文章编号: 1001 - 9081(2005)10 - 2413 - 02

基于 Adaboost的手写体数字识别

赵万鹏,古乐野
(中国科学院 成都计算机应用研究所,四川 成都 610041)
(wanpeng_zhao@yahoo.com.cn)

摘 要:提出了一种新的基于集成学习算法 Adaboost的手写体数字识别系统。Adaboost方法可以在仅比随机预测略好的弱分类器基础上构建高精度的强分类器。实验证明,基于 Adaboost的手写体数字识别系统具有较高的识别率和泛化能力,已经应用在 OCR识别软件中。

关键词: Adaboost;手写体数字识别;弱分类器
中图分类号: TP18 **文献标识码**: A

Handwritten digit recognition based on Adaboost

ZHAO Wan-peng, GU Le-ye
(*Chengdu Institute of Computer Application, Chinese Academy of Science, Chengdu Sichuan* 610041, *China*)

Abstract: A handwritten digit recognition system based on Adaboost algorithm was introduced in this paper. Adaboost could construct a highly accurate classifier by combining many weak classifiers that just had slightly better accurate than random prediction. Experiment proved that the handwritten system based on Adaboost have low error rate and good ability of generalization. And it was integrated in a ocr software.

Key words: Adaboost; handwritten digit recognition; weak classifier

0 引言

在 OCR(光学字符识别)技术中,手写体数字识别是一个特别的问题,在邮件的自动分拣、工商财务报表的自动录入、考试的自动化报名系统中,都有着广泛的应用背景。同时,手写体数字识别又是比较困难的,特别是脱机的手写体数字识别,成为模式识别领域的经典问题。而且,在很多具体的应用中,需要对扫描设备扫进的数字图像进行实时的识别。这就要求所采用的识别算法既有较高的识别率,又具备足够快的识别速度。

本文采用了模式识别领域较新颖的集成学习算法 Adaboost,提出一种将多分类的问题转换为一系列二分类问题的方案,使得 Adaboost可以应用在典型的多分类问题——手写体数字识别上,并取得了较好的效果。

1 Adaboost算法简介

1.1 Boosting算法

Boosting算法的基本思想就是,找出若干个、精度比随机预测略高的弱规则,再将这些弱规则组合成一个高精度的强规则。这个算法思想起源于 Valiant提出的 PAC学习模型,而且 Kearns和 Valiant证明,只要有足够多的数据,弱学习算法就能够通过集成的方式,生成任意高精度的估计。

1.2 Adaboost算法

在众多的 Boosting算法中, Freuch 和 Schapire 提出的 Adaboost算法最具有实用价值,也是目前研究的热点。Adaboost的主要思想是:给定一个训练集合 $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_m, Y_m)\}$,其中, X_i 属于某个域或者实例空间 X, $Y_i \in \{-1, +1\}$。初始化时, Adaboost指定训练集上的分布为 $1/m$,并按照该分布调用弱学习器对训练集进行训练。训练结束后,按照训练的结果更新权值在训练集上的分布,使得本次训练分类错误的样本,在下次训练中得到更多的关注。按照新的样本分布再进行训练,反复迭代 T轮,最终得到一个估计序列 $H1, H2\cdots\cdots: HT$,每个估计都具有一定的权重。最终的估计采用投票的方式获得。

Adaboost算法提出来以后,研究者采用了很多种类的弱分类器来检验 Boosting算法的性能。Drucker选择 Quinlan的 C4.5算法作为弱学习器对 NIST数据库中的样本进行分类。Freund, Schapire选择 C4.5, FindARrTest和 FindDecRule三个弱学习器对 UCI的 21个数据库中的样本进行了分类。Paul Viola和 Michael Jones采用了矩形特征分类器,实现了一个快速的人脸检测系统,取得了最好的分类结果和最快的分类速度。这也是目前为止, Adaboost算法在实际应用中最成功的范例之一。

本文采用的 Adaboost算法,是经过 Paul Viola和 Michael Jones变形之后的 Adaboost算法,其算法步骤如下:

1) 给定训练样本 $(X_1, Y_1), \cdots, (X_m, Y_m)$,其中 $\{0, 1\}$ 代表是反样本还是正样本;

2) 初始化权值 $W_{1,i} = 1/2m, 1/2l$,对应着 $Y_i = 0, 1$,其中 m, n为反、正样本个数;

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

权值归一化

对每一个特征 j,训练一个分类器 h_j,分类器的错误率用样本分布的权值 W_i来衡量:

$$E_j = \sum_i W_i \mid h_j(X_i) - Y_i \mid$$

选出一个错误率最低的分类器 h_t,其错误率为 E_t

收稿日期: 2005 - 04 - 19;**修订日期**: 2005 - 07 - 06

作者简介:赵万鹏(1981 -),男,山东莱芜人,硕士研究生,主要研究方向:模式识别、嵌入式系统; 古乐野(1960 -),男,重庆人,研究员,博士生导师,主要研究方向:嵌入式系统.

更新权值：$W_{t+1,i} = W_{t,i}\beta_t^{1-e_i}$

其中，当 h_t 对某样本分类正确时 $e_i = 0$，否则 $e_i = 1$。$\beta_t = E_t/1 - E_t$

最终的强分类器是：

$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T}\alpha_t h_t(x) \geq \frac{1}{2}\sum_{t=1}^{T}\alpha_t \\ 0 & \text{otherwis} \end{cases}$$

其中，$\alpha_t = \log\frac{1}{\beta_t}$

Adaboost算法相比其他机器学习算法，有很多优点。首先，寻找一个精度比随机预测略高的弱学习算法比寻找高精度的强学习算法要容易的多；其次，Adaboost算法不易过配，在训练误差达到零以后，算法仍能继续降低泛化误差，并没有出现由于迭代次数增加，而使泛化误差恶化的情况。

2 基于 Adaboost的手写体数字识别

2.1 基本原理

脱机手写体识别系统的基本原理如图 1。

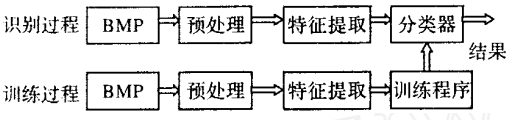

图 1 脱机手写体识别系统基本原理

2.2 预处理

本系统所做的预处理，是将大小不同的数字图像，缩放到 30×20像素的规范大小上。首先进行边界扫描，确定数字图像的实际大小，然后使用插值算法，将实际图像缩放到 30×20大小。这样做是为了适应数字图像大小不一的情况，在规范的大小上所做的分类器，有更理想的分类效果。

2.3 特征提取

特征提取是手写体数字识别系统的重要步骤，系统采用的是一种很简单的矩形特征。如图 2所示。

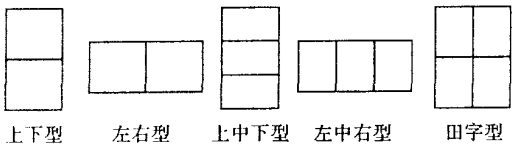

图 2 特征提取的矩形特征

总共有 5种形态的矩形特征。特征的特征值是这样计算的：1) 上下型特征的特征值 $V = N1 - N2$，其中 $N1$、$N2$分别为上、下矩形框内被污染像素的个数。2) 左右型特征的特征值 $V = N1 - N2$，其中 $N1$、$N2$分别为左、右矩形框内被污染像素的个数。3) 左中右型特征的特征值 $V = N1 + N3 - N2$，其中 $N1$、$N2$、$N3$分别为左、中、右矩形框内被污染像素的个数。4) 上中下型特征的特征值 $V = N1 + N3 - N2$，其中 $N1$、$N2$、$N3$分别为上、中、下矩形框内被污染像素的个数。5) 田字形特征的特征值 $V = N1 + N4 - N2 - N3$，其中 $N1$、$N2$、$N3$、$N4$分别为左上、右上、左下、右下矩形框内被污染像素的个数。

每种形态的矩形特征，随着起始点、宽度、高度的不同，会产生不同的特征。我们对在数字图像上的特征结构可以如下定义：

```
Struct aFeature
{
 CPoint StartPoint;                    //特征起始点
 Int Height;                           //高度
 Int Width;                            //宽度
 Int type;                             //五种类型之一
}
```

所以，在一个 30×20的数字图像上，可以取到许许多多个矩形特征，这些矩形特征提供了足够多的信息来表征这个数字图像。本系统取到的矩形特征是 13 932个。

特征的数量很多，计算特征值看起来是一个很大的工作量。然而，积分图像可以帮助我们解决这个问题。积分图像是图像的一种表示方式，其定义如下：

$$IImage(x, y) = \sum_{x' \leq x, y' \leq y} Image(x', y')$$

其中，$IImage(x, y)$ 是积分图像，$Image(x, y)$ 是原始图像。

计算积分图像是个效率很高的过程，只需对原始图像遍历一遍即可，其计算公式为：

$$S(x, y) = S(x, y - 1) + Image(x, y);$$

$$IImage(x, y) = IImage(x - 1, y) + S(x, y);$$

其中，$S(x, y)$ 是第 x列的上 y个像素中被污染像素的个数，且 $S(x, -1) = 0$，$IImage(-1, y) = 0$；

有了积分图像，计算矩形特征变的非常简单。假设一个矩形为 $Rect(left, top, right, down)$，可以用如下公式计算这个矩形框内被污染像素的个数：$NumOfBlackPoint = IImage(right, down) - IImage(left, down) - IImage(right, top) + IImage(left, top)$；计算矩形特征特征值的高效率使得在识别过程中的计算量很小，保证了足够快的识别速度。

2.4 分类器设计

前面介绍过本系统所采用的 Adaboost算法，这是一个两分类的学习算法，不能直接应用到手写体数字识别中。所以，本系统将数字识别的十分类问题转化为一系列的两分类问题来实现。最后形成的分类器包含第一层次上的 10个二分类器以及第二层次上的 45个二分类器。分类器的层次如图 3所示。

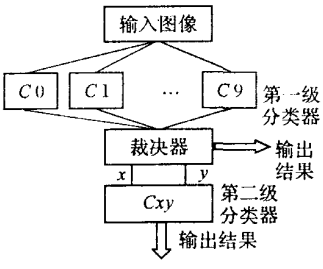

图 3 分类器的层次图

第一级分类器 Ci的作用是做一个“是 i”和“非 i”的二分类，分类器输出的结果 $h[0, 1]$，也就是说 h表征了输入图像对第 i类的隶属度。裁决器的作用是对第一级分类器的输出结果进行判决，如果输入图像对某个类别的隶属度特别高，则直接输出结果，认为输入图像隶属于该类别；如果没有一个类别的隶属度明显超出其他类别，裁决器选出两个隶属度最高的类别 x和 y，在第二级分类器中使用 Cxy再进行一次二分类，将输出结果作为最后的识别结果。

二分类器的训练，需要选取一种精度比随机预测略高的弱学习算法，本系统采用的弱分类器如下所示：

$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) < p_j\theta_j \\ 0 & \text{otherwise} \end{cases}$$

其中 p_j为方向系数，f_j为第 j个特征的特征值，θ_j为阈值。

每一个弱分类器的训练过程，实际上就是针对一个特征，找出相应的阈值，使得对训练样本的错误率最低。Adaboost的每一轮迭代，都可以在当前的样本分布上训练出 N个弱分类器（每个特征训练一个，N为特征数），然后从这 N个弱分类器

（下转第 2417页）

1) 当环境固定,规划结果如图 3所示。规划路径避开两个固定障碍物,到达目标点停止。

图 3　固定环境下的路径规划

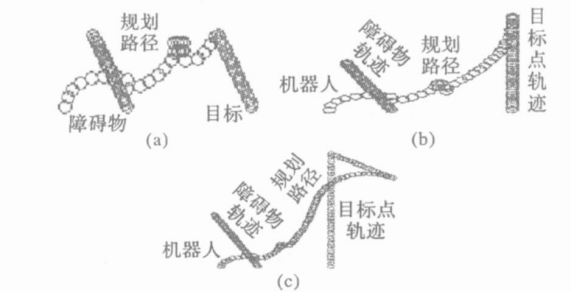

图 4　单个运动障碍物与运动目标的路径规划

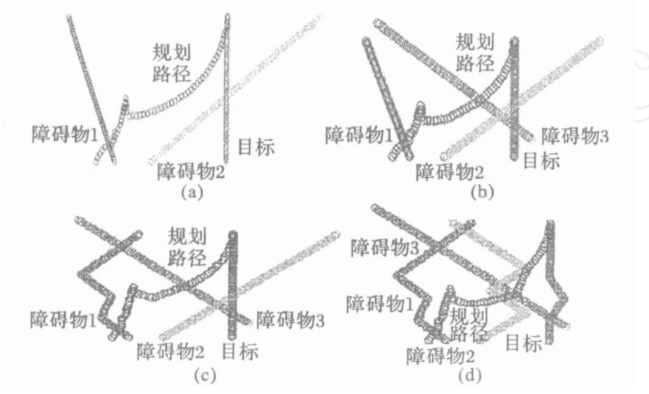

图 5　多个运动障碍物和运动目标的路径规划

2) 在存在单个运动障碍物与运动目标的环境下,使用改进势场法规划的结果如图 4所示。其中,图 (a)和 (b)是目标和障碍物分别往不同的方向运动的规划情况。图 (c)是当目标在运动过程中突然发生转向的情况。由图可以看出,使用本文所述的方法进行路径规划的时候,规划结果能够很好的跟踪目标的运动。

3)在目标点运动同时存在多个运动障碍物的环境下,实验结果如图 5所示。图 (a)、(b)、(c)分别是多个障碍物往不同方向运动的规划结果,图 (d)是在障碍物与目标点发生不规则运动情况下的规划结果。规划结果表明所规划的路径能够让机器人正确的避开环境中运动的障碍物,自动跟踪运动的目标点。同时从 (d)中可以看出,在目标点和障碍物运动规律完全未知的情况下,仍然能够进行正确的路径规划。

参考文献:

[1] 王耀南. 机器人智能控制工程 [M]. 北京:科学出版社, 2004.

[2] VADAKKEPAT P, TAN KC, WANG M-L. Evolutionary Artificial Potential Fields and Their Application in Real Time Robot Path[A]. Evolutionary Computation, 2000. Proceedings of the 2000 Congress on[C], 2000. 256 - 263.

[3] GE SS, CUI YJ. New Potential Functions for Mobile Robot Path Planning[J]. IEEE Transactions on robotics and automation, 2000, 16(5): 615 - 620.

[4] BARRAQUAND J, LANGLOIS B, Latom be J-C. Numerical Potential Field Techniques for Robot Path Planning[J]. IEEE Transactions on Systems, Manand Cybernetics, 1992, 22(2): 224 - 241.

[5] TARASSENKO L, BLAKE A. Analogue compution of collision-free paths[A]. Proceedings of the 1991 IEEE International Conference on Robotics and Automation[C]. 540 - 544.

[6] SUH S-H, SHIN KG. A Variational Dynamic Programming Approach to Robot-path Planning with a Distance-safety Criterion[J]. IEEE Journal of Robotics and Automation, 1998, 4(3): 334 - 349.

[7] PARK MG, LEE MC. Artificial Potential Field Based Path Planning for Mobile Robots Using a Virtual Obstacle Concept[J]. IEEE/ASME International conference on Advanced Intelligent Mechatronics(AIM), 2003: 735 - 740.

[8] 邹细勇,诸静. 一种考虑安全的移动机器人矢量场路径规划算法[J]. 中国机械工程, 2003, 14(14): 1205 - 1208.

(上接第 2414页)

中,挑出分类效果最好的一个,作为本轮选出的弱分类器。如此迭代 T次后,就得到了 T个弱分类器,然后将 T个弱分类器按权值组合成一个强分类器,完成了二分类器的训练。

2.4　实验结果

实验采用了 5 000个手写体数字图像作为训练样本。在训练程序的编制上,使用了一种快速的弱分类器训练方法,使得训练时间大大减少。在 PⅣ1.7G/256M的 PC机上,第一层每个二分类器的训练时间约为 3小时,第二层每个二分类器的训练时间约为 10分钟,总共训练时间为 40小时,训练正确率为 99.5%。

测试样本是另外 5 000个书写体数字图像,测试样本的书写者与训练样本的书写者不同,但由于选取的样本是按照"考试报名系统"的书写要求书写的数字,所以采集的测试样本较归整 (如图 4),测试正确率为 99.3%,测试错误如下所示。

图 4　实验结果

数字	0	1	2	3	4	5	6	7	8	9
错误个数	1	0	2	5	3	7	6	2	5	4

与其他训练方法产生的分类器相比,本实验的分类器的识别时间较快,单个数字的识别时间 < 3ms,可以满足实时的要求。

3　结语

本文提出一种将多分类问题转换为一系列二分类问题的方案,基于 Adaboost实现了手写体数字识别系统,取得了理想的识别效果。实践证明,Adaboost方法可以产生高精度的强分类器,并且具有较强的泛化能力。但是,其对训练样本中噪声的干扰比较敏感,训练的最后阶段往往将注意力集中到样本的噪声上面,这就对样本的选取提出了更高的要求。

参考文献:

[1] 于玲,吴铁军. 集成学习:Boosting算法综述 [J]. 模式识别与人工智能, 2004, 17(1): 52 - 59.

[2] 王海川,张丽明. 一种新的 Adaboost训练算法 [J]. 复旦学报 (自然科学版), 2004, 43(1): 27 - 32.

[3] FREUND Y, SCHAPIRE RE. Experiment With A New Boosting Algorithm[A]. Proc of the 13th International Conference On Machine Learning, San Francisco, CA[C], 1996. 148 - 156.

[4] VIOLA P, JONES M. Robust Real-time Object Dectection[A]. 8th IEEE International Co-nference On Computer Vision[C]. USA: IEEE Computer Society Press, 2001.