文章编号:1001 - 9081(2011)02 - 0396 - 03 doi:10.3724/SP. J. 1087. 2011. 00396

基于二叉树和 Adaboost 算法的纸币号码识别

潘 虎,陈 斌,李全文
(中国科学院 成都计算机应用研究所,成都 610041)
(tt. timmy@ hotmail. com)

摘 要:运用一种快速弱分类器训练算法和高速缓存策略来加速 Adaboost 算法的训练。集成学习算法 Adaboost 能够精确构建二分类器,运用二叉树型结构快速灵活地将纸币号码识别转化为一系列的 Adaboost 二分类问题。实验结果证明,快速 Adaboost 训练算法能加快训练速度,基于二叉树和 Adaboost 的纸币号码识别系统具有较好的识别率和处理速度,已经应用在点钞机、清分机和 ATM 中。

关键词:Adaboost 算法;快速 Adaboost 算法;二叉树;号码识别
中图分类号: TP391.41 **文献标志码**:A

Paper currency number recognition based on binary tree and Adaboost algorithm

PAN Hu, CHEN Bin, LI Quan-wen
(*Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu Sichuan* 610041, *China*)

Abstract: A fast weak classifier training algorithm and a fast caching strategy were used to accelerate Adaboost training. Integrated learning algorithm Adaboost can accurately construct two classifiers, so paper currency number recognition was formulated as a series of Adaboost two-class classification problems quickly and flexibly by using binary tree structure. The experimental results demonstrate that the fast Adaboost training algorithm can speed up the training and the paper currency number recognition system based on binary tree and Adaboost algorithm has good recognition rate and processing speed, and it has widely been used in currency counter, cash sorter and ATM.

Key words: Adaboost algorithm; fast Adaboost algorithm; binary tree; number recognition

0 引言

纸币是市场上一般等价物的主要形式,它在人们的社会生活中起着不可替代的作用。纸币上的号码是纸币的非常重要的标识之一,可以用来识别纸币的身份。如果能开发一种智能纸币号码识别系统,自动记录下通过点钞机的纸币的号码,并存储备案,就可以实现对纸币号码的便捷管理。这就要求所采用的识别算法既有很高的识别率,又具备足够快的识别速度。Adaboost 算法是一种构造准确分类器的学习算法,预先把负样本和正样本加入待学习样本,通过机器学习,将一系列比随机预测略好的弱分类器线性组合为识别能力很强、识别范围更广的强分类器。

Adaboost 广泛地应用在人脸检测[1]、汽车牌照号码快速定位[2]等领域。文献[3]将 Adaboost 成功用于手写体数字的识别。Adaboost 的训练随着样本的增加会非常耗时,常见的处理方法是对训练样本进行数据降维处理。文献[4]用到的随机投影和主成分分析(Principal Component Analysis, PCA)会降低分类器性能。快速阈值选择和高速缓存策略能够有效缩减 Adaboost 的训练时间;二叉树型分类树结构简单,能够充分发挥 Adaboost 解决二分类问题的优势,对于 k 类分类问题只需构造 $k-1$ 个 Adaboost 分类器。将两者结合起来应用到纸币号码的识别,能有效提高计算机自动号码识别的速度和准确率。

1 Adaboost 算法简介

1.1 弱可学习定理

1984 年,Valiant[5]提出机器学习的另类理念,即学习模型无需绝对精确,只需概率近似正确(Probably Approximately Correct,PAC)即可。1990 年,Schapire[6]通过构造性方法证明了弱学习算法和强学习算法的等价性,即可以将若干个略好于随机猜测的弱学习算法提升为强学习算法,而不必直接去找通常情况下很难获得的强学习算法,这就是著名的弱可学习定理。

1.2 Adaboost 算法

弱可学习定理是 Boosting 算法的理论基础。1990 年,Schapire[6]提出了 Boosting 算法,该算法的基本思想是找出若干个精度比随机预测略高的弱学习规则,再将这些弱学习规则组合成一个高精度的强学习规则。在众多的 Boosting 算法中,Freund 和 Schapire[7]在 1995 年提出的 Adaboost 算法最具有实用价值,也是目前研究的热点。Adaboost 算法的美妙之处在于它使用加权后选取的训练数据代替随机选取的训练样本;将弱分类器联合起来,使用加权的投票机制代替平均投票机制。

Adaboost 算法被 Viola 和 Jones 成功运用于人脸检测[1],其算法步骤如下:

1)给定训练样本 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,$y_i \in \{-1,1\}$。其中 $y_i=-1$ 代表负样本,$y_i=1$ 代表正样本。

2)初始化权值。对于负样本,其权重 $w_{1,i}=1/(2m)$,m 为负样本个数;对于正样本,其权重 $w_{1,i}=1/(2(n-m))$,$n-m$ 为正样本个数。

3)For $t=1,2,\cdots,T$

①归一化权值,$w_{t,i} \leftarrow w_{t,i}/\sum_{j=1}^{n} w_{t,j}$,确保 w_t 服从一个概率分布。

②对每一个特征 j,训练一个弱分类器 h_j,其误差由样本

收稿日期:2010 - 08 - 16;**修回日期**:2010 - 10 - 11。

作者简介:潘虎(1986 -),男,湖南岳阳人,硕士研究生,主要研究方向:图像处理、模式识别; 陈斌(1970 -),男,四川广汉人,研究员,博士,主要研究方向:图像处理、模式识别、工业视觉; 李全文(1984 -),男,广西桂林人,硕士研究生,主要研究方向:图像处理、模式识别。

分布的权值 w_t 来衡量，$\varepsilon_j = \sum_i w_i |h_j(x_i) - y_i|$。

③挑选具有最小误差 ε_t 的分类器 $h_t = \arg\min_{h_t} \varepsilon_j$。

④更新样本的权值 $w_{t+1,i} = w_{t,i}\beta_t^{1-e_i}$，$\beta_t = \varepsilon_t/(1-\varepsilon_t)$，当样本 x_i 被正确分类时，$\varepsilon_i = 0$，否则 $\varepsilon_i = 1$。

4)最终的强分类器为：

$$h(x) = \text{sign}\Big[\sum_{t=1}^{T}\alpha_t h_t(x) - \frac{1}{2}\sum_{t=1}^{T}\alpha_t\Big]$$

其中 $\alpha_t = \log_t \frac{1}{\beta_t}$。

2　纸币号码识别

2.1　预处理

本系统做的预处理是将在清分机 CIS 灯光下获取的纸币灰度图像进行搜边定位、倾斜校正、号码区域定位，分割得到号码图像，如图 1 所示。然后进行单个字符的分割得到单个号码的灰度图像，最后对单个字符归一化，其中阿拉伯数字图像全部归一化到 19 × 29，英文字母图像全部归一化到 23 × 35。Adaboost 是一种基于集成学习方法训练出来的分类器，具有很强的学习泛化能力，可以直接应用于灰度图像，字符不需要二值化，和传统的模板匹配、字符穿刺等方法相比，不仅节省了时间，也避免了对字符轮廓可能造成的偏差，降低了误识率。

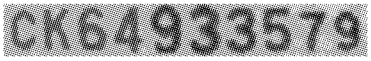

图 1　纸币号码图像

2.2　弱分类特征的选取

特征的选择和提取是纸币号码识别的重要步骤，纸币号码由 2 个英文字母和 8 个阿拉伯数字组成。Viola 将很简单的 Haar-Like 矩形特征成功应用于人脸检测[1]。纸币号码字符结构简单，特征提取用到几种常用的 Haar-Like 矩形特征，如图 2 所示，分别为 2-矩形特征、3-矩形特征和 4-矩形特征。

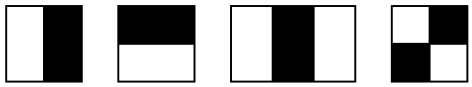

图 2　Haar-Like 矩形特征

对于每个样本，其矩形特征对应有一个特征值。2-矩形特征和 4-矩形特征的特征值为所有白色矩形区域的灰度值的和减去所有黑色矩形区域的灰度值的和所得到的差，3-矩形特征的特征值为所有白色矩形区域的灰度值的和减去所有黑色矩形区域的灰度值的和的 2 倍所得到的差。

特征值的计算是通过积分图像快速实现的，对于一幅灰度图像 I，它的积分图的定义为：

$$ii(x,y) = \sum_{x'=1}^{x}\sum_{y'=1}^{y} i(x',y')$$

其中：$i(x',y')$ 为像素值。将图像所有像素的位置的积分图用一个二维数组存储，则图 2 中矩形 D 内的灰度值的和可以由积分图上的四点的坐标快速计算出。不妨设图 3 中 1、2、3、4 共 4 个点的坐标分别为 (x_1,y_1)、(x_2,y_2)、(x_3,y_3) 和 (x_4,y_4)，则矩形 D 内的灰度值的和为 $ii(x_4,y_4) - ii(x_2,y_2) - ii(x_3,y_3) + ii(x_1,y_1)$。

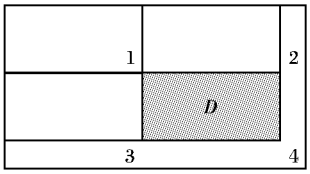

图 3　积分图像

2.3　弱分类器的构造

给定一个特征集和带标记的正负样本集，存在许多将它们分类的算法。采用最简单的单阈值加偏置值的函数来构建的弱分类器，具有很好的对称性。对于第 j 个特征 f_j 和样本 x，其对应的弱分类器 $h_j(x)$ 定义为：

$$h_j(x) = \text{sign}[pf_j(x) - p\theta_j]$$

其中：$f_j(x)$ 为特征值；θ_j 为阈值；偏置值 $p = \pm 1$。输入样本 x 为一幅数字图像，一个特征 f_j 对应一个弱分类器 $h_j(x)$。该弱分类器结构具有对称性，即当 θ_j 确定，且偏置值 $p = 1$ 和 $p = -1$ 时，分类器的误差之和为 1。

2.4　弱分类器的快速训练

Adaboost 是一种统计学习算法，需要大量的样本进行训练。一个弱分类器的训练就是寻找最优阈值 θ_j 的过程，一轮分类器的训练过程就是寻找最优弱分类器 $h_j(x)$ 的过程。训练所用的阿拉伯数字图像全部归一化到 19 × 29，将图 2 中的每种 Haar-Like 矩形特征用五元组 $(x,y,w,h,style)$ 表示，记录矩形特征左上角坐标、宽度高度和类型。根据组合原理，穷举计数所有的五元组，得到 19 × 29 的数字图像中 Haar-Like 特征的总数目高达 103 080。英文字母图像全部归一化到 23 × 35，特征数目更加巨大。如果按照传统的 Adaboost 弱分类器的训练算法，阈值的计算和选择将导致训练极其地耗费时间。利用弱分类器结构的对称性[8]，通过对特征值预排序并缓存可以实现弱分类器的快速训练，具体算法流程如下。

1)输入：n 个训练集 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，$y_i \in \{-1,1\}$，权值为 $\{w_i\}_{i=1}^n$，矩形特征 f_j。

2)将特征 f_j 在所有样本的特征值 $v_1,v_2,\cdots,v_n$，将它们从小到大排序，得到一个次序表 $v_{i,1} \leqslant v_{i,2} \leqslant \cdots \leqslant v_{i,n}$。

3)初始化 $p = +1$，$\theta_j = v_{i,1}$，$\varepsilon_1 = \varepsilon(v_{i,1}) = \sum_{y_i=1} w_i$，即阈值为 $v_{i,1}$ 时弱分类器的误差。

4) for $k = 1, 2, \ldots, n-1$ do
　if $y_{i,k} = 1$ then
　　$\varepsilon(v_{i,k+1}) = \varepsilon(v_{i,k}) - w_{i,k}$;
　else
　　$\varepsilon(v_{i,k+1}) = \varepsilon(v_{i,k}) + w_{i,k}$;
　end if
end for

5) $t^+ = \underset{1\leqslant k\leqslant n}{\text{argmin}}\,\varepsilon(v_{i,k})$，$t^- = \underset{1\leqslant k\leqslant n}{\text{argmax}}\,\varepsilon(v_{i,k})$，$\theta^+ = v_{i,t^+}$，$\theta^- = v_{i,t^-}$。

6) if $\varepsilon_{t^+} \leqslant 1 - \varepsilon_{t^-}$ then
　输出：弱分类器 $h_j(x) = \text{sign}[\theta^+ - f_j(x)]$
else
　输出：弱分类器 $h_j(x) = \text{sign}[f_j(x) - \theta^-]$

预先对特征值的排序可以在弱分类器训练过程中充分利用弱分类器结构的对称性，计算误差时可以利用上一次的结果。在每轮 Adaboost 训练中，样本 $\{(x_i,y_i)\}_{i=1}^n$ 本身并没有改变，更新的只是样本的权重 $\{w_i\}_{i=1}^n$；对于每一个特征，样本特征值和排序结果在训练过程中也不会发生变化。因此在第一次计算特征值和对特征值排序时，对每个特征的特征值 $\{v_i\}_{i=1}^n$ 的排序结果 $v_{i,1},v_{i,2},\cdots,v_{i,n}$ 新建一个 $M \times N$ 的序列表 G，M 为所有特征数目，N 为样本总数。此后的迭代训练都直接从缓存中读取特征值序列，进行阈值选择计算。除了第一次迭代需要消耗较长时间，之后每次迭代只需花费较少时间，以空间换时间的高速缓存策略是提升特征选择速度的关键。在算法时间复杂度上，原来算法训练的时间复杂度为 $O(NMT\log N)$，使用高速缓存策略训练的时间复杂度为

$O(NM \log N + NMT)$，T 为迭代次数。

2.5 二叉树型多分类结构

模式识别问题归根结底就是分类问题。由于纸币号码是字母的数字的组合，类别较多，Adaboost 用于多类问题时，思想还是转化为两类问题。以 Adaboost 分类器为节点的二叉树分类结构具有如下优势：Adaboost 解决二分类问题已经非常成熟，在人脸检测等领域的应用都取得了极大的成功，因此在此基础上实现的二叉树形多重分类器也较容易得到很高的识别率；同时二叉树型结构直观、便于理解，树形结构也在一定程度上减少了比较的次数，节约了时间。

对于阿拉伯数字 0～9 的识别，建立二叉树分类结构，每个节点处设置一个 Adaboost 分类器，共需要设计 9 个分类器，如图 4 所示，号码识别的过程就是从树根到任意叶子节点的分类过程。该二叉树型结构的设计也并不是唯一的，主要是依据一些先验知识。实验结果表明，这种结构模型很好地利用了先验知识来提高分类器性能。

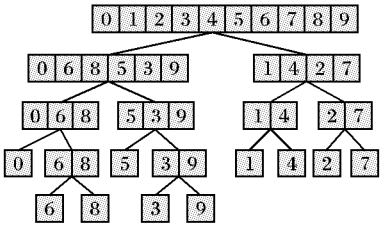

图4 阿拉伯数字二叉树分类结构

对于英文字母的识别，同样依据一些先验知识以及字符的结构特征差异如字符 4 个方向圆弧的饱满程度，对纸币号码里的 25 个英文字母（人民币编码冠字不包括 V）用 Adaboost 分类器分成 2 个子树，左子树节点为 B、C、D、E、G、O、Q、S、Z，其对应的二叉树分类结构如图 5 所示。右子树节点为 A、F、H、I、J、K、L、M、N、P、R、T、U、W、X、Y，其对应的二叉树分类结构如图 6 所示。

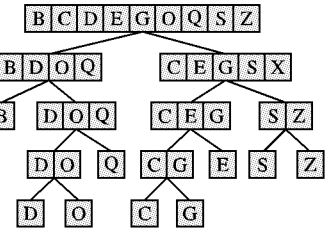

图5 英文字母左子树二叉树分类结构

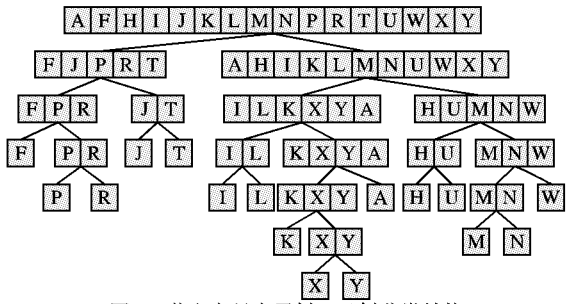

图6 英文字母右子树二叉树分类结构

3 实验结果

训练样本集是由高速纸币清分机 CIS 白光扫描通过的 100 元人民币获取的，包括 2116 个阿拉伯数字，5858 个字母，运用快速 Adaboost 弱分类器训练算法和缓存策略来训练每一层的 Adaboost 强分类器，每一层训练 35 个弱分类器。在 Intel Pentium E5300@2.6 GHz 机器和 VC2008 编译环境下分别对 OpenCV 训练算法（OpenCV 采用的是原始 Adaboost 训练算法）与改进后的训练算法在每一层训练所需要的时间进行了测试对比，结果如表 1，改进后的训练策略使 Adaboost 的训练时间大大缩减。

表1 OpenCV 和快速 Adaboost 的训练时间对比

正样本数	负样本数	OpenCV 方法时间/min	快速算法时间/min
86	221	12	4
251	471	35	13
1364	752	132	38
980	2078	257	51
2800	3058	678	166

测试用的纸币号码单个字符是由清分机 CIS 白光实时对随机高速通过的 1384 张人民币扫描，并经过一系列预处理获取的。测试中阿拉伯数字的识别率非常高，仅有 3 个识别错误，全部是把 5 错识别为 6；英文字母的识别率因其二叉树分类结构更加复杂而略有降低。纸币号码的识别率和识别速度如表 2 所示。

表2 纸币号码识别率和平均识别时间

符号	测试数	误识数	识别率/%	平均识别时间/ms
阿拉伯数字	13840	3	99.97	3.4
英文字母	3460	29	99.16	3.4

实验所采用的基于 Adaboost 的二叉树型多分类器对纸币号码的识别率很高，单个字符的平均识别时间在 3.4 ms，能够满足工业上实时应用的要求。

4 结语

本文提出的基于二叉树和 Adaboost 的纸币号码识别系统对纸币号码的实时识别取得了非常理想的识别效果。实践证明，快速阈值选择和缓存策略能够加速 Adaboost 训练，Adaboost 方法可以充分发掘样本间的差异，产生高精度的二分类器；二叉树分类结构能够充分发掘Adaboost的泛化能力，灵活地将多分类问题转换为一系列简单的二分类问题。沿着此应用思路，还可以把 Adaboost 学习算法应用到更多的识别、检测等领域。

参考文献：

[1] VIOLA P, JONES M. Robust real-time face detection[J]. International Journal of Computer Vision, 2004, 57(2): 137－154.

[2] 潘石柱, 殳伟群, 王令群. 基于 Adaboost 的汽车牌照快速定位[J]. 计算机工程, 2006, 32(12): 187－188.

[3] 赵万鹏, 古乐野. 基于 Adaboost 的手写体数字识别[J]. 计算机应用, 2005, 25(10): 2413－2417.

[4] PAUL B, ATHITHAN G, MURTY M N. Speeding up AdaBoost classifier with random projection[C] // 7th International Conference on Advances in Pattern Recognition. Washington, DC: IEEE Computer Society, 2009: 251－254.

[5] VALIANT L G. A theory of the learnable[J]. Communication of the ACM, 1984, 27(11): 1134－1142.

[6] SCHAPIRE R E. The strength of weak learnability[C]// 30th Annual Symposium on Foundations of Computer Science. Washington, DC: IEEE Computer Society, 1989: 28－33.

[7] FREUND Y, SCHAPIRE R E. A decision theoretic generalization of online learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119－139.

[8] WU JIAN XIN, BRUBAKER S C, MULLIN D M, *et al*. Fast asymmetric learning for cascade face detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(3): 369－382.