第 22 卷 第 2 期
2007 年 6 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.22 No. 2
Jun. 2007

基于改进交叉验证算法的支持向量机多类识别

邓 蕊[1]，马永军[2]，刘尧猛[2]
（1.天津科技大学电子信息与自动化学院，天津 300222;
2.天津科技大学计算机科学与信息工程学院，天津 300222）

摘 要：如何确定支持向量机最佳参数用以训练得到最优分类器，使之对未知样本同样具有良好的分类效果，一直是问题解决的关键. 针对传统的交叉验证算法仅仅从全局的角度寻找极值点作为最优参数，而忽略了局部信息使得分类效果受到限制问题，提出一种改进的交叉验证算法，在考虑全局极值点的同时，也记录了局部极值点，求取全部极值点对应参数的平均值，由此得到最优参数. 实验结果表明，该算法可以有效地确定最优参数，分类准确率有所提高.

关键词：支持向量机；统计学习理论；交叉验证
中图分类号：TP391.41 **文献标识码**：A **文章编号**：1672-6510（2007）02-0058-04

Support Vector Machine Multi-class Classification Based on an Improved Cross Validation Algorithm

DENG Rui[1]，MA Yong-jun[2]，LIU Yao-meng[2]
（1. College of Electronic Information and Automation，Tianjin University of Science & Technology，Tianjin 300222，China;
2. College of Computer Science and Information Engineering，Tianjin University of Science & Technology，
Tianjin 300222，China）

Abstract：How to find the optimal parameters of SVM to train the model in order to have a better recognition rate when treating unknown test samples has been the most important issue to solve practical problems. Traditional cross validation algorithm only find the optimal parameters in the whole scale of values. A new algorithm was proposed which take into account the extreme values in the limited scale. Results show that this algorithm has a better recognition rate.

Keywords：support vector machine；statistical learning theory；cross validation

机器学习旨在研究从观测样本出发寻找规律，利用这些规律对未来样本或无法观测的样本进行预测. 其重要理论基础之一是统计学.统计学习理论（Statistical Learning Theory, SLT）是一种专门研究小样本情况下机器学习规律的理论. V.Vapnik 等人[1]从六七十年代开始致力于此方面的研究，到 90 年代中期，随着其理论的不断发展和成熟，统计学习理论开始受到越来越广泛的重视. 统计学习理论是建立在一套较坚实的理论基础之上的，为解决有限样本学习问题提供了一个统一的框架. 同时，在这一理论基础之上发展了一种新的通用学习方法——支持向量机（Support Vector Machine, SVM），它已初步表现出很多优于已有方法的性能[2]. 最初的支持向量机是为二分类问题设计的，不能直接应用到多分类问题中，而实际应用中往往遇到的都是多分类问题，一些学者从两个方向研究用支持向量机解决多分类问题：一个方向是将基本的两类支持向量机（Binary-class SVM, BSVM）扩展为多类分类支持向量机（Multi-class SVM, MSVM），使支持向量机本身成为解决多类分类问题的多类分类器；另一方向则相反，将多类分类问题逐步转化为两类分类问题，即用多个两类分类支持向量机组成的多类分类器.

支持向量机模型选择实质上就是如何确定分类器模型参数的问题、交叉验证算法是较为经典的解决办法之一. 传统的交叉验证算法仅仅从全局的角度寻找极值点作为最优参数，而忽略了局部信息，使得分

收稿日期：2006-10-18；**修回日期**：2007-01-08
基金项目：天津市科技攻关重点项目（04310951R）；天津市高等学校科技发展基金资助项目（20061011）
作者简介：邓 蕊（1982—），女，天津人，硕士研究生.

类效果受到限制，本文提出一种改进的交叉验证算法，并以字符识别问题为例，对比了采用传统交叉验证算法和改进算法的字符识别效果.

1　多分类支持向量机及交叉验证算法

依据统计学习理论中的结构风险最小化原则,Vapnik 等人首先提出了模式分类算法[3]. 假设训练样本集为

$$(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l) \tag{1}$$

式中：$x_i\in R^n, i=1,2,\cdots,l$ (R 为实数域)；对于两类的分类问题，$y_i\in\{+1,-1\}$；支持向量机分类算法的原始形式可归结为下列二次规划问题[3]：

$$\min\frac{1}{2}(w,w)+C\sum_{i=1}^{l}\xi_i,$$
$$s.t.\quad y_i((w,x_i)+b)-1+\xi_i\geqslant 0, \tag{2}$$

式中：$(\bullet,\bullet)$ 为两向量之间的内积；$\xi_i\geqslant 0$ 为松弛项，表示错分样本的惩罚程度；C 为常数，用于控制对错分样本惩罚的程度，实现在错分样本数与模型复杂性之间的折衷；w 和 b 为判决函数 $f(x)=(w,x)+b$ 中的权向量和阈值，当无错分样本时，最小化目标函数的第一项等价于最大化两类间的间隔，可降低分类器的 VC 维，实现结构风险最小化原则，上述二次规划的对偶形式为

$$\max\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i,x_j),$$
$$s.t.\sum_{i=1}^{l}\alpha_i y_i=0,\quad 0\leqslant\alpha_i\leqslant C,\quad i=1,2,\cdots,l, \tag{3}$$

其中 α_i 为 Lagrange 乘子.根据最优化理论中的 KKT（Karush-Kuhn-Tucker）条件，只有少量样本（判决函数值等于 ±1 的样本和错分样本）的 α_i 值不为零，称之为支持向量.

由于对偶形式（3）中只出现两向量间的内积运算，用满足 Mercer 条件的核函数 $k(x_i,x_j)$ 来代替内积运算 (x_i,x_j)，实现线形算法的非线性化. 常用的核函数包括：多项式核、径向基核以及二层神经网.

由二分类支持向量机推广可得[4]：

$$\min_{w,\xi,b}\Phi(w,\xi)=\frac{1}{2}\sum_{m=1}^{k}\|w\|^2+C\sum_{i=1}^{l}\sum_{m\neq y_i}\xi_i^m w_{y_i}^T\Phi(x_i)+$$
$$b_{y_i}\geqslant w_m^T\Phi(x_i)+b_m+2-\xi_i^m$$
$$\xi_i^m\geqslant 0,\quad I=1,\cdots,l\quad m\in\{1,\cdots,k\} \tag{4}$$

最终可得到目标决策函数：

$$f(x)=\arg\max_{m=1,\cdots,k}[w_m^T\Phi(x)+b_m] \tag{5}$$

为了确定支持向量机最佳参数用以训练得到最优分类器，最常用的方法就是多次交叉验证（n-fold cross validation）. n 是分组数，如 n=3 就是拆成 3 组，然后先用 1 和 2 来训练分类器并预测 3 以得到正确率；再用 2 和 3 训练并测试 1，最后用 1 跟 3 训练并测试 2. 其他以此类推. 用来做 cross validation 的数据组数对参数的选择影响并不太大——就是说选为 5 和 10 并不会导致最后选到的参数大相径庭. 通常比较重要的参数就是 g 和 c，一般每个参数均遍历 2^{-10}~2^{10} 足以找到最优参数，对于每一个给定的参数对（c, g），均进行 n 次交叉验证，将这 n 次的测试结果取平均值作为该参数对的指标，最后选择指标最高的一组参数对作为最终的最优参数对训练分类器，并对未知样本进行测试.

2　改进的交叉验证算法

算法的基本思想是在进行交叉验证的内层循环中，每次当识别率出现第一次局部最大值时就记录下当时的参数值，并且跳出内层循环. 最后通过计算全部局部最大值的算术平均来估计最优参数对. 算法如下：

输入：数据集 D
输出：用以训练分类器的参数（c,g）
for c=2^{-10}: 2^{10}, g=2^{-10}: 2^{10}
　　for i=1: n
　　取第 i 份作为测试样本 D_{test},剩余 $n-1$ 作为训练样本 D_{train}
　　利用训练样本D_{train}训练分类器
　　利用训练后的分类器对未知样本D_{test}进行测试得到识别率 r
　　$r_{\text{ave}}(j)$=sum(r)/n
　　if $r_{\text{ave}}(j-1)>r_{\text{ave}}(j)$
　　记录$r_{\text{ave}}(j-1)$对应的参数（c,g），即局部极值点
　　break
　　最后计算全部局部极值点（c,g）的算术平均值并取整.

将原始已知样本平均分成 n 份，依次取其中的每一份作为测试样本，而用剩余 $n-1$ 份样本作为训练样本训练产生分类器对测试样本进行测试，所以一次循环结束，对于某一组给定的参数对，一共得到 n 个识别率，然后求取这 n 个识别率的平均值，以此作为该参数对所对应的性能指标，当遍历了全部可能的参数对（c,g）之后，比较相邻参数对的性能指标，记录

下全部局部极值点，对极值点所对应的参数对求取平均值，由此得到了最优参数，作为后续步骤的输入，用以测试未知样本.

3　实验与分析

以多分类问题中最典型的字符识别为例测试算法效果，训练样本为 624 个无噪声标准字符样本，然后利用 MATLAB 图像处理工具箱中的 IMNOISE 函数分别生成包含椒盐、高斯和斑点噪声的测试样本，每类测试样本数均为 624，样本图片如图 1 所示. 核函数选择解决非线性问题常用的 RBF 核函数. 分别利用传统的 n–fold 交叉验证算法和改进算法进行最优参数的选取，本实验中取 n=6，利用选好的参数对（c, g）进行对未知测试样本的识别，得到识别率. 无噪声情况下识别率对比如表 1 所示，有噪声情况下识别率对比曲线如图 2 所示.

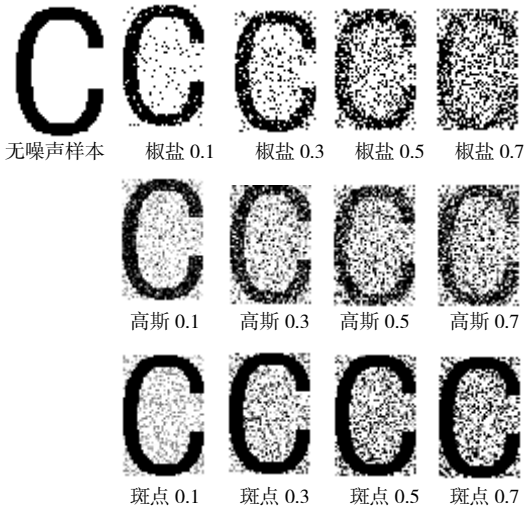

图 1　添加不同等级噪声的样本图片示例

Fig.1　Test samples with noisy input

表 1 无噪声测试样本识别结果

Tab.1　Recognition result of test samples with no noisy input

算法	传统交叉验证	改进交叉验证
最优参数	c=300, g=50	c=14, g=439
识别率/%	99.52	99.87

由实验结果可见，对目前较难处理的一类噪声——椒盐噪声，随着噪声等级的增加，识别率迅速下降，噪声等级达到 0.5 以后，识别率明显下降. 而包含高斯或斑点噪声的测试样本在相同噪声等级的情况下，识别率明显高于高斯噪声，特别是斑点噪声，当噪声等级达到 1.0，仍能保持 95%的识别率. 可见，利用改进交叉验证算法求得的最优参数训练分类器，对未知样本的识别效果有所提高.

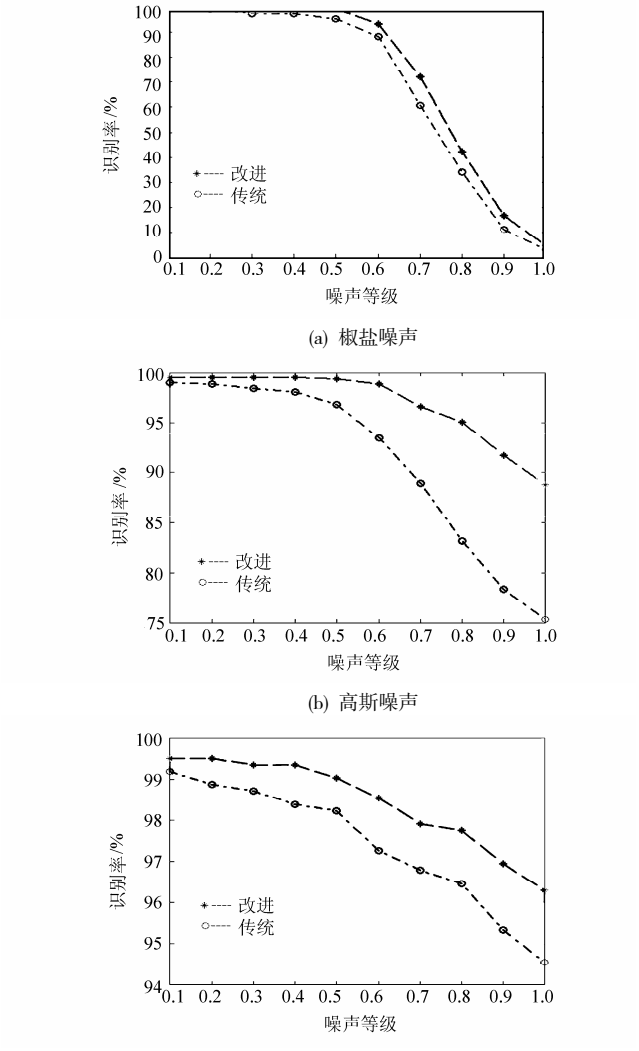

图 2　含有噪声的测试样本识别率

Fig.2　Recognition rate of test samples with noisy input

4　结　论

分析了传统用于确定支持向量机最优参数的 n–fold 交叉验证方法，传统算法只考虑了全局最优参数而未考虑局部信息，针对此缺陷本文提出了一种改进算法，最后以字符识别问题为例对比了两者的分类器识别效果. 实验表明，改进算法与传统算法相对识别率有所提高.

由于交叉验证算法的时间复杂度很高，在普通的机器配置下，要运行数日才能得出结论，所以下一阶段工作重点着眼于如何利用有效的方法降低时间复

杂度，有限时间范围内求取出最佳参数.

参　考　文　献：

［1］ Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.

［2］ 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2001，26（1）：32—42.

［3］ 许建华，张学工，李衍达. 支持向量机的新发展[J]. 控制与决策, 2004，19（5）：481—484.

［4］ Hsu C-W, Lin C-J. A comparison of methods for multi-class support vector machine[J]. IEEE Transactions on Neural Networks，2002（13）：415—425.

（上接第 23 页）

［7］ 王丹丽，王恩德. 针铁矿及腐殖质对水体重金属离子的吸附作用[J]. 安全与环境学报，2001，1（4）：1—4.

［8］ 吕福荣，刘　艳. 腐殖酸对镉、钴作用的研究[J]. 大连大学学报，2002，23（4）：63—67.

［9］ 傅平青，刘丛强，万鹰昕，等. 水环境中腐殖质对重金属吸附行为的影响[J]. 矿物岩石地球化学通报，2002，21（4）：277—281.

［10］王丹丽，关子川，王恩德. 腐殖质对重金属离子的吸附作用[J]. 黄金，2003，24（1）：47—49.

［11］于瑞莲，胡恭任. 泉州湾滩涂沉积物对 Cu（Ⅱ）的吸附实验[J]. 地球与环境，2005，23（1）：93—96.

［12］杨　敏，王红斌，罗秀红，等. 焦磷酸钠法从沼泽土中提取腐殖酸的实验研究[J]. 云南民族学院学报，2002，11（2）：100—102.

［13］杨　敏，王红斌，宁　平，等. 云南沼泽土中提取腐殖酸的研究[J]. 化学世界，2002，7：351—353.

［14］于瑞莲，胡恭任. 苯酚在滩涂沉积物上的吸附特性[J]. 生态环境，2004，13（4）：535—537.

［15］吴　萍，杨桂朋，赵润德. 苯酚在海洋沉积物上的吸附作用[J]. 海洋与湖沼，2003，34（4）：345—352.

（上接第 35 页）

方面，分散效果好，颜料和涂料的黏度和流动性有所改善，使得涂布纸张表面性能较使用 DC 的纸张好；另一方面，使用 poly-DADMAC 的纸张对喷墨打印墨水固色性强，颜色密度较高，颜色饱和度高，密度范围较大. 另外，实验观察得知，采用 poly-DADMAC 分散的涂料能较长时间保持涂料的悬浮状态. 因此，poly-DADMAC 用作阳离子分散剂，可以有效地保持涂料的稳定性，提高喷墨打印性能.

参　考　文　献：

［1］ 钱鹭生，曹丽云，凌永龙，等. 涂布加工纸技术手册[M]. 北京：中国轻工业出版社，2000.

［2］ 张运展. 加工纸与特种纸[M]. 北京：中国轻工业出版社，2001.

［3］ Prakash B Malla, Siva Devisetti. Novel kaolin pigment for high solids ink jet coating[J]. PITA Coating Conference，2005（3）：1—19.

［4］ Quintin Parker. Binder choice in coating formulations balances gluring，print characteristics[J]. Pulp and Paper. 2004（5）：53—56.

［5］ 王玉丰，黄红生，陆建辉. 彩色喷墨打印纸性能研究[J]. 中国造纸，2005（9）：74—75.

［6］ 刘新艳，赵传山. 普通喷墨打印纸涂料配方的研制[J]. 造纸化学品，2004（3）：13—18.

［7］ 曹邦威. 纸张颜料涂布与表面施胶[M]. 北京：中国轻工业出版社，2001.

［8］ 王玉珑，赵传山，杨　飞. 影响颜料分散的几个因素[J]. 纸和造纸，2003（4）：49—51.