基于 Adaboost 和 CART 结合的优化分类算法

丁 雍,李小霞
(西南科技大学 信息工程学院 模式识别与图像处理实验室,四川 绵阳 621010)

摘 要:提出了一种基于 Adaboost 算法和 CART 算法结合的分类算法。以特征为节点生成 CART 二叉树,用 CART 二叉树代替传统 Adaboost 算法中的弱分类器,再由这些弱分类器生成强分类器。将强分类器对数字样本和人脸样本分类,与传统 Adaboost 算法相比,该方法的错误率分别减少 20%和 86.5%。将分类器应用于目标检测上,实现了对这两种目标的快速检测和定位。结果表明,改进算法既减小了对样本分类的错误率,又保持了传统 Adboost 算法对目标检测的快速性。

关键词:Adaboost;CART;数据挖掘;目标识别;模式分类
中图分类号:TP391.41 文献标识码:A 文章编号:1674-7720(2011)23-0046-05

Optimization of classification based on combination of Adaboost and CART algorithm

Ding Yong, Li Xiaoxia
(School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China)

Abstract: This paper presents a method based on combination of Adaboost and CART algorithm. The method firstly uses CART binary tree as a weak classifier, and then combines these weak classifiers to generate a strong classifier. Compared with the conventional Adaboost algorithm, using the strong classifier in face and digital number classification, the error rates are reduced by 20% and 86.5%. Using the strong classifier on object detection, targets' positions are quickly found out in pictures. The results show that the improved algorithm can not only reduce the classification error, but also maintain the rapidity feature in object detection of traditional Adboost algorithm.

Key words: Adaboost; CART; data mining; object recognition; pattern classification

数据挖掘是从大量的数据中提取出隐含有用信息的过程[1]。分类是数据挖掘的一种重要形式,在分类算法中,Adaboost 算法和 CART (Classification and Regression Tree) 算法在对数据的分类中都有着重要的作用。Adaboost 算法是一种迭代算法,其核心思想是针对同一个分类集训练不同的弱分类器,然后把这些弱分类器结合起来形成一个强分类器进而实现对数据分类,其分类速度快、精度高。2001 年,由 VIOLA P 和 JONES M 将该算法应用于人脸定位[2],算法开始得到快速的发展。此后,LIENHART R 和 MAYDT J 又用此算法成功实现了对不同方位人脸的检测[3]。决策树算法最早是由 HUNT 等人于 1966 年提出的 CLS(Concept Learning System)。当前,最有影响的决策树算法是 QUINLAN 于 1986 年提出的 ID3 和 1993 年提出的 C4.5。CART 算法是基于以上两种方法的改进算法,它采用一种二分递归分割的技术,将当前的样本集分为两个子样本集,使得生成的决策树的每个非叶子节点都有两个分支。因此,CART 算法生成的决策树是结构简洁的二叉树[4],比 ID3 和 C4.5 算法具有更好的抗噪声性能。

本算法是基于以上两种算法的改进算法,在算法的训练过程中,用 CART 算法生成的二叉树代替传统 Adaboost 算法中的弱分类器,然后级联成最终的强分类器,最后通过以实验验证了该算法的可靠性。实验分别以数字图像和人脸图像为样本,训练生成分类器,再分别对若干张测试样本分类并计算出分类误差及误差减小率。在目标检测的实验上,比较了改进算法和传统 Adaboost 算法的优越性,两种算法都能完全检测到目标,且耗时相当。

1 Adaboost 和 CART 算法

1.1 Adaboost 算法

Adaboost 算法的训练过程就是找出若干个弱分类器[5]。设 n 个弱分类器$(h_1,h_2,\cdots,h_n)$是由相同的学习算法形成的,每个弱分类器能单独对未知样本分类成正样本或负样本(二分类情况),通过加权统计弱分类器的分类结果得出最终的分类结果。选择弱分类器的过程中,只要求分类器对样本的分类能力大于自然选择就可以了,即分类错误率小于 0.5。凡是分类错误率低于 0.5 的分类器都可以作为弱分类器,但在实际的训练过程中,还是选择错误率最低的分类器作为该轮选择的弱分类器,表示如下:

$$h_j(x)=\begin{cases}1, & pf_j(x)<p_j\theta_j\\ -1, & \text{其他}\end{cases} \tag{1}$$

其中,$p=\pm1$,用于改变不等式的方向,θ_j 代表某个特征 j 的阈值。Adaboost 算法模型如图 1 所示。

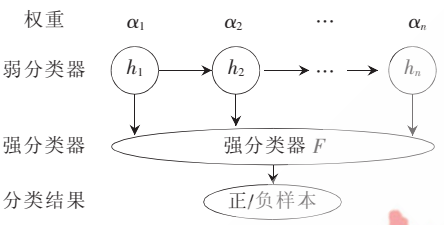


图 1 Adaboost 算法模型

图 1 中,权重代表弱分类器对样本分类的贡献大小,其值越大,表明特征对样本的分类能力越好。分类结果是由 n 个弱分类器加权“投票”的结果,投票结果与某一阈值比较,得出最终对样本的分类。强分类器 F 表示为:

$$F(x)=\begin{cases}1, \sum_{t=1}^{T}\alpha_t h_t(x)\geqslant\frac{1}{2}\sum_{t=1}^{T}\alpha_t\\ -1, \quad \text{其他}\end{cases} \tag{2}$$

1.1.1 Haar-Like 特征

为了应用 Adaboost 算法实现对目标的检测,VIOLA P 和 JONES M 首次引入 Haar-Like 特征表示人脸目标[1],并取得成功。实践证明,对其他目标的表示也可以采用特定的 Haar-Like 特征。Haar-Like 特征表示为一定大小的矩形模板,根据具体待检测的目标形状的不同[6],有不同的特征模板,如图 2 所示。

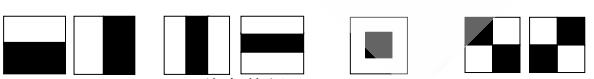


图 2 Haar-Like 特征模板

特征为矩形图像中白色区域内的像素总和减去黑色区域的像素总和,它反映了白色区域到黑色区域的梯度变化情况。

试验中对特征的提取一般都是基于特征图的,特征图可以使计算量大大减少。积分图就是对要处理的图像二次积分,表示如下:

$$f(x,y)=\sum_{i=0}^{x}\sum_{j=0}^{y}g(i,j) \tag{3}$$

其中,$f(x,y)$表示积分图,$g(i,j)$表示原图像。对数字图像而言,点(x,y)处的积分图为该像素左上所有像素的和。

1.1.2 特征生成

特征生成即是将样本图像表示成矢量的形式。以 24×24 样本图为例,生成积分图之后,选择有效的 Haar-Like 特征模板,在积分图中移动,并保存特征值。当一次移动完之后,改变模板大小继续移动取特征值,然后将所有特征按先后顺序排列成一维向量成为代表样本的特征向量。由于模板是在积分图上移动的,因此,每次只需要知道模板的 4 个顶点坐标就可以通过加减法轻松计算出特征值。生成的特征数量相对较多,参考文献[3]具体分析了每个模板对应的特征的个数及其计算公式,统计了所有模板在 24×24 图像上移动生成的特征总数为 117 941 个,即以 117 941 维的矢量表示一个样本图。

1.2 CART 算法

CART 算法是决策树的一种,所不同的是,它的分支始终是二分的。用变量 y 表示分类结果,用 X 表示 p 维特征,该算法首先找出 p 维特征中对分类有效的某个特征 x,将样本分成两个本集子样,以树的左右枝表示,并将此特征作为根节点。接下来判断左右子样本集是否只包含纯样本(全部正样本或全部负样本),如果是,则将此样本集定义为叶子;否则,再次在此子样本集中找出有效特征,继续将子样本集空间划分成左右枝,直到被划分的子样本集中只包含纯样本为止。在同一等级的节点中,可以选取相同属性的特征作为节点,这个划分是以递归方式实现的,最终形成一棵二叉树,形状如图 3 所示。

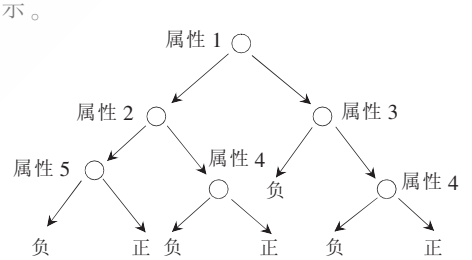


图 3 CART 模型

从根节点到每一个叶子节点,都对应一个规则。分类时,将待测样本的对应特征逐一在此树上从上到下搜索,直到叶子节点,此时,就将该样本的属性划分为该叶节点所表征的类(正样本或负样本)。

在决策树的分支中,常用的分支准则为信息熵法和信息增益法。其中,信息熵是 ID3 算法中常用的分支方法,而信息增益法主要是 C4.5 和 CART 中常用的分支方法。

信息熵本为通信电路携带信息量的大小,在这里反映的是某一个特征阈值对样本的划分准确率。对于训练例集 U,假设有 m 个类别,全局信息熵表示为:

$$E=\sum_{k=1}^{m}-\frac{s_i}{s}\log_2\frac{s_i}{s} \tag{4}$$

其中，s_i 表示 i 类中正样本的个数，s 为总样本个数。在 CART 算法中，因为每个节点都是二分的，即将样本分成两部分，所以熵的表示也就相对简单。假设其中的正样本出现的概率为 p_+，则负样本出现的概率就是 $p_-=1-p_+$，信息熵的公式表示为：

$$E=-p_+\log_2 p_+-p_-\log_2 p_- \tag{5}$$

如某一特征阈值将正负样本完全分开，此时被分开的每个子集的信息熵就达到最小。设训练样本空间为 U，以某一特征 A 将样本空间划分为 U_1 和 U_2 两个子集，在子空间，如果包含 20 个样本，10 个正样本和 10 个负样本，则正样本的概率等于负样本的概率，即 $p_+=p_-=0.5$，带入式(5)可计算得到此空间的信息熵达到最大值 1。与此类似，如果样本空间 U_1 为同一样本，则计算得到熵的最小值 0。如果用属性 A 将训练集划分为两个子集 S_1 和 S_2，每个子集中的信息熵又按照式(5)计算，分别用 E_{s_1} 和 E_{s_2} 表示，此时的正负样本概率都以该子集中的样本为依据统计。然后得出信息期望熵：

$$E(U,A)=\frac{E_{s_1}}{E}\times E_{s_1}+\frac{E_{s_2}}{E}\times E_{s_2} \tag{6}$$

CART 算法对节点的分支依赖于信息熵增益，即选取信息增益熵最大的特征作为一个节点。信息熵增益反映了全局信息熵降低的程度，信息熵增益越大，表明特征对样本分类越有利，信息熵增益表示如下：

$$G=E-E(U,A) \tag{7}$$

由于噪声的存在，决策树往往出现枝叶过于茂盛或者树干过长的情况，在分类的过程中，这会导致对训练数据过度拟合，使分类的错误率升高，反而不能对验证数据很好地分类。所以，一棵优秀的决策树应该包含剪枝的过程，即用验证数据将树的叶子或节点修剪，防止其对训练数据的过度拟合。剪枝算法有多种，常见的有前向剪枝和后向剪枝两种，CART 算法采用的是后向剪枝算法。

2 改进算法

2.1 算法原理

Adaboost 算法在每一轮的训练过程中都会判断某一单独特征对训练样本的分类能力，然后加大被错误分类样本的权重，减少被正确分类样本的权重。由于权重在每一轮训练完成之后都在改变，因此，每次选择的特征并不一定是最好特征，只是在当前权值条件下分类最好的特征。为了改善弱分类器对样本的分类能力，选择一棵具有 3 个节点的二叉树代替原来的弱分类器，即每轮训练都找出 3 个对分类最优的特征，构成一棵树。弱分类器的分类结果由这 3 个特征共同决定，比起只用单独特征分类的弱分类器而言，它对样本的分类能力更高。由于每个弱分类器对样本的分类能力提高了，因此，最终的强分类器的分类能力也将提高。为了与 Adaboost 算法中的弱分类器 h 区别，改进算法中的弱分类器用 H 表示。图 4 描述了本算法形成的分类器模型。

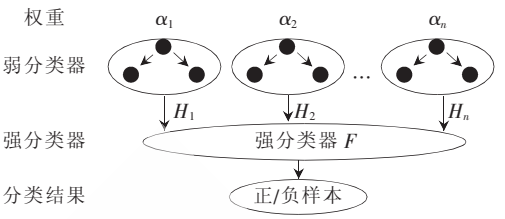


图 4 改进算法的分类器模型

2.2 算法步骤

根据图 4 的分类器模型，算法步骤如下：

(1) 给定训练集图像 $I_1, I_2, \cdots, I_n$，规范化到相同尺寸，计算积分图，利用 Haar-like 矩阵生成特征。

(2)给定训练集 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，其中，$x_1, x_2, \cdots, x_n$ 代表图像的特征；$y_i=0,1$，分别代表负样本图像和正样本图像。

(3)为正负样本分别初始化权重 $w_{1i}=\frac{1}{2m},\frac{1}{2n}$，其中 m、n 分别为训练集中正样本和负样本的个数。

(4)For $t=1,\cdots,T$

①归一化权重：

$$w_{t,j}\leftarrow\frac{w_{t,j}}{\sum_{j=1}^{n}w_{t,j}}$$

②用 CART 算法构建二叉树，作为弱分类器：

(a)计算全局信息熵 E。

(b)对每一个特征 j，训练一个小分类 h_j 器，小分类器与传统的 Adaboost 算法中的弱分类器相同，用该小分类器对样本分类，并计算分类错误率 $\varepsilon_j=\sum_i w_i|h_j(x_i)-y_i|$。

(c)找出错误率最小的 3 个小分类器 h_1、h_2、h_3，计算由它们划分的两个子样本集的信息熵，计算信息期望熵 $E(U,A)$，并计算信息增益 $G=E-E(U,A)$。

(d)选择信息熵增益最大的特征作为根节点，将该节点分支，分成两个子集。

(e)在两个子集中分别按照步骤(b)~(d)再次找到两分支节点，生成有 3 个节点的二叉树。

③保存上面生成的 3 个节点的二叉树和权值 w_i，作为此轮训练得出的弱分类器 H_t。

④更新权重：$w_{t+1,j}=w_{t,i}\beta_t^{1-e_i}$。其中，如果样本被正确分类，$e_i=0$；否则 $e_i=1$，$\beta_t=\frac{\in_t}{1-\in_t}$。

⑤最终的强分类器为：

$$F(x)=\begin{cases}1, & \sum_{t=1}^{T}\alpha_t H_t(x)\geqslant\frac{1}{2}\sum_{t=1}^{T}\alpha_t\\ -1, & \text{其他}\end{cases}$$

其中，$\alpha_t=\log\frac{1}{\beta_t}$。

在训练步骤(c)中，考虑了分类错误率和信息熵增益两个因素对分类的影响。算法在每一轮训练中都选择了对分类错误率最小的3个特征，然后再从其中计算信息熵增益最大的特征作为节点。这样的选择保证了弱分类器也能有较小的分类误差，因此最终的强分类器也有较小的分类误差。

3 实验结果与分析

3.1 实验描述

为了说明改进算法的效果，在相同条件下得出了两种算法的实验结果并进行了比较。实验一以人民币图像中的数字0作为样本，样本图像均由人工采集，在不同面额的人民币图像上采集得到正样本图像和负样本图像各500张。实验二以人脸图像作为样本，样本来源于AR (AR Face Database.http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html)人脸数据库，正负样本各1 000张。

实验均选择以图2(a)和图2(b)的Haar-Like模板生成特征。实验过程采用交叉验证的方式完成，所有实验均在同一条件下进行。实验条件：PC机采用AMD Athlon™II x2220 2.81 GHz处理器和2 GB内存；代码执行平台为MATLAB7.0。

3.2 实验结果

训练样本的数量越大，越能够反映真实样本的分布情况，在训练的过程中，也更能提取出对分类有效的特征。实验首先以数字图像样本为研究对象，以不同数量的样本训练分类器，然后将生成的分类器对200个测试样本分类，得到图5。可以明显看出，随着训练样本数量的增加，分类误差呈现下降的趋势，其下降的速率先快后慢，最后基本稳定在某一数值。实验还发现，当训练样本数量远远大于测试样本时，能够使测试样本的分类误差达到最小。试验中，在选取900个训练样本、100个测试样本的条件下，能够将100个测试样本完全分开，分类准确率达到100%。

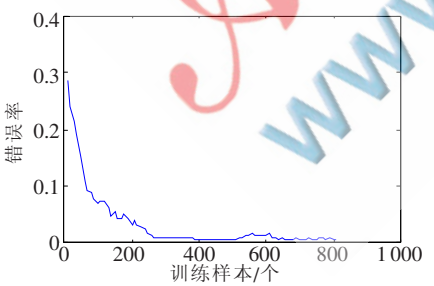


图5 训练样本数量对分类器性能的影响

由以上实验结果可知，当训练样本数量高于300个时，其分类误差基本保持在某一数值。为此，实验中将全部1 000个样本分成500个训练样本和500个测试样本(训练样本和测试样本中均各含250个正样本和负样本)，分别用传统的Adabost算法和改进算法生成强分类器，对测试样本分类。图6显示了两种分类器的分类误差随着训练次数的变化情况。

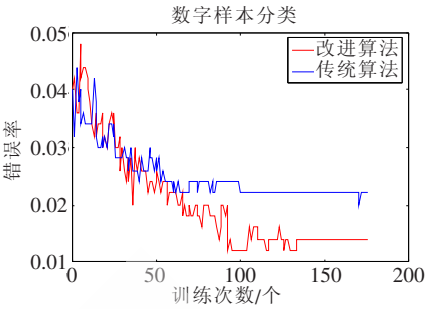


图6 两种算法对数字样本的分类

由图6可以看出，随着训练次数的增加，两种分类器对测试样本的分类误差逐渐减小。在训练次数高于某个数值之后，改进算法的错误率明显低于普通的Adaboost算法，说明改进算法的分类能力较强。由于实验中所选样本的可分性较强，因此，无论是传统的Adaboost算法还是改进算法，其分类误差都比较低(小于1%)。

为了验证算法鲁棒性，实验从AR人脸库中得到正负人脸样本各1 000张，再次比较两种算法的分类情况，如图7所示。从图中可以看出，改进算法对特征不明显的人脸图像分类照样能达到较高的分类精度(99.3%)，高于普通的Adaboost算法(94.8%)，这说明本算法的鲁棒性较强。

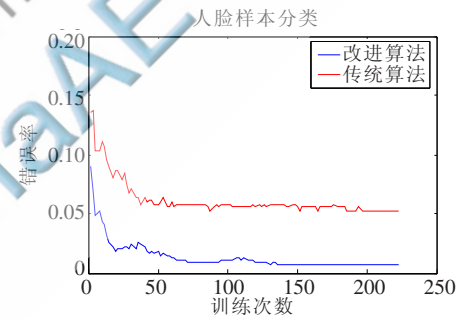


图7 两种算法对人脸样本的分类

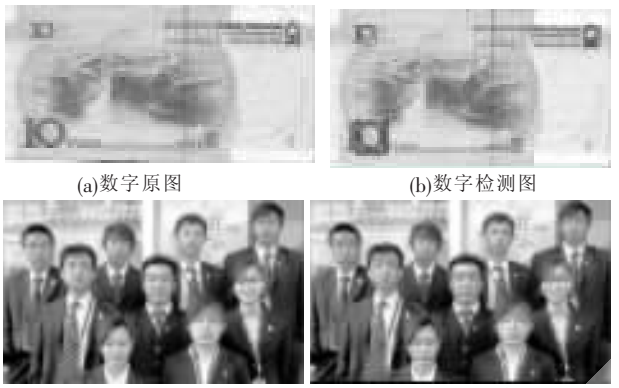
表1统计了两种算法对两类样本分类的一些参数。训练样本和测试样本各占总样本的1/2，均训练300次。其中，误差减小率表示改进算法的分类误差相对于传统Adaboost算法分类误差的减小程度。

表1 算法误差比较

样本	训练样本数/个	测试样本数/个	改进算法误差	Adaboost误差	误差减小率
数字	500	500	0.016	0.02	20%
人脸	1 000	1 000	0.007	0.052	86.50%

从表1可以看出，改进算法对不同的目标样本分类能力均有所提高，并且提高的程度有所不同。数字样本的Harr-like特征较明显，所以，无论是改进算法还是普通的Adaboost算法，分类误差都较小，而且误差减小率也相对较小。而从两种算法对人脸样本的分类可以看出，改进算法能明显减小分类误差，提高分类器的分类能力。

将生成的分类器应用于目标检测，能够快速检测出目标在图像中的位置。由于改进算法的实现过程保留了传统 Adaboost 算法中以 Haar-Like 模板提取特征的过程，因此两种算法耗时相当。图 8(b)和图 8(d)分别是以上生成的数字分类器和人脸分类器对两种目标的检测结果。



(a)数字原图 (b)数字检测图

(c)人脸原图 (d)人脸检测图

图 8 两种分类器应用于目标检测结果

本文以 Adaboost 算法和 CART 算法为基础，提出了将这两种算法相结合的改进算法，从理论上详细阐述了算法的原理和步骤。算法的关键在于，在训练样本的每一轮训练中寻找出对分类最有利的 3 个特征，形成二叉树，用来代替传统 Adaboost 算法中的弱分类器。树的形状是根据 CART 算法改进的，提高了单个弱分类器对样本的分类能力，由于强分类器由弱分类构成，因此，强分类器的分类能力也得到提高。最后以人民币图像上的数字 0 和人脸图像为样本，验证了本算法的可靠性和鲁棒性。较普通的 Adaboost 算法而言，改进算法对数字样本和人脸样本的分类误差率分别减少 20%和 86.5%，说明算法对样本的分类能力有所提高。改进算法的每轮训练都要生成有 3 个节点的二叉树，其训练过程将更加耗时，约等于普通算法的 3 倍。可以说，改进算法是以更长的训练耗时换取更高的分类精确度。由于改进算法在特征提取过程中保持了传统 Adaboost 算法的步骤，因此两种算法在目标检测的应用中耗时是相当的。

参考文献

[1] 毛国君. 数据挖掘的概念、系统结构和方法[J]. 计算机工程与设计，2002，23(8)：13-17.

[2] VIOLA P，JONES M. Rapid object detection using a boosted cascade of simple features [C]. Accepted Conference on Computer Vision and Pattern Recognition，2001(5)：511-518.

[3] LIENHART R，MAYDT J. An extended set of Haar-like features for rapid object detection [C]. IEEE ICIP 2002，2002，1：900-903.

[4] YOHANNES Y，HODDINOTT J. Classification and regression trees：an introduction [C]. International Food Policy Research Institute 2033 K Street，N.W. Washington，D.C. 20006 U.S.A.，1999.

[5] HORE U W. Comparative implementation of colour analysis based methods for face detection algorithm [C]. Emerging Trends in Engineering and Technology (ICETET)，2010(3)：176-179.

[6] LISU L. Research on face detection classifier using an improved adaboost algorithm [C]. International Symposium on Computer Science，2009(2)：199-204.

[7] FREUND Y，SCHAPIRE R E. Experiments with a new boosting algorithm[C]. Machine Learning：Proceedings of the Thirteenth International Conference，San Francisco，1996(5)：148-156.

(收稿日期：2011-09-13)

作者简介：

丁雍，男，1985 年生，硕士研究生，主要研究方向：目标检测算法、机器学习算法、人工智能算法。

李小霞，女，1976 年生，副教授，博士后，主要研究方向：生物信息学、生物特征识别、生物医学光子学、光谱检测与仪器。