基于 SVM 的手写体数字快速识别方法研究

李 琼[1],陈 利[2],王维虎[1]
(1. 汉口学院 信息科学与技术学院,湖北 武汉 430212;
2. 汉口学院 实验中心,湖北 武汉 430212)

摘 要:手写体数字识别是图像处理与模式识别中具有较高实用价值的研究热点之一。在保证较高识别精度的前提下,为提高手写体数字的识别速度,提出了一种基于 SVM 的快速手写体数字识别方法。该方法通过各类别在特征空间中的可分性强度确定 SVM 最优核参数,快速训练出 SVM 分类器对手写体数字进行分类识别。由于可分性强度的计算是一个简单的迭代过程,所需时间远小于传统参数优化方法中训练相应 SVM 分类器所需时间,故参数确定时间被大大缩减,训练速度得到相应提高,从而加快了手写体数字的识别过程,同时保证了较好的分类准确率。通过对 MNIST 手写体数字库的实验验证,结果表明该算法是可行有效的。

关键词:手写体数字识别;支持向量机;核参数;可分性强度
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2014)02-0205-04
doi:10.3969/j.issn.1673-629X.2014.02.051

Research on Method of Fast Handwritten Digits Recognition Based on SVM

LI Qiong[1],CHEN Li[2],WANG Wei-hu[1]
(1. School of Information Science and Technology,Hankou University,Wuhan 430212,China;
2. Dept. of Experiment Center,Hankou University,Wuhan 430212,China)

Abstract:Handwritten digits recognition has high practical value in the field of image processing and pattern recognition. In order to improve the recognition speed,at the premise of high recognition accuracy,a fast handwritten digits recognition method based on SVM is proposed. The new method which uses the separability measure between classes in the feature space to choose the best kernel parameters,can train SVM classifiers fast to recognize the handwritten digits. Due to the computation of separability measure is a simple iterative process,the time required for computing is far less than the time required for training SVM classifiers in traditional parameter optimization methods. Thus,the time for kernel parameters selection will be reduced greatly. Accordingly,the training speed will be increased,and so that the process of recognizing handwritten digits will also be speeded up,while ensuring better classification accuracy. The experiment results of testing MNIST show that the improved algorithm is feasible and effective.

Key words:handwritten digits recognition;support vector machine;kernel parameter;separability measure

0 引 言

手写体数字识别是光学字符识别技术(Optical Character Recognition,OCR)的一个分支,它的研究目标是让计算机模拟人自动识别纸张上的手写体阿拉伯数字[1]。目前手写体数字识别技术广泛应用于邮政编码、财务报表、统计报表、银行票据等方面,是图像处理和模式识别领域的一个研究热点。传统的手写体数字识别技术如人工分类、神经网络、决策树等识别方法普遍存在识别速度较低、识别正确率不高等问题,因此提出了一种基于支持向量机(Support Vector Machine,SVM)的快速手写体数字识别方法,并在 MNIST 数据库上进行了实验验证。

结果表明,该方法加快了 SVM 分类器的训练过程,提高了手写体数字的识别速度,并得到了较好的识别正确率。

收稿日期:2013-04-29 **修回日期**:2013-08-03 **网络出版时间**:2013-11-29
基金项目:2012 年湖北省教育科学技术研究计划指导性项目(B20128103)
作者简介:李 琼(1981-),女,湖北天门人,讲师,硕士,研究方向为中文信息处理、计算机应用;陈 利,教授,研究方向为中文信息处理、计算机应用。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20131129.0912.030.html

1 手写体数字的识别原理

手写体数字识别综合了图像处理、模式识别、机器学习等多个领域的知识,是一个跨学科的复杂问题,其识别系统通常由图像预处理、特征提取以及分类识别三部分组成,如图 1 所示。

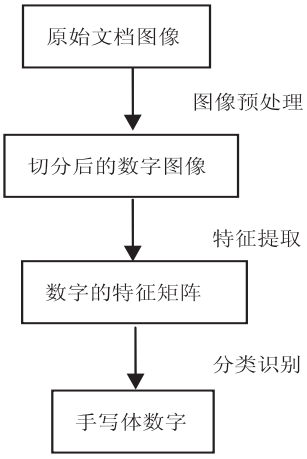

图 1 手写体数字识别原理图

1.1 图像预处理

手写体数字识别的第一步是图像预处理。通常,待识别的手写体数字图像在扫描过程中,常会带来一些噪声,用不同的扫描分辨率得到的数字图像,其质量也各不相同。另外,还需要正确分割整幅文档图像中的手写体数字,而分割后的数字大小、字体常各不相同,故还需进行归一化处理。所以,在图像预处理过程中需要解决的主要问题有:图像二值化、平滑化(去噪)、字符切分、规范化等。

图像预处理在整个手写体数字识别中,占据十分重要的地位。预处理做得好,能将反映文字本质特征的部分得以保留,使得后期识别相对容易,识别正确率高,识别速度快;反之,将会导致后期识别变得困难,或是出现误识、拒识等不良后果。

1.2 特征提取

特征提取的目的是从图像预处理后的数字图像中,提取出用来区分其他数字类别的本质属性并数值化,形成特征矢量的过程。常见的手写体数字特征有:统计特征和结构特征。

统计特征是在二值或灰度值点阵图像的基础上,对数字图像点阵进行数学变换后提取的特征。常用的数学变换有小波变换、样条曲线拟合、矩、傅立叶描绘子等。统计特征常与神经网络分类器或统计匹配方法结合使用。

结构特征主要描述手写体数字的几何结构,侧重于体现数字结构的本质特征。该方法区分相似字的能力较强,可以得到识别正确率较高的分类结果,但易受到噪声等因素的干扰。

通常,采用单一的特征提取方法可以利用的手写体数字信息量有限,这样不可避免地存在一些识别的死角,即存在利用该特征向量难以区分的手写体数字。因此,现在普遍采用将多种特征提取方法相结合,取长补短的方法来获取最优特征向量。

1.3 分类识别

分类识别是利用训练出的分类器,对特征提取后的手写体数字进行分类识别。分类器的识别原理是通过其拓扑结构和内置参数定义了特征空间上的一组曲面或超曲面,利用这组曲面或超曲面将特征空间划分为不同的区域,从而达到分类识别的目的。目前常用的手写体数字分类器有:

(1)基于距离的分类器:该分类器先从训练样本集中计算出类中心,再通过不同的距离测度,计算出相应参数矩阵。分类识别时,计算待识别数字样本与各类别中心点的距离,取距离最小者对应的数字类别为识别结果。距离分类器中常用的距离度量方式有欧式距离、城市块距离、加权距离及马氏距离等。基于距离的分类器常用作粗分类。

(2)人工神经网络分类器:是按照人脑的组织及活动原理,构造的一种数据驱动型非线性模型。它由神经元结构模型、网络连接模型、网络学习算法等几个基本要素构成,是具有某些智能功能的系统。人工神经网络分类器具有自适应功能、泛化功能及非线性映射功能等特点,但训练时间往往过长,且收敛速度及泛化能力也有待深入研究。常见人工神经网络分类器有 RBF 网络及 BP 网络等。

(3)支持向量机分类器:是根据 Vapnik 提出的结构风险最小化原则,通过最大化分类间隔或边缘,来提高分类性能的一种分类器。SVM 分类器通过选择训练一组称为支持向量的特征子集,使得对支持向量集的线性划分等同于对整个数据集的分割,实现了降低运算复杂度的同时,保证了分类识别的精度[2-4]。由于 SVM 分类器具有很强的小规模细分能力,故文中采用 SVM 分类器来自动识别手写体数字,并做了改进,采用各类别在特征空间中的可分性强度来决定最优核参数,加快了训练过程,提高了手写体数字的识别速度,并保证了良好的识别准确率。

2 基于 SVM 的快速手写体数字识别方法

手写体数字属于多类分类问题(共 10 类),然而最初的 SVM 只能解决两类别的分类问题,为此,常用两种解决方法[5-6]:一种是将多个分类面的参数求解合并在一个大的最优化问题中,通过求解该最优化问题“一次性”地解决多类分类问题。另一种是将多个两类 SVM 分类器组合在一起实现多类分类。第一种

方法在求解最优化问题的过程中,由于变量太多,严重影响了运算时间,故一般采用第二种方法来解决多类分类问题。第二种方法中常用的有 1-a-r(1-against-rest)方法、1-a-1(1-against-1)方法、DAG-SVM 有向无环图方法及二叉树 SVM 方法[7-9]。文中采用 1-a-1 多类分类方法来解决手写体数字的分类识别问题。1-a-1 多类分类方法的分类原理是在每两类间训练一个两值 SVM 分类器,这样将得到 $k(k-1)/2$ 个两值分类器(假设总类别数是 k,$k\geqslant 2$)。当对一个未知数字样本进行分类时,每个分类器都对其类别进行判断,并采用"投票法"为其相应的类别投票,最后将得票最多的类别作为该未知数字样本的类别。

由于 SVM 是一种基于核的机器学习方法,在分类识别前还需解决 SVM 核函数和核参数的确定问题[10-11]。这将直接影响到训练出的 SVM 分类器的分类性能。三种常用的 SVM 核函数中,采用高斯核函数的 SVM 分类器表现出很强的学习能力,故文中采用高斯函数作为 SVM 分类器的核函数。确定了核函数类型后,就面临着核参数的选择问题。目前,最常用的核参数选择方法是网格搜索法,该方法通过估计核参数(C,σ)的变化范围和变化粒度,将其划分为二维网格,网格中的每个节点作为一组候选核参数,然后利用交叉验证方法获取各组核参数的验证精度,选择验证精度最高的那组核参数作为最优核参数。该方法的缺点是:由于模型参数的可选范围很大,在训练样本集较大时优化过程非常慢,达不到实际需求中的实时处理要求[12-13]。

为了提高优化速度,文中提出采用高维特征空间 Hilbert 中,各类别间的可分性强度来确定最优核参数,由于不用训练相应参数的 SVM 分类模型,故节省了训练时间,加快了训练过程,提高了手写体数字的识别速度。

假设输入样本空间的两个手写体数字样本 x_1,x_2,被某一非线性映射函数 φ 映射到高维 Hilbert 特征空间中,得到 $\varphi(x_1)$,$\varphi(x_2)$,则输入样本空间中的点积在高维 Hilbert 特征空间中,可以用 Mercer 核表示为:

$$K(x_1,x_2)=\varphi(x_1)\cdot\varphi(x_2)$$

x_1,x_2 在特征空间中的欧氏距离为:

$$d^{\mathrm{H}}(x_1,x_2)=\sqrt{K(x_1,x_1)-2K(x_1,x_2)+K(x_2,x_2)}$$

类中心点 m_φ 在特征空间中变为:

$$m_\varphi=\frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)$$

其中,n 为类中样本的总数。

设两类训练样本 $\{x_1,x_2,\cdots,x_{n_1}\}$ 和 $\{y_1,y_2,\cdots,y_{n_2}\}$,被 φ 映射到特征空间中后,两类的中心点分别是 m_φ 和 m_φ',则特征空间中 m_φ 和 m_φ' 之间的欧式距离为:

$$d^{\mathrm{H}}(m_\varphi,m_\varphi')=\sqrt{\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}K(x_i,x_j)-\frac{2}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}K(x_i,y_j)+\frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}K(y_i,y_j)}$$

相应训练样本 x 到类中心 m_φ 的距离为:

$$d^{\mathrm{H}}(x,m_\varphi)=\sqrt{K(x,x)-\frac{2}{n}\sum_{i=1}^{n}K(x,x_i)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i,x_j)}$$

则:特征空间中类 i 和类 j 之间的可分性强度定义为:

$$\mathrm{sm}_{ij}^{\mathrm{H}}=\frac{d^{\mathrm{H}}(m_\varphi^i,m_\varphi^j)}{\sigma_i^{\mathrm{H}}+\sigma_j^{\mathrm{H}}}$$

其中,$\sigma^{\mathrm{H}}=\max d^{\mathrm{H}}(x_i,m_\varphi)$ 是类中各样本点到相应中心点之间的距离的最大值,代表各类别的紧凑程度。

设训练样本集为 L,验证样本集为 M,测试样本集为 N,总类别数为 K($K\geqslant 2$),快速优化核参数方法步骤如下:

Step1:预先给定一组核参数候选值,对于每个核参数候选值 σ,在训练样本集 L 上,计算上述定义中的 sm^{H};

Step2:按 sm^{H} 值由小到大的顺序排序,取使 sm^{H} 值最大的核参数为最终的最优核参数 σ_{Best};

Step3:给定一组惩罚因子值,对每个 C,将它与 σ_{Best} 组合,训练出相应的 SVM 分类器;

Step4:将 Step3 得到的所有 SVM 分类器,在验证样本集 M 上进行验证,取使验证精度最大的 SVM 分类器对应的(C_{Best},σ_{Best})为最优核参数组合;

Step5:采用 Step4 得到的最优 SVM 分类器,对测试样本集 N 进行分类识别。

3　实验结果与分析

实验数据来源于 MNIST 数据库,这是美国国家标准和技术研究所为满足学术界对基准数据的要求而提供的一个专门用于手写体数字识别研究的数据库(可从 http://yann.lecun.com/exdb/mnist/处下载)。从中任意选取1 000个训练图像作为训练样本,再任取1 000个测试图像作为测试数据,其中所有字符是用 20*20=400 像素空间中的向量来表示。实验采用 5-折交叉验证法获取验证精度,采用 SMO 训练算法和一对一分类算法解决多类分类问题。实验中采用的相关软件可从 http://www.csie.ntu.edu.tw/~cjlin/libsvm/处下载[14]。改进算法采用高斯核函数,预先设定的核参

数搜索范围为：$C \in [2^{-1},2^{0},\cdots,2^{7}]$，$\sigma \in [2^{-15},2^{-14},\cdots,2^{0}]$。

实验采用网格搜索法和文中算法做了对比验证，具体结果如表 1。

表 1 实验结果

	核参数(C,σ)	分类精度/%	训练时间/s
网格搜索法	$(2^{3},2^{-14})$	94.63	267
改进算法	$(2^{2.5},2^{-12})$	93.21	73

实验结果表明，采用改进算法优化核参数，在保证分类精度维持较高水平的同时，训练时间明显优于网格搜索法。这是因为该方法所需训练时间由计算各类别 sm^{H} 的时间及将 σ_{Best} 和每个候选 C 组合来训练 SVM 分类器的时间两部分组成。与传统的网格搜索法相比，计算 sm^{H} 所需时间，远远少于训练相应 SVM 分类器所需时间，故核参数的优化时间被大大缩短，训练速度得到显著提高，相应手写体数字的分类识别速度也得到明显提高。

4 结束语

文中首先简要介绍了手写体数字识别的一般过程，然后在现有识别技术的基础上，提出了一种基于 SVM 的快速手写体数字识别方法。该方法通过计算各类别在特征空间中的可分性强度来选择最优核参数组合，与传统的网格搜索法相比，由于可分性强度的计算是一个简单且不需迭代的过程，故可以显著缩短训练时间，加快训练过程，从而快速识别出手写体数字。最后，通过对 MNIST 手写体数字库的实验验证可知，该算法是可行有效的。

文中下一步要做的工作是如何预设合理的核参数搜索范围来保证在尽可能短的时间内搜索到最优核参数。

参考文献：

[1] 边肇祺，张学工. 模式识别[M]. 北京：清华大学出版社，2000.

[2] Vapnik V N. 统计学习理论[M]. 许建华，张学工，译. 北京：电子工业出版社，2004.

[3] Astorino A，Gorgone E，Gaudioso M，et al. Data preprocessing in semi－supervised SVM classification[J]. Optimization，2011，60(1－2)：143－151.

[4] Hu Guyu，Gong Yong，Chen Yande，et al. Semi－supervised radio transmitter classification based on elastic sparsity regularized SVM[J]. Journal of electronics (China)，2012，29(6)：501－508.

[5] 李文趋. SVM 在手写数字识别中的应用[J]. 泉州师范学院学报(自然科学)，2010，28(4)：18－21.

[6] 石会芳，胡小兵，刘瑞杰，等. 基于启发式 GA－SVM 的手写数字字符识别的研究[J]. 计算机技术与发展，2012，22(10)：5－9.

[7] 蒋 华，戚玉顺. 基于球结构 SVM 的多标签分类[J]. 计算机工程，2013，39(1)：294－297.

[8] 刘端阳，邱卫杰. 基于 SVM 期望间隔的多标签分类的主动学习[J]. 计算机科学，2011，38(4)：230－232.

[9] Hsu C W，Lin C J. Acomparison of methods for multi－class support vector machines[J]. IEEE transactions on neural networks，2002(13)：415－425.

[10] 奉国和. SVM 分类核函数及参数选择比较[J]. 计算机工程与应用，2011，47(3)：123－124.

[11] 王 佳，徐蔚鸿. 基于动量粒子群的混合核 SVM 参数优化方法[J]. 计算机应用，2011，31(2)：501－503.

[12] 陈圣兵，王晓峰. 基于样本差异度的 SVM 训练样本缩减算法[J]. 计算机工程与应用，2012，48(7)：20－22.

[13] 王 涛，程良伦. 基于快速 SVM 的大规模网络流量分类方法[J]. 计算机应用研究，2012，29(6)：2301－2305.

[14] Chang C C，Lin C J. LIBSVM：A library for support vector machines[EB/OL]. 2001[2013－03－04]. http://www.csie.ntu.edu.tw/～cjlin/papers/libsvm.pdf.

(上接第 204 页)

同机制的分析[J]. 计算机应用研究，2011，28(6)：2006－2010.

[3] Deng H，Varanasi M，Swigger K，et al. Design of sensor－embedded radio frequency identification (SE－RFID) systems [C]//Proc of IEEE international conference on mechatronics and automation.[s.l.]：[s.n.]，2006.

[4] 刘福铭. RFID 与无线传感器网络集成技术研究与开发[D]. 上海：上海交通大学，2007.

[5] 朱嵘涛，徐爱钧，陈 超. 基于 ZigBee 的无线位移测量系统的设计[J]. 石油仪器，2011，25(2)：1－3.

[6] 皮桂英，张维波，王春霞. 浅谈无线传感器网络技术[J]. 仪表技术，2010(7)：72－73.

[7] Sung J，Lopez T S，Kim D. The EPC sensor network for RFID and WSN integration infrastructure[C]//Proc of the 5th annual IEEE international conference on pervasive computing and communication.[s.l.]：[s.n.]，2007.

[8] 李 杰. 物联网中无线传感器节点和 RFID 数据融合的方法[J]. 电子设计工程，2011，19(7)：103－106.

[9] 唐承佩，唐焯宜. 一种基于 RFID 与 WSN 融合的异构网络识别平台[J]. 制造业自动化，2011，33(3)：31－33.

[10] 薛丽莹，王 健. WSN 节点定位算法与 RFID 在食品安全监测中的应用[J]. 森林工程，2012，28(4)：89－92.

[11] 莫 西，吴云洁. RFID 与 WSN 融合技术中软硬件接口的研究与实现[J]. 微型电脑应用，2012，28(7)：5－10.

[12] 曲涛涛，阎 芳. RFID 与 WSN 技术融合理论研究[J]. 物联网技术，2013(3)：30－34.

基于SVM的手写体数字快速识别方法研究

作者： 李琼， 陈利， 王维虎， LI Qiong， CHEN Li， WANG Wei-hu

作者单位： 李琼,王维虎,LI Qiong,WANG Wei-hu(汉口学院 信息科学与技术学院,湖北 武汉,430212)
， 陈利,CHEN Li(汉口学院 实验中心,湖北 武汉,430212)

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(2)