基于并行组合分类器的脱机手写体数字识别

楚浩宇,高 萌,刘永生
(东北农业大学 电气与信息学院,黑龙江 哈尔滨 150030)

摘 要:为了提高脱机手写体数字识别的识别率和可靠性,并且考虑到传统的单一分类器对数字之间差异的敏感性不同,综合 K-近邻算法、广义回归神经网络、支持向量机三种机器学习算法,提出了一种并行组织结构的组合分类器。并行组合分类器通过改进的投票机制来判定识别结果。以 MNIST 数据库为数据来源,在 MATLAB 平台上开展各种分类器的性能对比实验。组合后的识别率、拒识率、误识率、可靠性分别可达到 97.48%、1.55%、0.97%、99.02%。实验结果表明,并行组合分类器在鲁棒性方面优于传统的单一分类器,在识别率、拒识率、算法的时间复杂度上均优于其他组合分类器。并行组合分类器以简易结构实现了脱机手写体数字的快速、高效识别。

关键词:模式识别;组合分类器;LR;广义回归神经网络;支持向量机;手写体数字
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1673-629X(2018)03-0105-04
doi:10.3969/j.issn.1673-629X.2018.03.022

Off-line Handwritten Digit Recognition Based on Parallel Combined Classifiers

CHU Hao-yu,GAO Meng,LIU Yong-sheng
(School of Electrical and Information,Northeast Agricultural University,Harbin 150030,China)

Abstract:In order to improve the recognition rate and reliability of off-line handwritten digit recognition,considering that traditional single classifiers have different sensitivity to the differences between digital,we propose a combined classifier of parallel organizational structure combining three machine learning algorithms of K-nearest neighbor,general regression neural network and support vector machine. It uses the improved voting mechanism to determine the recognition result. Using MNIST database as data source,the comparable experiment on the performance of classifiers is carried out on MATLAB,whose results (recognition rate,rejection rate,false accept rate,reliability) are 97.48%,1.55%,0.97% and 99.02%. The experiments indicate that the parallel combined classifier is superior to the traditional single classifier in terms of robustness,and other combined classifiers in terms of recognition rate,rejection rate and time complexity. With a simple structure,it can achieve fast and efficient off-line handwritten digit recognition.

Key words:pattern recognition;combined classifiers;LR;GRNN;SVM;handwritten digit

0 引 言

手写体数字识别是光学字符识别的一个重要分支,分为联机手写体数字识别和脱机手写体数字识别。在联机手写体数字识别中,计算机可以通过与之相连的输入设备得到关于笔尖运动轨迹和速度的有效信息,所以识别相对较易[1]。由于数字图像的数据量庞大且书写风格的迥异等干扰因素对识别会产生很大的影响,因此脱机手写体数字识别难度较大,但其应用领域更加宽泛。因此这是一项意义重大的研究课题。

鉴于传统的单一分类器对数字之间差异的敏感性不同,许多学者开始研究组合分类器所产生的效果[2]。文献[3]使用四种特征和三种传统分类器构造了九种不同的分类器进行组合。文献[4]构造了两级的组合分类器,第一级是最小距离分类器,第二级由三个反向传播网络并联而成。文献[5]提出了一种基于量子神经网络的二级识别系统。这些方法虽然在一定程度上提高了识别率与可靠性,但分类器的组合结构却十分复杂,因此识别速度随之下降。

文中提出使用一种特征、三种分类器、并行结构组织的组合分类器,与传统方法相比,在提高识别率与可靠性的同时,极大地减少了算法的时间复杂度。

收稿日期:2017-04-17 **修回日期**:2017-08-22 **网络出版时间**:2017-12-05
基金项目:国家"863"高技术发展计划项目(2013AA10230304)
作者简介:楚浩宇(1996-),男,研究方向为机器学习与人工智能;高 萌,讲师,博士,研究方向为智能信息处理。

1 手写体数字识别的基本原理

手写体数字识别一般包括图像预处理、特征提取、分类识别等模块[6],其结构如图 1 所示。

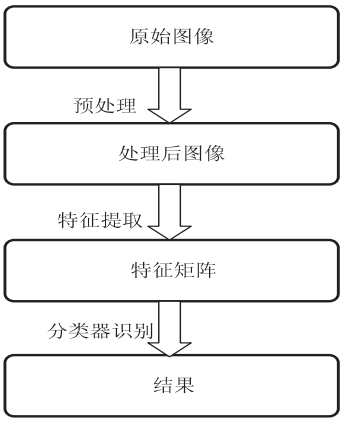

图 1 数字识别的步骤

1.1 图像的预处理

预处理的主要目的是去除字符图像中的噪声,并采用灰度化方法处理读入的图像,图像中的每个像素就对应唯一一个灰度值,得到规范化的点阵,为识别做好准备[7]。

1.2 特征提取

数字图像用元素为灰度值的矩阵表示,直接用维数很高的矩阵进行计算无论是时间复杂度还是空间复杂度都很大,因此需要进行特征提取。一般对预处理后的图像进行统计特征提取,常用的有主成分分析、独立成分分析和 Fisher 线性鉴别分析[8]。

1.3 识别方法

分类器主要分为基于概率分布的分类器,例如逻辑回归分类器[9];基于距离的分类器,例如最近邻分类器[10]; 人工神经网络分类器,例如 BP 神经网络分类器;支持向量机分类器[11]。近年来,也有学者提出应用分类识别的伯努利隐马尔可夫模型与基于活动图的半监督学习模型[12-13]。

2 并行结构组合分类器的设计

为了能够从不同的角度观察不同数字之间变化的规律,综合考虑各种单一分类器的优缺点,同时不能由于结构太过复杂化而导致训练分类器的时间和综合决策时间过长。对此,文中采用 K-近邻算法、广义回归神经网络及支持向量机,提出一种基于改进的投票机制的并行组合分类器,实现脱机手写体数字的快速、高效识别。

2.1 单一分类器的分类原理

2.1.1 K-近邻(KNN)分类器

K-近邻算法是最近邻算法的一个推广,样本类别的判断是由距离最近的 K 个样本投票来决定[10]。给定训练样本集 $S=\{x_1,x_2,\cdots,x_n\}$,假设需要识别的样本为 $\boldsymbol{x}$。计算出与 S 中每个样本的距离,并寻找出与 $\boldsymbol{x}$ 距离最近的前 K 个样本,则 K-近邻算法分类规则为:如果 $j=\underset{1\leqslant i\leqslant c}{\operatorname{argmax}}k_i$,则判别 $\boldsymbol{x}\in w_j$。其中,k_i 是与 $\boldsymbol{x}$ 距离最近的 K 个样本中属于 w_i 类的样本数。

2.1.2 广义回归神经网络(GRNN)分类器

广义回归神经网络是径向基神经网络的一种[14]。假设随机变量 $\boldsymbol{x}$ 和随机变量 y 的联合概率密函数为 $f(x,y)$。设 $\boldsymbol{X}$ 是随机变量 $\boldsymbol{x}$ 的测量值,则 y 在给定 $\boldsymbol{X}$ 下的条件均值为:

$$E[y\mid\boldsymbol{X}]=\frac{\int_{-\infty}^{\infty}yf(\boldsymbol{X},y)\,\mathrm{d}y}{\int_{-\infty}^{\infty}f(\boldsymbol{X},y)\,\mathrm{d}y} \tag{1}$$

基于随机变量 $\boldsymbol{x}$ 和 y 的样本值 X^i 和 Y^i 的概率估计 $\hat{f}(\boldsymbol{X},Y)$ 为:

$$\hat{f}(\boldsymbol{X},Y)=\frac{1}{(2\pi)^{(p+1)/2}\sigma^{(p+1)}}\cdot\frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{(\boldsymbol{X}-\boldsymbol{X}^i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}^i)}{2\sigma^2}\right)\cdot\exp\left(-\frac{(Y-Y^i)^2}{2\sigma^2}\right) \tag{2}$$

其中,n 为样本观测值的个数;p 为随机变量 $\boldsymbol{x}$ 的维数;σ 为高斯函数的宽度系数。

将式(2)中的 $\hat{f}$ 代入式(1),交换积分和求和的次序,可得网络的输出 $\hat{Y}(\boldsymbol{X})$ 为:

$$\hat{Y}(\boldsymbol{X})=\frac{\sum_{i=1}^{n}Y^i\exp\left(-\frac{(\boldsymbol{X}-\boldsymbol{X}^i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}^i)}{2\sigma^2}\right)}{\sum_{i=1}^{n}\exp\left(-\frac{(\boldsymbol{X}-\boldsymbol{X}^i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}^i)}{2\sigma^2}\right)} \tag{3}$$

2.1.3 支持向量机(SVM)分类器

支持向量机是一种建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的模式识别方法。支持向量机的主要思想是通过某种事先选择的非线性映射将输入向量 x 映射到高维特征空间,在这个空间中构造最优分类超平面[11]。

对于给定的训练集 $(x^{(i)},y^{(i)})$,为了寻找最优分类超平面,支持向量机需要求解以下二次规划问题:

$$\max_{\alpha}W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_jK(x^{(i)},x^{(j)})$$

$$0\leqslant\alpha_i\leqslant C,i=1,2,\cdots,m$$

$$\sum_{i=1}^{m}\alpha_iy^{(i)}=0 \tag{4}$$

其中,m 为训练样本数;α_i 为拉格朗日乘子;C 为惩罚系数;$(x^{(i)},x^{(j)})$ 为核函数。

给定测试数据 $\boldsymbol{x}$,通过式(5) 的值来确定标签 y。

$$f(\boldsymbol{x}) = \sum_{\alpha_i = 0} K(x^{\langle i \rangle}, \boldsymbol{x}) \tag{5}$$

2.2　组合分类器的工作原理及流程

2.2.1　改进的投票机制

分类器识别状况的权值矩阵为:

$$\boldsymbol{W} = \begin{bmatrix} w_{00} & w_{01} & \cdots & w_{09} \\ w_{10} & w_{11} & \cdots & w_{19} \\ \vdots & \vdots & \vdots & \vdots \\ w_{n0} & w_{n1} & \cdots & w_{n9} \end{bmatrix} \tag{6}$$

其中, w_{ij} 为第 $i+1$ 个分类器在识别数字 j 时设定的权值。

权值矩阵中每个元素的取值为:

$$w_{ij} = \begin{cases} \omega, j = c \\ 0, j \neq c \end{cases} \tag{7}$$

其中, ω 为该分类器的预设定权值; c 为该分类器的识别结果。

在识别数字时的判定依据为:

$$R_j = \sum_{i=0}^{n} w_{ij}, j = 0, 1, \cdots, 9 \tag{8}$$

给定拒识阈值 t ,若满足 $R_j < t$ 则拒绝识别,否则输出识别的数字。

2.2.2　工作流程

三个分类器,即 KNN 分类器、GRNN 分类器、SVM 分类器分别简称为 C_1、C_2、C_3。

(1) C_1、C_2、C_3 分别在训练集中进行学习并进行参数寻优,得出各分类器的参数;

(2)将各分类器的识别结果与测试集中的结果进行比对,根据正确率排出各分类器的可信度的排名;

(3)收集 C_1、C_2、C_3 识别数字的结果,设定矩阵中各元素的取值范围;

$$w_{0c} = \begin{cases} 1, c = j \\ 0, c \neq j \end{cases} \tag{9}$$

$$w_{1c} = \begin{cases} 1, c = j \\ 0, c \neq j \end{cases} \tag{10}$$

$$w_{2c} = \begin{cases} 1.5, c = j \\ 0, c \neq j \end{cases} \tag{11}$$

(4)列出三行十列的混淆矩阵;

$$\boldsymbol{w} = \begin{bmatrix} w_{00} & w_{01} & \cdots & w_{09} \\ w_{10} & w_{11} & \cdots & w_{19} \\ w_{20} & w_{21} & \cdots & w_{29} \end{bmatrix} \tag{12}$$

(5)根据式(13)计算 R_j 的值;

$$R_j = \sum_{i=0}^{2} w_{ij}, j = 0, 1, \cdots, 9 \tag{13}$$

(6)根据各分类器的可靠性预设拒识阈值 $t = 2.5$,若满足 $R_j < 2.5$ 拒绝识别,否则输出识别的数字。

3　实　验

3.1　数据来源

文中应用的实验数据来源于 MNIST 数据库,这是一个广泛使用于各种图像处理系统和机器学习领域的大型手写体数字数据库。该数据库包含了 6 万组训练图片和 1 万组测试图片,其中每张图片已经经过预处理压缩至 784(28 * 28)像素[15]。特征矩阵即是由灰度值作为元素的 28 * 28 维矩阵。图 2 是 MNIST 数据库中的部分训练数据。

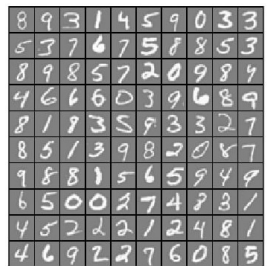

图 2　MNIST 训练集的部分数据

3.2　性能指标

分类器的水平由性能指标来评价,主要的性能指标如下:

(1)识别率(recognition rate)。

$$\text{Recognition Rate} = \frac{\text{The number of samples correctly recognized}}{\text{Total sample size}} \times 100\% \tag{14}$$

(2)误识率(false accept rate)。

$$\text{False Accept Rate} = \frac{\text{The number of samples falsely recognized}}{\text{Total sample size}} \times 100\% \tag{15}$$

(3)可靠性(reliability)。

$$\text{Reliability} = \frac{\text{Recognition Rate}}{\text{Recognition Rate} + \text{False Accept Rate}} \times 100\% \tag{16}$$

(4)拒识率(rejection rate)。

多数情况下,在输入待测样本之后,分类器会给出对应数字的分类结果。但是对于某些特定领域,分类结果发生错误可能会带来极其严重的后果,因此需要对没有把握的样本拒绝识别,由此降低误识率,提高可靠性。

$$\text{RejectionRate} = \frac{\text{The number of samples rejected to recognize}}{\text{Total sample size}} \times 100\% \tag{17}$$

3.3　实验结果及分析

评价一个分类器性能的优劣主要看其是否具有低

误识率、低拒识率和高识别率、高可靠性。实验结果如表 1 所示。

表 1 分类器性能对比实验结果 %

分类器	识别率	拒识率	误识率	可靠性
KNN 分类器	96.65	0	3.35	96.65
GRNN 分类器	97.00	0	3.00	97.00
SVM 分类器	98.58	0	1.42	98.58
文献[3]的结果	97.05	2.05	0.90	99.08
文献[4]的结果	88.60	9.80	2.30	97.50
QNNs 多级分类器[5]	96.50	2.30	1.20	98.80
并行组合分类器	97.48	1.55	0.97	99.02

比较表 1 中单一分类器和组合分类器的结果,单一分类器的拒识率这一指标全部为 0。组合分类器由于有些数字样本因手写不谨慎而与别的数字产生混淆的原因,才拒绝识别了这一小部分样本,所以组合分类器存在单一分类器所没有的“噪音”过滤功能。从表 1 中可以看出,并行组合分类器的误识率要比单一分类器低 0.45% ~2.38%,而且可靠性要高于单一分类器 0.5% ~2.43%,所以其鲁棒性优于单一分类器。

文中提出的并行组合分类器在识别率、拒识率、识别算法的时间复杂度上均优于文献[3-5]中所采用的组合分类器。由于文中选取了识别率较高的单一分类器、改进了传统的投票机制,这使得只有极少测试样本会因同时在各种分类器中识别不佳导致权值之和达不到预设的阈值而被拒识,因此组合分类器具有高识别率、低拒识率。

文献[3]需要事先训练九种分类器再进行组合判断;文献[4]中组合分类器第二级采用了三个并联 BP 神经网络,虽然 BP 神经网络在数学上已经被证明具有实现任何复杂非线性映射的功能,但收敛速度十分缓慢;文献[5]使用的 QNNs 多级分类器包含十四个子网络;而文中采用的三个单一分类器均具有较快的训练速度且并行组合分类器的结构简单,因此在识别算法的时间复杂度上小于以上文献的同时,又取得了高识别率、低拒识率。

并行组合分类器在误识率和可靠性方面高于文献[4-5],较之文献[3]略低。考虑到文献[3]中九个组合分类器极大地增加了训练时间和综合决策时间,而且其较高的拒识率也在一定程度上减少了误识率、增加了可靠性,而文中提出的并行组合分类器在牺牲较少时间,拒识较少样本的情况下与其只有不到 0.1% 的差距,所以并行组合分类器在总体性能上要优于其他组合分类器。

4 结束语

提出了一种并行结构的组合分类器,通过改进的投票机制得出最终的结果,实验结果表明,对单一分类器进行组合,在保证了低误识率、低拒识率、高识别率、高可靠性的前提下,利用简易结构即可实现脱机手写体数字的快速、高效识别。组合分类器的鲁棒性比单一分类器的要强,而且组合分类器有着较强的灵活性和可拓展性,找到更好的组织结构以及判断机制将是今后研究的重点。手写体数字识别是字符识别中的一个研究方向,提出的组合分类器起到了抛砖引玉的作用,绝不只是应用于数字识别,稍加改变,便可应用于其他字符识别。

参考文献:

[1] KHERALLAH M,HADDAD L,ALIMI A M,et al. On-line handwritten digit recognition based on trajectory and velocity modeling[J]. Pattern Recognition Letters,2008,29(5):580-594.

[2] 柳回春,马树元,吴平东,等. 手写体数字识别技术的研究[J]. 计算机工程,2003,29(4):24-25.

[3] 胡钟山,娄 震,杨静宇,等. 基于多分类器组合的手写体数字识别[J]. 计算机学报,1999,22(4):369-374.

[4] 傅德胜,谢忠红,苏 坚. 基于组合分类器的自由手写体数字识别方法[J]. 计算机工程与设计,2004,25(10):1713-1715.

[5] 吴茹石,彭 力. 基于量子神经网络的手写体数字识别方法研究[J]. 计算机工程与设计,2007,28(18):4462-4465.

[6] 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.

[7] 张 猛,余仲秋,姚绍文. 手写体数字识别中图像预处理的研究[J]. 微计算机信息,2006,22(6-1):256-258.

[8] 杨 健,杨静宇,叶 晖. Fisher 线性鉴别分析的理论研究及其应用[J]. 自动化学报,2003,29(4):481-493.

[9] BISHOP C M. Pattern recognition and machine learning[M]. New York:Springer-Verlag,2006:205-206.

[10] 刘家峰,赵 巍,朱海龙,等. 模式识别[M]. 哈尔滨:哈尔滨工业大学出版社,2014:15-17.

[11] VAPNIK V N. The nature of statistical learning theory[M]. New York:Springer-Verlag,2000:133-141.

[12] GIMERNEZ A,ANDRES-FERRER J,JUAN A. Discriminative Bernoulli HMMs for isolated handwritten word[J]. Pattern Recognition Letters,2014,35:157-168.

[13] CECOTTI H. Active graph based semi-supervised learning using image matching:application to handwritten digit recognition[J]. Pattern Recognition Letters,2016,73:76-82.

[14] SPECHT D F. A general regression neural network[J]. IEEE Transactions on Neural Networks,1991,2(6):568-576.

[15] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.