基于启发式 GA-SVM 的手写数字字符识别的研究

石会芳,胡小兵,刘瑞杰,叶剑英
(重庆大学 数学与统计学院,重庆 401331)

摘 要:数字字符识别技术是图像处理领域中的一个重要研究方向,并且在社会经济生活的许多方面有着越来越广泛的应用。预处理是手写数字识别中重要的一个环节,它的优劣直接关系到识别算法的性能。文中提出了一种改进后的数字字符识别处理方案。该方案首先对输入的数字字符图像进行预处理,然后使用大津法进行全局图像阈值选取,通过阈值处理将灰度图像转化为二值图像。最后利用 libsvm 加强工具箱和启发式遗传算法进行手写数字的识别。实验结果显示该方法提高了识别数字的准确度。

关键词:数字字符识别;阈值处理;支持向量机;遗传算法
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2012)10-0005-05

Research on Handwritten Numeral Character Recognition Based on Heuristic GA-SVM

SHI Hui-fang, HU Xiao-bing, LIU Rui-jie, YE Jian-ying
(College of Mathematics and Statistics, Chongqing University, Chongqing 401331, China)

Abstract: The technology of numeral character recognition is an important research direction in image processing area and the application in various fields is more and more widely. Pretreatment is a vital step in recognition of handwritten numerals, its strengths and weaknesses is directly related to the performance of the recognition algorithm. It describes an improved treating plan of numeral character recognition. Firstly, the method is to do the pretreatment for the input numeral character image, and then Otsu method is used to finish the global image threshold selection, through threshold processing shift grayscale image into binary image. Finally, use libsvm enhance toolbox and the method of genetic algorithm to discriminate digit. Experimental results show that the accuracy of the numeral recognition is improved.

Key words: numeral character recognition; threshold processing; support vector machine; genetic algorithm

0 引 言

手写数字识别(Handwritten Numeral Recognition)是光学字符识别技术(Optical Character Recognition,简称 OCR)的一个分支,它研究的对象是:如何利用电子计算机自动辨认人手写在纸张上的阿拉伯数字。经过多年研究,已经开始向各种实际应用推广[1],如银行储蓄、邮政编码。

字符识别系统一般要分为预处理、特征抽取、判别、后处理等模块[1]。预处理是字符识别重要的一环,它把原始的图像转换成识别器所能接受的二进制形式。要识别手写体数字首先要对其字符图像进行预处理[1]。预处理通常包括数字图像的二值化处理、细化处理等步骤。数字图像的二值化处理是将上一步骤所得到的灰度数字图像转化为二值数字图像,即在数字图像中区分出字符和背景[2,3]。二值化处理方法很多,但根据现实生活中大量数字识别的需要,一般采用 Otsu(大津法) 方法进行阈值处理从而获得二值化数字图像。阈值选取是图像处理中的基本问题,选取的阈值直接影响计算机视觉的后续处理效果,在国内外有很多学者针对这一课题进行了研究,并且提出了很多阈值选取方法,其中 Otsu 方法是无监督、非参数的自适应阈值选取方法[4,5]。由于它不需要其他先验知识,利用图像中的灰度直方图,确定图像分割门限值的依据是目标与背景之间的方差最大而且是动态的,因

收稿日期:2012-02-12;**修回日期**:2012-05-20
基金项目:重庆市自然科学基金资助项目(CSPC,2005BB2197);重庆大学"211 工程"三期创新人才培养计划建设基金资助项目(S-09110)
作者简介:石会芳(1986-),女,河南开封人,硕士研究生,主要研究方向为支持向量机的研究与应用;胡小兵,副教授,博士,主要研究方向为现代优化算法、机器人控制技术和计算机软件设计。

此 Otsu 方法的应用范围非常广,目前仍然是最常用的图像分割方法之一。

遗传算法(Genetic Algorithm,简称 GA)是由美国 Holland 教授在 1975 年提出的一类借鉴生物界自然选择和自然遗传机制的随机搜索算法,较以往传统的搜索算法具有使用方便、鲁棒性强、便于并行处理等特点,广泛应用于各种领域[6,7]。其主要特点是不依赖于梯度信息的隐含并行随机群体搜索,可以在搜索过程中自动获取并积累有关搜索空间的知识,并且可以自适应地控制搜索进程,进而得到全局最优解[8]。

支持向量机(Support Vector Machine, SVM)是借助于最优化方法解决数据挖掘中若干问题的有力工具,它在一定程度上克服了"维数灾难"和"过学习"等传统困难,并在文本分类、生物信息、语音识别、遥感图像分析、故障识别和预测、时间序列预测、信息安全等诸多领域有了成功的应用[9]。

近年来 SVM 已经成为模式识别领域的研究热点,因此一些学者也开始把支持向量机技术应用到手写数字识别中[2],例如将遗传算法(GA)用于 SVM 分类识别中 SVM 参数优化和样本的特征选择已有学者进行了研究[10],并且取得了良好的应用效果。文中采用 Otsu 方法进行阈值处理以获得二值化数字图像,并且运用 libsvm 工具箱自带的有关支持向量机的函数,利用启发式遗传算法实现基于 RBF 的 SVM 参数优化来识别手写数字,这个方法提高了识别数字的准确度,并通过实验证明了其可行性。

1　图像处理与分析

以下介绍对灰度图像进行处理与分析的技术要点,即对灰度图像进行如下操作:用 Otsu 方法选择使黑白像素内方差最小的阈值,用设置阈值的方法(thresholding)将结果图像转换成二值图像。其算法步骤描述如下:

1)用 MATLAB 工具箱函数 imread( )读入一个图片。

2)对图像进行反色处理。

3)计算全局阈值,对图像进行阈值处理。先调用 graythresh,自动计算一个适当的阈值;然后调用 im2bw 执行阈值处理,将灰度图像转换成二值图像。

4)查找数字上所有像素点坐标。调用 find 函数,查找数字上所有像素点的行标 y 和列标 x。

5)截取包含完成数字的最小区域图像,并用 imresize 函数将截取的包含完整数字的最小区域图像转成 16 × 16 的标准化图像,转化后的图像是白色的字黑色的背景,图中数字区域像素值为 1,背景区域为 0。

6)生成训练样本矩阵。将标准化图像按列拉成一个向量后并转置,生成 50 × 256 的训练样本矩阵 training,并将第 k 个数字的训练图像存在 training{ k }中。

文中的改进之一是将图像处理中的整体阈值 0.4 改为 3)中的方法,经试验可得,这种阈值处理方法提高了手写数字识别的准确度。例如,对于训练图像"8",原始图像经过以上图像标准化处理可得图 1。

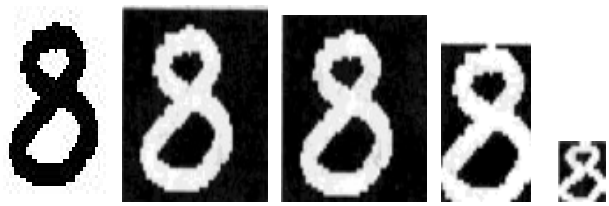

图 1　图像标准化处理示例

2　支持向量机

支持向量机是 Vapnik 教授等人在多年研究统计学理论基础上提出的一种用于解决线性不可分这样的分类问题的理论[11]。它不仅能解决两分类问题,而且能解决现实中经常面临的多分类问题,这里首先介绍两分类支持向量机,即 C-支持向量机,它的具体算法如下:

1)给定训练集。

$$T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in (R^n \times Y)^l \quad (1)$$

其中 $x_i \in R^n, y_i \in Y = \{1, -1\}, i = 1, \cdots, l$; x_i 为特征向量;

2)选取适当的核函数 $K(x, x')$ 以及惩罚参数 $C > 0$,其中 $K(x, x') = (\varphi(x), \varphi(x'))$,其中非线性映射 $\varphi(x)$ 为从 R^n 到 Hilbert 空间 H 的变换,将输入数据映射到高维空间;

3)构造并求解凸二次规划问题。

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=i}^{l} y_i y_j K(x_i, x_j) \alpha_i \alpha_j - \sum_{j=1}^{l} \alpha_j \quad (2)$$

$$\text{s.t.} \sum_{i=1}^{l} y_i \alpha_i = 0 \quad 0 \leqslant \alpha_i \leqslant C, i = 1, \cdots, l \quad (3)$$

得解 $\alpha^* = (\alpha_1^*, \cdots, \alpha_l^*)^{\mathrm{T}}$

4)计算 b^*:选取位于开区间 $(0, C)$ 中的 α^* 的分量 α_j^*,据此计算

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x_j) \quad (4)$$

5)构造决策函数。

$$f(x) = \operatorname{sgn}(g(x)) \quad (5)$$

其中

$$g(x) = \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x) + b^* \quad (6)$$

直接利用上述算法只能解决两分类问题,由于平常遇到的实际问题中,大多为多分类问题,求解决多分

类问题的一个途径是构造一系列两分类问题,并建立相应的两分类机,然后根据这些两分类机对输入 x 判定的结果推断 x 的属性,然而构造不同的两类分类问题导致不同的方法,主要有以下几种:成对分类(one versus one)、一类对余类(one versus the rest)、纠错输出编码等[12]。文中运用 libsvm 工具箱自带的有关支持向量机的函数进行手写数字识别,因此解决多分类问题采用成对分类算法,同时求解与训练集(1)对应的二次规划问题(2)、(3)的解决算法是序列最小最优化算法(SMO)。

支持向量分类机中核函数的意义很广泛,它蕴含着特征选择,并且蕴含着特征提取中的线性降维,因此对和函数的确定是一个很重要的问题。常用的核函数有很多种,如线性核函数(Linear kernel)、多项式核函数(Polynomial kernel)、sigmoid 核、二次核函数(Quadratic kernel)和 RBF 核等[13]。经过多年来国内外的研究表明,SVM 以 RBF 核为核函数具有很强的学习能力和分类效果。因此,文中选用径向基内积函数(RBF),也称为高斯核:

$$K(x,x') = \exp(-\| x - x' \|^2/2\sigma^2) \tag{7}$$

3　C 和 σ 参数的遗传算法选择

RBF 核 SVM 所包含的主要参数有误差惩罚因子 C 及宽度函数 σ。其中 C 和 σ 是两个可以人为调节的参数,参数取值不同,对应的分类器性质以及推广识别率也将有很大的差别。通过对参数空间中不同区域所对应的分类器性质的分析,有助于人们使用启发式的方法寻找最优参数,从而获得更高的推广识别率[2]。关于 SVM 参数的优化选取,目前在国际上还没有公认统一的最好的方法,由于过高的 C 会导致过学习状态发生,即训练集分类准确率很高而测试集分类准确率很低,所以在能够达到最高验证分类准确率中的所有的成对的参数 C 和 σ 中认为较小的惩罚参数 C 是更佳的选择对象。由于遗传算法的主要特点是不依赖于梯度信息的隐含,并且能够对群体进行随机搜索,从而得到全局最优解,因此文中利用遗传算法的全局搜索能力进行参数选取,从而取得最优参数。

3.1　验证性能指标的选取——k-折交叉验证

在利用支持向量机进行分类的情况下,为了优化参数 C 和 σ,通常应该将初始的含有 l 个样本点的总训练样本集,分为 k 个样本子集,其中这 k 个样本子集的交集为空集。分别取每一个样本子集作为新的验证样本集,并且在取每一个样本子集作为验证样本时,总训练样本集中剩余的其它样本作为新的训练样本。假设对第 i 个样本子集进行测试,可以得到正确分类的训练点个数为 l_i,那么在 k 次迭代完成后,便可以得到 $l_1, \cdots, l_k$,其中每个样本子集验证性能指标定义为:第 i 个样本集的正确分类数为 l_i,最后的验证性能指标为所有样本子集正确分类概率的平均值[14]。

最后得总的验证性能指标为:

$$\min_{C,\sigma} L = (l_1 + \cdots + l_k)/k \tag{8}$$

通过对式(1)的优化即可求得调整参数 C 和径向基函数的宽度 σ。其中设定惩罚因子和核参数的范围为:$C \in (0,100]$,$\sigma \in [0,1000]$。

3.2　遗传算法的设计过程

3.2.1　群体规模的选择过程

群体太大计算过程可能相当复杂,群体太小不容易求得满意的结果,因此合适的群体规模对遗传算法的收敛具有相当重要的意义。种群的大小应结合具体问题的计算复杂度来确定,一般选取区间为[20,100]。

文中的支持向量机参数对分类模型的正确率有很大影响,计算量大,故不适合建立大规模种群,因此,将种群规模定位为 20。

3.2.2　适应度函数的设计过程

在运用遗传算法来进行参数选择中,通常采用适应度函数值来评估个体性能进而指导搜索,在此过程中很少使用搜索空间的知识,因此适应度函数的选取在遗传算法的设计中占有相当重要的位置。所以利用上面所述的 k-折交叉确认作为验证性能的指标,选择如下所示的适应度函数:

$$f = (l_1 + \cdots + l_k)/l \times 100\% \tag{9}$$

其中,l_i 为测试样本集中第 i 个样本集的正确分类样本数目,l 为测试样本集总的样本数。文中在后面的试验中取 k 为 5,即选取 5 折交叉确认的平均误差作为参数性能的评价标准。

3.2.3　选择与复制

选择具有从上一代遗传结果中以一定的概率选择适应度较大的个体从而进入下一代的操作的作用。所以,适应度值越大的个体被选择的概率相对的也就会越大。其中个体的选择概率为:

$$P_i = f_i / \sum_{j=1}^{N} f_j \tag{10}$$

其中 f_i 表示个体 i 的适应度,N 表示种群规模。个体 i 被复制的个数为 $R(i) = N \times P_i$。经过选择和复制,从初始化种群中形成一个新的子群。

3.2.4　交叉和变异

交叉操作过程在遗传算法中属于最主要的遗传操作过程,由于参数基因采用的是实值编码,所以,为了保证交叉后产生新的参数值,并且能够生成新的搜索空间,参数基因的交叉操作过程采用线性组合的方式,通过将两个基因串对应交叉位的值相结合从而生成两个新的基因串,从而使这两个新的基因串含有其父代

的特征。

选择和交叉这两个操作过程基本上完成了遗传算法的大部分搜索功能,而变异则增加了遗传算法找到比较接近最优解的能力,所以,变异在遗传操作过程中属于辅助性的搜索操作,它的主要目的是保持群体的多样性。由于较低的变异概率可以在一定程度上防止群体中重要的单一基因的丢失,从而降低遗传算法开辟新搜索空间的能力;较高的变异概率将使遗传操作趋向单纯的随机搜索,从而大大地降低算法的稳定性和收敛速度。因此,要根据实际问题来选取变异概率,一般取值区间为[0.001,0.5]。

4　应用和比较

4.1　手写数字识别模型

文中选取参考文献[2]中的手写数字训练样本和测试样本,并且采集一些手写样本。其中选取1050幅手写数字图片作为训练样本,每个数字均有105幅图片;另外选取1520幅手写数字图片作为测试样本,每个数字对应152幅测试图片,每幅图片大小均为50×50像素。

文中采用libsvm加强工具箱,利用Matlab进行编程进而构造支持向量机,算法步骤如下所示[15]:

1)对输入数据按照第2部分所示的算法步骤进行图像处理与图像分析。

2)对运行参数进行设置。取进化最大代数为 $G=100$;种群规模数 $N=20$;变异概率 $Pm=0.01$;交叉概率 $Pc=0.9$[16]。

3)采用二进制编码,随机生成个染色体作为初始种群。

4)对每个染色体解码,并且按照式(9)计算每个染色体的适应度。

5)按其适应度选择、复制个体,从而生成新的种群。

6)对上述新的种群进行遗传操作过程。

7)判断结果是否满足停止准则或是否满足最大进化代数,如果满足则转到步骤8);如果不满足上述条件则返回步骤4)。

8)对每个染色体解码,构造支持向量机,并利用验证性能标准评价支持向量机性能。

9)利用训练好的支持向量机对手写数字进行识别。

其中利用遗传算法迭代50次,在 CV 意义下的最高分类准确率为94.2857%,训练样本的识别精度为100%,适应度函数曲线如图2所示。

4.2　比　较

在图像处理时阈值的选择中运用给定整体阈值法,设定整体阈值为0.4[11],此时训练样本的识别精度仍为100%,在 CV 意义下的最高分类准确率为93.3333%,适应度函数曲线如图3所示。这种方法与文中方法对比如表1所示,可以看出用大津法进行阈值处理与给定整体阈值法相比,测试结果非常理想,对各种情况的表现都较为良好。

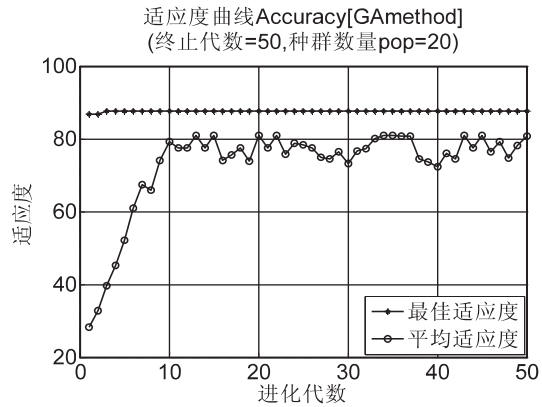

图2　给定整体阈值法基础上参数选择的适应度曲线

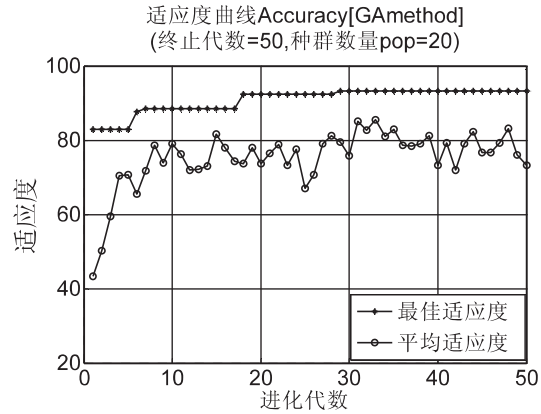

图3　大津法基础上参数选择的适应度曲线

表1　遗传算法与网格搜索测试结果比较

参数选择算法	阈值处理算法	训练样本数	测试样本数	C	σ	测试集识别率
遗传算法	大津法	1050	1520	3.4897	0.0165	97.3684%
遗传算法	给定整体阈值法	1050	1520	59.2507	0.0346	94.0789%
网格搜索	给定整体阈值法	1050	1520	1.0000	0.0359	94.0789%

在运用大津法进行阈值处理的基础上,应用网格搜索参数选择后的支持向量机进行手写数字分类,可得如图4和图5所示的SVC参数选择结果,并与遗传算法优选参数后的支持向量机分类结果相比,由表1可知利用遗传算法优选参数后的支持向量机的分类精度是比较高的[17]。

5　结束语

文中尝试了采用libsvm工具箱利用Matlab编程进行基于启发式GA-SVM的手写数字识别。首先用大津法选取出阈值,虽然它在很多情况下都不是最佳

的分割,但分割质量通常都有一定的保障,而且比给定整体阈值法测试集识别率高的多,可以说是最稳定的分割。因此,大津算法是一种较为通用的分割算法。并且在 Otsu 法图像阈值处理的基础上,利用遗传算法解决支持向量机的参数选择问题,与网格搜索法相比提高了手写数字的识别率。

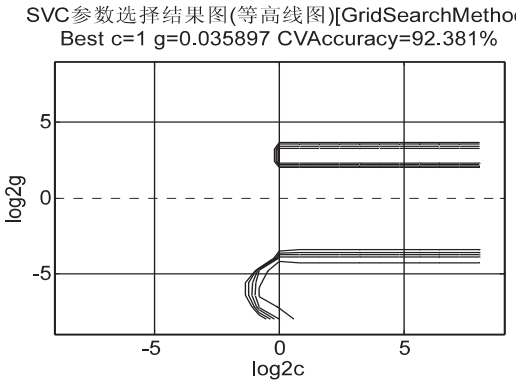

图 4　SVM 参数选择的等高线图

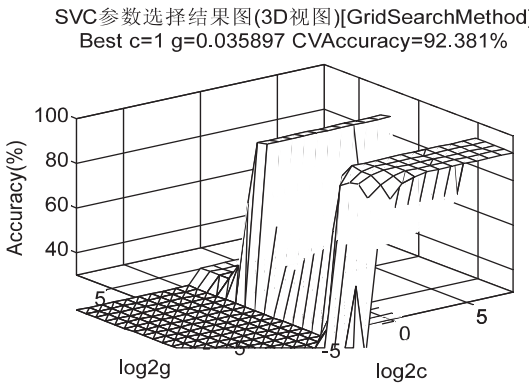

图 5　SVM 参数选择的 3D 视图

参考文献:

[1] 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.

[2] 蒙庚祥,方景龙. 基于支持向量机的手写体数字识别系统设计[J]. 计算机工程与设计,2005,26(6):1592-1598.

[3] 李　晶,姚明海. 基于支持向量机的语义图像分类研究[J]. 计算机技术与发展,2010,20(2):75-78.

[4] 付忠良. 图像阈值选取方法-Otsu 方法的推广[J]. 计算机应用,2000,20(5):37-39.

[5] 李雪峰,李灵峰,刘　芳. 基于遗传算法和 Otsu 理论的图像阈值自动选取[J]. 信息技术,2006(8):52-55.

[6] 郑春红,焦李成,丁爱玲. 基于启发式遗传算法的 SVM 模型自动选择[J]. 控制理论与应用,2006,23(2):187-192.

[7] 陈国良,王煦法,庄镇泉,等. 遗传算法及其应用[M]. 北京:人民邮电出版社,1996.

[8] 胡　康,万金泉. 基于遗传算法的控制系统在废水处理中的应用[J]. 计算机技术与发展,2011,21(2):18-21.

[9] 邓乃扬,田英杰. 支持向量机:理论、算法与拓展[M]. 北京:科学出版社,2009.

[10] 连　可,陈世杰,周建明,等. 基于遗传算法的 SVM 多分类决策树优化算法研究[J]. 控制与决策,2009,24(1):7-12.

[11] 吴　鹏. MATLAB 高效编程技巧与应用:25 个案例分析[M]. 北京:北京航空航天大学出版社,2010.

[12] Feng Jun, Yang Yang, Wang Hong, et al. Feature selection based on genetic algorithms and support vector machines for handwritten similar Chinese characters recognition[J]. Machine Learning and Cybernetics,2004(6):3600-3605.

[13] Wu Chih-Hung, Tzeng Gwo-Hshiung, Goo Yeong-Jia, et al. A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy[J]. Expert Systems with Applications,2007,32(2):397-408.

[14] 周辉仁,郑丕谔,赵春秀. 基于遗传算法的 LS-SVM 参数优选及其在经济预测中的应用[J]. 计算机应用,2007,27(6):1418-1429.

(上接第 4 页)

event-based infrastructure and its application to the development of the OPSS WFMS[J]. IEEE Transactions on Software Engineering,2001,27(9):827-850.

[5] Carzaniga A, Rosenblum D, Wolf A. Design and evaluation of a wide-area event notification service[J]. ACM Transactions on Computer Systems,2001,19(3):332-383.

[6] Diao Y, Altinel M, Franklin M, et al. Path sharing and predicate evaluation for high performance XML filtering[J]. ACM Transactions on Database Systems,2003,28(4):467-516.

[7] Gupta A, Suciu D. Stream processing of XPath queries with predicates[C]//Proceeding of the 2003 ACM SIGMOD International Conference on Management of Data. San Diego: ACM,2003:419-430.

[8] 胡昔祥. 面向大规模分布式计算的发布/订阅系统[J]. 浙江大学学报,2008,42(5):1145-1151.

[9] 尹建伟,施冬材,钱剑锋,等. 结构化 P2P 网络上语义发布/订阅事件路由算法[J]. 浙江大学学报,2008,42(9):1156-1161.

[10] 汪锦岭,金蓓弘,李　京,等. 基于本体的发布/订阅系统的数据模型和匹配算法[J]. 软件学报,2005,16(9):1124-1130.

[11] 薛　涛,冯博琴,李　波,等. 基于内容的发布订阅系统中快速匹配算法的研究[J]. 小型微型计算机系统,2006,27(3):529-533.

基于启发式GA-SVM的手写数字字符识别的研究

作者： 石会芳， 胡小兵， 刘瑞杰， 叶剑英
作者单位： 重庆大学 数学与统计学院,重庆401331
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2012(10)

参考文献(14条)

1. 张猛;余仲秋;姚绍文 手写体数字识别中图像预处理的研究 2006(6-1)

2. 蒙庚祥;方景龙 基于支持向量机的手写体数字识别系统设计 2005(06)

3. 李晶;姚明海 基于支持向量机的语义图像分类研究 2010(02)

4. 付忠良 图像阈值选取方法-Otsu方法的推广 2000(05)

5. 李雪峰;李灵峰;刘芳 基于遗传算法和Otsu理论的图像阈值自动选取 2006(08)

6. 郑春红;焦李成;丁爱玲 基于启发式遗传算法的SVM模型自动选择 2006(02)

7. 陈国良;王煦法;庄镇泉 遗传算法及其应用 1996

8. 胡康;万金泉 基于遗传算法的控制系统在废水处理中的应用 2011(02)

9. 邓乃扬;田英杰 支持向量机:理论、算法与拓展 2009

10. 连可;陈世杰;周建明 基于遗传算法的SVM多分类决策树优化算法研究 2009(01)

11. 吴鹏 MATLAB高效编程技巧与应用:25个案例分析 2010

12. Feng Jun;Yang Yang;Wang Hong Feature selection based on genetic algorithms and support vector machines for handwritten similar Chinese characters recognition 2004(06)

13. Wu Chih-Hung;Tzeng Gwo-Hshiung;Goo Yeong-Jia A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy 2007(02)

14. 周辉仁;郑丕谔;赵春秀 基于遗传算法的LS-SVM参数优选及其在经济预测中的应用 2007(06)