	36 648	约 41 秒	18.3 小时	1 835
传统训练	6 000	83 699	约 93 秒	39.5 小时	2 418

从实验结果中可以看出,改进后的算法虽然训练样本有所增加,但是单个弱分类器的训练时间却下降了 50% 左右,表明改进后的算法提高了训练速度.

5.3 改进前后检测性能比较

本文分别用传统的 AdaBoost 算法和改进的 AdaBoost 算法构造级联车牌检测器.两种方法构造的检测器在弱分器数目分别在 100 和 150 时,检测率和误检率的比较如表 2 所示.

表 2 传统 AdaBoost 与改进 AdaBoost 算法检测性能的比较

弱分类器个数	算法	检测率/%	误检率/%
100	传统 AdaBoost 算法	81.2	18.4
	本文改进的算法	86.1	15.0
150	传统 AdaBoost 算法	81.4	17.7
	本文改进的算法	89.0	7.3

从表中数据可以看出,改进 AdaBoost 算法明显

比传统 AdaBoost 算法的性能要好,说明采用新算法避免了过训练现象的发生.只要当前强分类器的平均误差低于给定的误差阈值,在输出强分类器的同时随即更新样本训练集,保证了每一轮训练都有新的样本予以补充.所以训练出来的分类器具有很强的分类能力.图 4 为使用本文算法构造的车牌检测器的部分检测结果图.



图 4 不同条件下的检测效果

6 结束语

基于级联结构的 AdaBoost 分类器是解决目标检测问题的有效途径之一.但是,传统 AdaBoost 算法在训练过程要重复计算特征值及排序,它将占用相当多的训练时间;另外,传统算法容易出现过训练的问题,导致训练出来的分类器检测精度不高.

本文提出了一种改进 AdaBoost 算法的方法,新方法在训练过程中对训练样本及时更新,使每一轮循环都有更新后的新样本补充到样本训练集中,这样避免了过训练问题的发生,从而使训练出来的分类器具有较高的检测精度.同时,将特征值和排序结果进行缓存以减少不必要的重复计算,大大缩短了分类器训练时间.因此,改进后的 AdaBoost 算法的整体性能得到了提高.

实验结果证明了改进算法的性能优于传统算法,采用改进算法训练基于级联结构的分类器并将其应用到汽车牌照检测,取得了较好的结果.

参考文献:

[1] Wu Qiang, Huafeng Zhang, Wenjing Jia, et al. Carplate detection using cascaded tree - style learner based on hybrid object features[EB/OL]. [2009 - 03 - 15]. <http://portal.acm.org>.

[2] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C]//Proc of the 13th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1996: 148 - 156.

[3] Schapire R E. A brief introduction to boosting[C]// Proc of the 16th International Joint Conference of Artificial Intelligence. San Francisco: Publishers Inc, 1999: 1401 - 1406.

(下转第 48 页)

参考文献:

- [1] 黄秋元, 雷艳敏, 李卫国. DVB-T 接收系统中 2k-8kFFT 处理器的设计及 ROM 优化[J]. 微电子学与计算机, 2008, 25(11):16-20.
- [2] 元中瑞, 张浩, 邱昕. 802.16e 系统变长可配置 FFT 的设计与实现[J]. 微电子学与计算机, 2009, 26(2):1-4.
- [3] Liu G H, Feng Q Y. ASIC design of low-power reconfigurable FFT processor[C]//ASICON'07. China: Changsha, IEEE, 2007:44-47.
- [4] Wang Chuachin, Huang Jianming, Cheng Hsian chang. A 2K/8K mode small-area FFT processor for OFDM demodulation of DVB-T receivers[J]. IEEE Transactions on Consumer Electronics, 2005, 51(1):28-32.
- [5] Chang Weihsin Chang, Truong Nguyen. An OFDM-specified lossless FFT architecture[J]. IEEE Transactions on circuits and systems, 2006, 53(6):1235-1243.
- [6] Kim H G, Youn J S, Yoon K T, et al. 8K-point pipelined FFT/IFFT with compact memory for DVB-T using block floating-point scaling technique[C]//Information and Communications Technology (ICICT 2007). Dhaka, Bangladesh University, IEEE, 2007:41-44.

作者简介:

姜鑫男, (1983-), 硕士研究生. 研究方向为数字集成电路设计.

唐广男, (1958-), 副教授. 研究方向为电路参数优化设计、模数混合集成电路设计和数字信号处理.

吴斌男, (1976-), 副研究员. 研究方向为宽带通信 SOC.

(上接第 43 页)

- [4] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[J]. IEEE Computer Soc. Computer Vision and Pattern Recognition, 2001(1): 511-518.
- [5] Xiao Rong, Zhu Long, Hongjiang Zhang. Boosting chain learning for object detection[C]//ICCV 2003. China: Beijing, 2003:709-715.
- [6] Wu J, Rehg J M, Mullin M D. Learning a rare event detection cascade by direct feature selection[C]//NIPS. Canada: Vancouver, 2004.
- [7] Freund Y, Schapire R E. A decision-theoretic generalization of online learning and an application to boosting[J]. Computer and System Sciences, 1997, 55(1): 119-139.
- [8] Kim J H, Kwon B G, Kim J Y, et al. Method to improve the performance of the AdaBoost algorithm by combining weak classifiers[EB/OL]. [2009-03-10]. <http://www.ieee.org/portal/site>.
- [9] 蔡津津, 赵杰煜, 王乐珩. AproPhos: 基于 AdaBoost 方法的蛋白质磷酸化修饰预测系统[J]. 微电子学与计算机, 2007, 24(7):35-39.

作者简介:

方义秋女, (1963-), 副教授. 研究方向为模式识别、软件工程.

卢道兵男, (1974-), 硕士. 研究方向为模式识别.

葛君伟男, (1961-), 博士, 教授. 研究方向为模式识别、图像处理、软件工程.