文章编号:1001 - 9081(2005)12 - 2925 - 03

基于 HMM 的分类器集成方法在脱机手写大写金额识别中的应用

王先梅,杨 扬,王 宏
(北京科技大学 电子信息工程系,北京 100083)
(Plum-wang@ tom. com)

摘 要:以脱机手写大写金额为研究对象,对其分别提取归一化小波特征、笔划密度特征和黑像素百分比特征,在分别形成独立的 HMM 分类器的基础上,将其按照一定的规则进行集成。实验结果表明,该方法能有效提高系统的识别率。

关键词:无约束手写大写金额识别; HMM; 分类器集成
中图分类号: TP319.43 **文献标识码**:A

Off-line handwritten amount in words recognition using classifiers combination method based on HMM

WANG Xian-mei, YANG Yang, WANG Hong
(*Department of Electronic Information Engineering, University of Science and Technology, Beijing* 100083, *China*)

Abstract: A new combination scheme was proposed. First, three individual classifiers were constructed with normalized wavelet, stroke density and the percentage of the number of black pixels. Then the combination based on different methods was studied. The experiment results show that the proposed approach can achieve high performance.

Key words: unconstrained off-line handwritten amount in words; HMM; classifier combination

0 引言

无约束脱机手写体汉字识别是一个非常复杂的多模式识别问题,为了提高识别率,人们从不同角度对各种特征和分类方法进行了广泛的研究。多年的实践表明,每种分类方法都有自己的优缺点和不同的适用范围,因此仅靠单独的分类器进行分类无法达到很高的识别率和可靠度。多特征融合和多分类器集成是进一步提高识别性能的必由之路,它可以弥补每个方案的不足之处,使识别率得到显著的提高[1,3,4]。从目前的情况来看,尽管分类器集成的方法很多,但在应用到具体应用领域时还需要根据实际情况选择适当的集成方案。

HMM 是一种对时序变化性信号进行处理的概率模型,其模型可以通过对大量的训练样本学习得到,因此对信号变化的适应能力强。本文以票据中的脱机手写大写金额为研究对象,对其提取基于小波特征、笔划密度特征和黑像素百分比特征,在 HMM 框架下分别进行参数训练,得到独立的 HMM 参数,分类时将不同的 HMM 分类器进行集成,取得了较为满意的识别效果。

1 系统模型

本文采用的系统是票据识别系统的一部分,用来识别各类票据中的大写金额,是基于 HMM 的手写汉字识别系统。整个系统包括预处理、序列特征提取与量化、训练和识别四个部分。其中预处理包括图像定位、去边框、二值化、切分、去噪、平滑、归一化处理等;特征提取从不同的角度提出了小波、笔划密度、黑像素百分比特征三种特征;量化时采用经典的 K-means 算法;训练时分别采用 Baum-Welch[2] 算法建立独立的隐马尔可夫模型;单分类器识别时采用 Viterbi[2] 算法进行;分类器集成时基于投票法按照一定的规则将三个分类器进行有机融合。

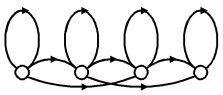

图 1 带跨越的 2 阶模型结构

票据系统中的大写金额识别是一个小类别识别系统,因此本文对每类识别单元建立独立的模型。尽管 2-D HMM 能够对汉字进行更精确的描述,但是基于是否有成熟算法、计算复杂度、样本数量以及识别速度等方面的要求,本文采取了左右型带跨越的 2 阶 1-D 离散隐马尔可夫模型,具体结构如图 1 所示。每一个状态除了可以自循环与向后一相邻状态转移外,还可实现状态的隔位转移,这样可以较好地模拟书写过程中发生的笔划冗余与丢失情况。

2 特征提取

2.1 基于多尺度小波变换的特征提取

二维正交小波变换分量具有方向选择性,可将原始图像分解成平滑子图、水平子图、垂直子图和斜向子图。所以,无须经过复杂的笔画方向特征提取,就可以得到图像字符的结构特征分量图。标准的小波分解过程采用金字塔算法,递归分解信号的低频部分,以生成下一尺度的各频带输出。图 2 表示了一个对图像进行 2 级正交小波分解的示意图,其中 W_{LL} 表示分解后的平滑子图,W_{LH} 表示水平方向子图,W_{HL} 表示垂直方向子图,W_{HH} 表示斜方向子图。

2 级小波分解后得到的信号的维数依然是比较高的。为了降低维数,取每一个尺度下各子图像的 $l1$ 范数作为特征

收稿日期:2005 - 06 - 24;**修订日期**:2005 - 09 - 03

作者简介:王先梅(1974 -),女,山东烟台人,讲师,博士研究生,主要研究方向:模式识别、虚拟现实; 杨扬(1955 -),男,河北承德人,教授,博士生导师,主要研究方向:模式识别、图像处理、多媒体通信; 王宏(1973 -),男,湖南衡阳人,博士研究生,讲师,主要研究方向:模式识别、图像处理.

值。具体计算公式如式(1) 所示:

$$e = \frac{1}{M \times N}\sum_{m=1}^{M}\sum_{n=1}^{N}|x(m,n)| \quad (1)$$

其中 $M \times N$ 为频带图像的大小,$x(m,n)$ 为该频带的小波系数。从图 2 可知,对于 K 级小波分解,特征矢量的维数为 $3K+1$。

2.2 笔划密度特征

笔划密度特征就是指以固定扫描次数沿水平、垂直或对角线方向对汉字进行扫描,计算扫描过程中扫描线与笔划相交的次数。对于二值图像而言,如果扫描线上相邻图像像素的值发生了变化,则认为有一个笔划边缘穿过。

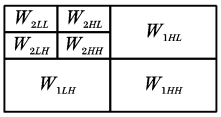

图 2 图像的金字塔小波分解

本文分别沿 0° 和 90° 两个方向对图像进行扫描,从而得到水平方向和垂直方向的笔划密度特征。

2.3 黑像素百分比特征

所谓的黑像素百分比特征就是求某个区域范围内黑像素的数目占整个图像黑像素总数的百分比。其计算公式如(2),(3) 所示:

$$weight = \sum_{m=1}^{M'}\sum_{n=1}^{N'}x(m,n) \quad (2)$$

$$sumt = \sum_{m=1}^{M}\sum_{n=1}^{N}x(m,n) \quad (3)$$

$M' \times N'$ 表示所求区域图像的大小,$M \times N$ 表示整个图像的大小。

2.4 特征归一化

对于小波特征而言,由于图像各频带的 $l1$ 范数取值范围不同,如果简单地将其组合,则绝对值小的特征的作用有可能会被淹没,因此一般需要进行特征归一化。本文采用基于均值和方差的归一化方法[5] 具体计算公式如式(4) ~ (6) 所示。

$$m(i) = \frac{1}{S_N}\sum_{j=1}^{S_N}x_j(i) \quad (4)$$

$$\sigma(i) = \sqrt{\frac{1}{S_N}\sum_{j=1}^{S_N}[x_j^2(i)] - [m(i)]^2} \quad (5)$$

$$\overline{x_j(i)} = \frac{x_j(i) - m(i)}{\sigma(i)} \quad (6)$$

各参数的含义为:

S_N:训练样本集中某一类汉字的所有子图像样本数;

$x_j = [x_j(1), x_j(2), \cdots, x_j(i), \cdots, x_j(D)]$:第 j 个子图像的特征矢量,其中 $j = 1,2,\cdots,S_N$,$i = 1,2,\cdots,D$,D 表示特征维数;

$m(i)$:子图像第 i 维特征的均值;

$\sigma(i)$:第 i 维特征矢量的方差;

$\overline{x_j(i)}$:第 j 个子图像归一化后的第 i 维特征。

3 分类器集成

分类器集成从体系结构上分为并行集成与串行集成两种。所谓并行是指各个识别子系统都独立地接受原始图像并给出自己的识别结果,而后在相互独立的识别结果基础上得到最终答案,主要的方法有投票法、贝叶斯方法以及神经网络合成等。而串行方法则是将前一级识别子系统的结果作为后一级的输入,其关系相对于并行方法要复杂得多。考虑到计算的简单性,本文采用了投票法进行基于 HMM 的分类器集成[3,4],其集成结构如图 3 所示。

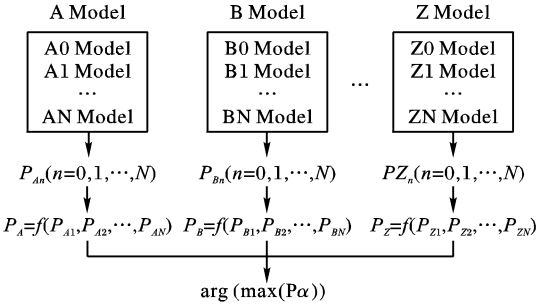

图 3 投票法分类器集成方法

(1)集成分类器构造问题描述

假设训练过程中已经生成 N 个独立的分类器,每个分类器产生过程中生成了 Z 个不同汉字的模型。对于给定的未知样本 x,每一独立分类器 K 分别输出其与第 α 个模型的相似度 $P_{\alpha k}(k = 1,\cdots,N)$,那么计算该未知模式属于某类汉字 α 的可能性 P_α 的过程是:将样本 x 分别与该类汉字已知的 N 个模型进行比较,将输出结果 $P_{\alpha n}(n = 1,2,\cdots,N)$ 按照某种准则 $f(P_{\alpha 1},P_{\alpha 2},\cdots,P_{\alpha N})$ 进行集成,从而得到待识别样本与该类别的新相似性度量 P_α。然后按照某种判决规则将 x 归入相应的类别中去。该集成方法的性能与集成函数 $f(P_{\alpha 1},P_{\alpha 2},\cdots,P_{\alpha N})$ 的选择有关。

(2) 集成函数

最大值规则:

$$f(P_{\alpha 1},P_{\alpha 2},\cdots,P_{\alpha N}) = \max(P_{\alpha 1},P_{\alpha 2},\cdots,P_{\alpha N})$$

平均值规则:

$$f(P_{\alpha 1},P_{\alpha 2},\cdots,P_{\alpha N}) = \frac{1}{N}\sum_{i=1}^{N}P_{\alpha i}$$

加权平均规则:

$$f(P_{\alpha 1},P_{\alpha 2},\cdots,P_{\alpha N}) = \frac{1}{N}\sum_{i=1}^{N}C_iP_{\alpha i}(C_i \text{ 表示加权值})$$

(3) 判决准则

若 $P_k = \max(P_\alpha)(\alpha = A,B..Z)$,则 $x \in W_k$

4 实验结果与分析

本文自行收集了零、壹、贰…. 玖、拾、元等 12 个字符的 11373 个样本,对手写方式没有限制。其中用于训练的样本数为 7773 个,测试样本为 3600 个。

在提取序列化特征的过程中,采用了分区滑动窗口技术。将每个图像沿水平方向分成 8 个滑动窗口,每个滑动窗口又沿垂直方向分成 4 个子区域,将每个子区域的特征按顺序连接起来就形成了该滑动窗口的特征矢量。

本文为每类汉字建立独立的模型,因此需要建立 12 个模型。系统中的状态数和码本大小根据经验设定为 $N = 9$,$M = 64$。模型结构采用如图 2 所示的允许隔位跨越的离散隐马尔科夫模型。

为了检验不同集成方法的性能,本文将三个分类器的输出结果分别按照所述的三种方法进行集成。实验结果如表 1 所示。加权平均集成的性能与权值有关,实验中将各个独立分类器对测试样本集的识别率作为权值。

从表 1 中可以看出,虽然分类器有不同的集成方法,但是通过分类器集成技术,各种不同机理的分类器得到了有效的综合,集成分类器的性能较各子分类器都有显著的提高。同时还可以看出,在本文的应用背景下,平均值集成和加权平均

集成的性能要优于最大值集成。

表 1　不同分类器及按不同规则集成的实验结果

分类方法	识别性能	
	训练集	测试集
归一化小波	94.61	92.11
笔划密度	95.45	93.88
黑像素百分比	91.73	88.08
最大值集成	96.67	95.69
平均值集成	98.88	98.06
加权平均集成	98.89	98.08

5　结语

模式识别中的多分类器集成方法得到了越来越多的关注和应用。多分类器集成的关键是根据应用的背景,寻找一种合适的组合准则。

本文基于 HMM 模型,在构建基于归一化小波特征、笔划密度特征以及黑像素百分比特征的独立分类器基础上,通过多分类器集成技术,将三种分类器进行有效集成。本文所采用的多种集成方法,都保持了比较高的识别率。集成后的系统显著改善了系统整体的识别能力。

参考文献:

[1] 朱小燕, 史一凡, 马少平. 手写体字符识别研究[J]. 模式识别与人工智能, 2000, 13(2): 174 - 180.

[2] RABINER LR. A Tutorial on Hidden Markov Models and Select Applications in Speech Recognition[J]. IEEE, 1989, 77(2): 257 - 286.

[3] NISHIMURA H, KOBAYASHI M, MARUYAMA M, *et al*. Off-line character recognition using HMM by multiple directional feature extraction and voting with bagging algorithm[A]. Fifth International Conference on Document Analysis and Recognition[C]. Bangalore, 1999. 49 - 52.

[4] WANG WW, BRAKENSIEK A, RIGOLL G. Combining HMM-based two-pass classifiers for off-line word recognition[A]. Proceedings 16th International Conference on Pattern Recognition[C]. Quebec, 2002. 151 - 154.

[5] FAN GL, XIA XG. Wavelet-Based Texture Analysis and Synthesis Using Hidden Markov Models[J]. IEEE Transaction on Circuits and System, 2003, 50(1): 106 - 119.

(上接第 2918 页)

模拟实验(1)、(2)比较了当输入流量是指数分布的 on/off 流时,在没有 Hash 调整的情况下,$\zeta_{max} = 1.41 > 1$,表明部分时刻已经出现外部链路过载现象,节点域的负载均衡性能很差,进行 Hash 调整时则没有出现链路过载现象。模拟实验(3)、(4)比较了当输入流量是常数速率流时,进行 Hash 表调整的负载均衡机制效果明显,在同一时刻,不同外部链路的 ζ 值更接近。

表 1　实验结果

实验序号	实验情景	ζ_{max}	$\Delta_{max}\zeta$	$\overline{\sigma_\zeta^i}$
实验(1)	指数流量分布,没有 Hash 表调整	1.41	0.85	0.19
实验(2)	指数流量分布,进行 Hash 表调整	0.91	0.18	0.049
实验(3)	常数速率流量分布,没有 Hash 表调整	0.575	0.23	0.05
实验(5)	指数流量分布,调整间隔 $\tau = 0.1$s	0.91	0.18	0.04919
实验(6)	指数流量分布,调整间隔 $\tau = 0.5$s	0.96	0.415	0.079
实验(7)	指数流量分布,调整间隔 $\tau = 1.0$s	1.22	0.46	0.11
实验(8)	指数流量分布,判定阈值 $\Delta = 0.01$	0.72	0.24	0.052
实验(9)	指数流量分布,判定阈值 $\Delta = 0.05$	0.80	0.45	0.088
实验(10)	指数流量分布,判定阈值 $\Delta = 0.1$	0.80	0.67	0.18

可以看出,执行 Hash 表调整的负载均衡机制可以有效地改善网络性能。由于 Internet 流量的突发性,DDLBM 中 Hash 表调整是必要的。

模拟实验(5)、(6)、(7)比较了 Hash 表调整时间间隔 τ 不同对负载均衡性能的影响。随着调整时间间隔的增加,同一时刻链路的 ζ 差异增加,表明负载均衡效果减弱。调整时间间隔太小,负载均衡依赖负载的即时速率进行负载均衡操作,当流量具有较高的突发特征时,即时速度变化幅度大、变化频率快可能使得 Hash 表调整频繁,容易造成流在逻辑链路内不同外部链路之间频繁切换,最终导致流内包的乱序。调整时间间隔太长,负载均衡依赖负载的平均速率进行负载均衡操作,负载均衡依赖的流量速率估计过于乐观,可能导致链路过载。确定 Hash 表调整时间间隔需要在负载均衡效果与调整的频率之间折中。

模拟实验(8)、(9)、(10)比较了负载均衡算法中判定是否需要 Hash 表调整的阈值发生变化时对负载均衡性能的影响。随着判定阈值 Δ 的增大,同一时刻链路的 ζ 差异增加,表明负载均衡效果减弱。判定阈值 Δ 的大小影响 Hash 表调整的频次,进而影响负载均衡算法的效果。可以根据负载均衡的需要,灵活地设置阈值 Δ。

5　结语

给出了一个通用的域间负载均衡模型,并以此为基础提出一种基于 Hash 表的分布式动态调整的域间负载均衡机制 DDLBM,算法详细描述了 Hash 表调整的方法,计算需要调整的链路以及调整的力度。算法在域的各个入口处独立地进行负载的分配,依赖局部的流量历史信息调整分配到各外部链路的流量,没有消息的传递和全局的协同,适合应用于大型的高速网络。模拟实验揭示了采用 Hash 表调整的算法负载均衡效果明显优于没有 Hash 表调整的算法。

参考文献:

[1] MOY J. OSPF Version 2, RFC 2328[S]. 1998.

[2] ISO 10589. Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473) (Also republished as RFC 1142)[S].

[3] HEDRICK CL. An introduction to igrp[R]. Technical report, Rutgers University, August 1991.

[4] REKHTER Y, LI T. A Border Gateway Protocol 4(BGP-4), RFC 1771[S]. 1995.

[5] THALER D, HOPPS C. Multipath Issues in Unicast and Multicast, RFC 2991[S]. 2000.

[6] HOPPS C. Analysis of an Equal-Cost Multi-Path Algorithm, RFC 2992[S]. 2000.

[7] CAO ZR, WANG Z, ZEGURA E. Performance of Hashing-Based Schemes for Internet Load Balancing[A]. Proceedings of IEEE Infocom[C]. 2000. 332 - 341.

[8] UCB/LBNL/VINT Network Simulator-NS (version 2)[EB/OL]. http://www-mash.cs.berkerley.edu/ns, 2005 - 05.