第 20 卷 第 7 期　　控 制 与 决 策　　2005 年 7 月
Vol. 20 No. 7　　Control and Decision　　July 2005

文章编号: 1001-0920(2005)07-0746-04

支持向量机多类分类算法研究

唐发明, 王仲东, 陈绵云
(华中科技大学 控制科学与工程系, 武汉 430074)

摘　要: 提出一种新的基于二叉树结构的支持向量(SVM)多类分类算法. 该算法解决了现有主要算法所存在的不可分区域问题. 为了获得较高的推广能力, 必须让样本分布广的类处于二叉树的上层节点, 才能获得更大的划分空间. 所以, 该算法采用最小超立方体和最小超球体类包含作为二叉树的生成算法. 实验结果表明, 该算法具有一定的优越性.

关键词: 支持向量机; 多类分类; 二叉树; 多类支持向量机
中图分类号: TP391　　**文献标识码**: A

On Multiclass Classification Methods for Support Vector Machines

TANG Fa-ming, WANG Zhong-dong, CHEN Mian-yun
(Department of Control Science and Engineering, Huazhong University of Science and Technolgy, Wuhan 430074, China. Correspondent: TANG Fa-ming, E-mail: trobust@263.net)

Abstract: The multiclass SVM methods based on binary tree are proposed. The new methods can resolve the unclassifiable region problems in the conventional multiclass SVM methods. To maintain high generalization ability, the most widespread class should be separated at the upper nodes of a binary tree. Hypercuboid and hypersphere class least covers are used to be rules of constructing binary tree. Numerical experiment results show that the multiclass SVM methods are suitable for practical use.

Key words: Support vector machines; Multiclass classification; Binary tree; Multiclass support vector machines

1 引　　言

支持向量机(SVM)方法最初是针对二类别的分类而提出的, 如何将其有效的推广到多类别分类仍是当前支持向量机研究的重要内容之一. 目前, 对于多类分类问题, SVM 的解决途径有两种: 一种是通过构造多个 SVM 二值分类器并将它们组合起来实现多类分类, 例如 one-against-rest[1], one-against-one 和 DAGSVM[2]. 虽然这三种方法是当前最常用且性能较优的, 但 one-against-rest 和 one-against-one 方法的泛化误差是无界的. 再者, one-against-one 所需构造的子分类器的数量关于类别数 k 成超线性增长, 共 $k(k-1)/2$ 个, 且在测试阶段, 都必须计算所有子分类判据函数. One-against-one 方法还有一个最明显的缺点就是, 每个子分类器必须都要非常仔细地调整, 如果某个子分类器不规范化, 则整个分类系统将趋于过学习. DAGSVM 方法解决了不可分区域问题, 而且不一定要计算所有的子分类判决函数, 但各个子分类器在有向无环图中的位置也会对分类系统产生较大的影响.

另一种是直接在一个优化公式中同时考虑所有子分类器的参数优化. 严格的讲, 其思想类似于 one-against-rest 方法, 只不过是把 k 个二值 SVM 优化问题放在一个最优化公式中同时优化, 所以它也存在 one-against-rest 方法相同的缺点. 另外, 这

收稿日期: 2004-08-29; **修回日期**: 2004-11-19.
基金项目: 国家自然科学基金项目(79970025); 国防科技预研基金项目(00J15.3.3.JW0528).
作者简介: 唐发明(1977—), 男, 江西吉安人, 博士生, 从事模糊建模、计算机集成控制等研究; 陈绵云(1937—), 男, 湖北竹山人, 教授, 博士生导师, 从事一般系统论、模糊系统论等研究.

种思想尽管看起来简洁, 但在最优化问题求解过程中的变量远远多于第 1 种, 训练速度不及第 1 种, 且在分类精度上也不占优[3]. 当训练样本数非常大时, 这一问题更加突出.

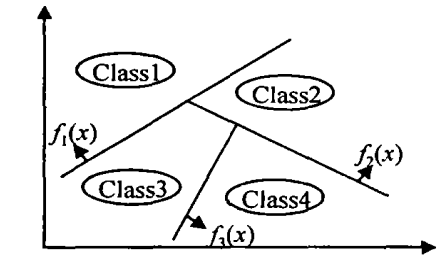
因此, 本文在对现有主要的 SVM 多类分类算法作简单介绍的基础上, 提出了新的基于二叉树的 SVM 多类分类方法, 并通过一系列实验分析、比较了各种算法的特点.

2　基于二叉树的多类 SVM

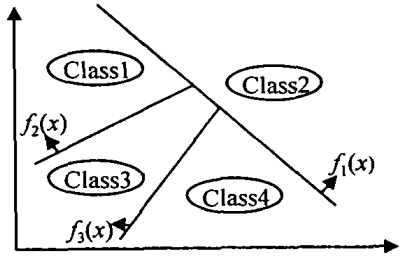
基于二叉树的多类 SVM 是先将所有类别分成两个子类, 再将子类进一步划分成两个次级子类, 如此循环下去, 直到所有的节点都只包含一个单独的类别为止, 此节点也是决策树中的叶子. 该方法将原有的多类问题同样分解成了一系列的两类分类问题, 其中两个子类间的分类函数采用 SVM. 二叉树方法可以避免传统方法的不可分情况, 并且只需构造 $k-1$ 个 SVM 分类器, 测试时并不一定需要计算所有的分类器判别函数, 从而可节省测试时间.

二叉树的结构对整个分类模型的分类精度有较大的影响. 图 1 是一个 4 类问题的不同的二叉树法构造示意图. 在图 1(a) 中, 第 1 个分割面是由第 1 类和第 2, 第 3, 第 4 类构成, 第 2 个分割面是由第 2 类和第 3, 第 4 类构成, 最后一个分割面是由第 3 类和第 4 类构成; 而图 1(b) 的分割顺序是第 2 类, 第 1 类, 第 3 类. 从此例可看出, 分割顺序不一样, 每个类的分割区域也不同. 因此, 此多类 SVM 方法的每个类的区域依赖于二叉树的结构, 主要是二叉树节点所代表的二值 SVM 分类器的位置.

二叉树的结构有两种: 一种是在每个内节点处, 由一个类与剩下的类构造分割面; 另一种是在内节点处, 可以是多个类与多个类的分割. 本文只考虑前一种情况, 即每次分割只分割出一个类. 基于二叉树的多类 SVM, 在测试阶段类似 DAGSVM, 从根节点开始计算决策函数, 根据值的正负决定下一节点, 如此下去, 直到到达某一叶节点为止, 此叶节点所代表的类别就是测试样本的所属类别.



(a)　分割顺序为 1, 2, 4, 3



(b)　分割顺序为 2, 1, 3, 4

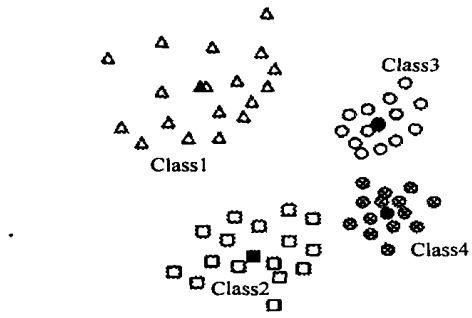
图 1　4 类问题的不同划分顺序

目前, 基于二叉树的多类 SVM 分类方法已有学者提出, 文献[4～6]的基本思想都是基于二叉树的分类. 但这些方法不是随机地生成二叉树, 就是采用的二叉树生成算法并不能很好的提高整个分类模型的推广能力. 从前面的分析可看出, 越上层节点的分类性能对整个分类模型的推广性影响越大. 因此, 在生成二叉树的过程中, 应该让最易分割的类最早分割出来, 即在二叉树的上层节点处分割. 基于此, 本文提出了根据训练样本在属性空间的几何分布情况来生成二叉树的方法, 从而建立一个推广性高的多类 SVM 分类模型. 由于支持向量机的思想是在样本的属性空间中构造最优超平面, 线性 SVM 的属性空间等价于输入空间, 但非线性的 SVM 却无法得到具体的属性空间表达式. 事实上, 样本在输入空间中的物理联系在属性空间也同样存在. 所以, 只需在输入空间中考虑样本的几何分布.

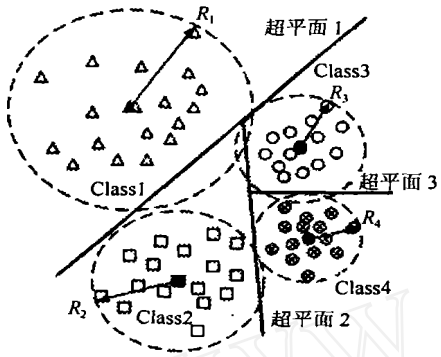
3　改进的多分类二叉树法

二叉树多类分类法的每个类的区域依赖于二叉树的生成顺序, 主要是二值 SVM 分类器所在的内节点位置. 为了提高分类模型的推广能力, 必须利用合理的策略来生成二叉树结构. 所以, 本文提出以类样本分布情况作为二叉树的生成算法, 从而构造推广能力好的基于二叉树的 SVM 多类分类模型. 改进算法的基本思想就是在每次生成二叉树内节点时, 选择最易分割的情况来构造当前节点的二值 SVM.

分割顺序不一样, 每个类的分割区域是不同的, 先分割出来的类更容易有较大的分割区域. 为了让分布广的类拥有较大的分割区域, 就应最先把这些类分割出来. 因为各类数据的真实分布无法得知, 所以用有限样本数据的分布来对真实分布作近似估计. 样本分布范围的度量可采用超长方体和超球体最小包含某一类的样本, 超长方体(或超球体)的体积就是此类样本的分布度量. 图 2(a) 为 4 类样本数据的二维输入空间分布图, 直观上看, 最好的分割顺序为: 以第 1 类与其他类构造分割超平面, 接着第 2 类与第 3 和第 4 类构造分割超平面, 最后是第 3 与第 4 类构造分割超平面, 这主要是考虑到各类样本空间分布范围的大小.



(a) 各类样本的分布



(b) 最小包含圆及分割

图 2 分布范围法示意图(黑体为各类的重心)

定义 1(超长方体最小类包含) 设类 S 有 n 个 d 维样本向量 $x_1, x_2, \cdots, x_n$, $x_i \in R^d$, 则最小包含类 S 样本的超长方体体积为

$$v = \prod_{i=1}^{d} (\max\{x_j^i, j = 1, 2, \cdots, n\} - \min\{x_j^i, j = 1, 2, \cdots, n\}). \quad (1)$$

定义 2(超球体最小类包含) 设类 S 有 n 个样本 $x_1, x_2, \cdots, x_n$, 此类样本集的重心为 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, 则最小包含这些样本的超球体半径为

$$R = \max_{x_i \in S}\{\| \bar{x} - x_i \|\}. \quad (2)$$

式中 $\| \cdot \|$ 表示欧氏距离运算. 超球体的体积计算为 $v = \pi R^d$, 因为目的只是比较各个类别分布范围的相对大小, 所以为了简化计算, 可用半径代替体积, 只需计算式(2) 即可, 见图 2(b).

除了样本分布范围度量定义不同之外, 基于超长方体最小类包含和超球体最小类包含二叉树生成法的多类别 SVM 分类算法流程几乎是相同的, 具体的算法流程为:

Step 1: 根据式(1) (或(2)) 计算各类样本数据的分布体积 v_i ($i = 1, 2, \cdots, k$).

Step 2: 根据各类的分布体积由大到小的顺序, 对类别进行排序. 当存在两个或两个以上的类别具有相同分布体积时(这种情况很少出现), 把类标号小的类排在前面. 最后得到所有类别的排列 $n_1, n_2, \cdots, n_k$, 此处 $n_m \in \{1, 2, \cdots, k\}, m = 1, 2, \cdots, k$ 为类标号.

Step 3: 利用二值分类的 SVM 训练算法构造二叉树各内节点的最优超平面. 在根节点处, 从样本集中选择第 n_1 类样本为正样本集, 其他样本为负样本集, 利用 SVM 训练算法构造最优超平面, 然后把属于第 n_1 类的样本从样本集中删除. 在第 2 个节点处, 从样本集中选择第 n_2 类样本为正样本集, 其他剩余的样本为负样本集, 利用 SVM 训练算法构造最优超平面, 然后把属于第 n_2 类的样本从样本集中删除. 依次下去, 最终可得到如图 3 所示的基于二叉树的多类别 SVM 分类模型.

Step 4: 算法结束.

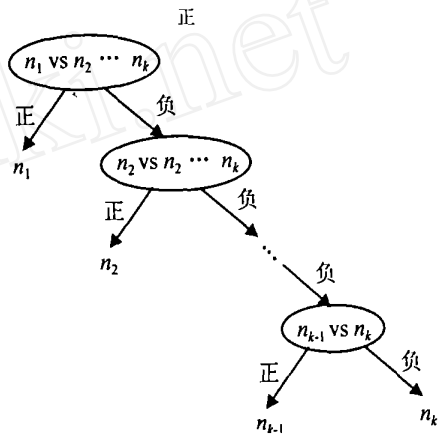


图 3 多类别 SVM 分类模型的二叉树结构

4 实 验

4.1 实验数据和实现

为了比较各种多类 SVM 算法的性能, 本文使用 Statlog[7] 数据库中的多类别数据集 letter、satimage 和 shuttle 对 one-against-one, one-against-rest 以及本文提出的两种算法分别进行实验测试. 表 1 列出了各数据集的一些信息, 最后一列是 Statlog 主页上给出的根据不同学习算法所取得的最好测试结果.

为了避免取值范围大的属性比取值范围小的属性更占优势, 所以必须对样本数据各属性进行归一化预处理, 线性调整到[- 1, + 1]. 实验中本文只使用 RBF 核函数 $K(x_i, x_j) = \exp(-\gamma\| x_i - x_j \|^2)$, 以便减少最优参数集的搜索空间. 考虑模型的推广能力与常数 (C, γ) 有关, 对于每一个优化问

表 1 本文实验数据集统计表

问题	训练样本数	测试样本数	类别数	属性数	Statlog rate/%
letter	15 000	5 000	26	16	93.6
satimage	4 435	2 000	6	36	90.6
shuttle	43 500	14 500	7	9	99.99

表 2　实验中具有最好推广能力的参数值和测试精度

Problem	1vs1 (C, γ) rate	1vsRest (C, γ) rate	超长方体最小包含 (C, γ) rate	超球体最小包含 (C, γ) rate
Letter	(2^7, 2^1) 97.76	(2^3, 2^1) 95.68	(2^4, 2^1) 96.84	(2^3, 2^1) 96.86
Satimage	(2^3, 2^1) 92.00	(2^3, 2^0) 89.90	(2^1, 2^1) 91.45	(2^1, 2^1) 91.50
Shuttle	(2^{12}, 2^3) 99.92	(2^{12}, 2^2) 99.91	(2^{12}, 2^3) 99.92	(2^{12}, 2^3) 99.93

表 3　训练时间、测试时间以及支持向量数　(s)

Problem	one-against-one		one-against-rest		超长方体		超球体	
	训练时间 测试时间	# SVs u-SVs	训练时间 测试时间	# SVs u-SVs	训练时间 测试时间	# SVs u-SVs	训练时间 测试时间	# SVs u-SVs
Letter	111 42	7 619 7 558	258 59	13 909 7 098	123 26	10 891 6 185	120 27	11 032 6 286
satimage	12 4	2 048 2 048	23 8	4 034 1 706	15 6.4	3 991 2 023	19 4.1	3 337 1 680
Shuttle	126 2	273 273	464 3	388 258	175 1.2	250 174	181 1.2	250 174

题, 可利用不同的核参数 γ 和惩罚系数 C 组合, 即 $\gamma = [2^{-1}, 2^0, \cdots, 2^4]$, $C = [2^0, 2^1, \cdots, 2^{12}]$. 从而对于每个问题, 可以尝试 $6 \times 13 = 78$ 种参数组合来建立模型, 从中得到推广精度高的 γ 和 C 参数. 当存在多个相同的最高测试率时, 取支持向量数少的那组 γ 和 C 参数. 每个实验的 KKT 停止条件的容许误差为 0.001. 本文所有算法均采用 C++ 实现, 并利用 VC++ 编译, 二值 SVM 分类算法是在 LIBSVM 工具包的基础上修改实现的. 实验平台为 P4 2.0G, 512M RAM 的 Dell PC, 操作系统为 Windows 2000. 对于每个优化问题, 分配 40MB 的内存作为核函数计算值的缓存, 同时采用 Shrinking 软件实现技术来加快训练.

4.2　结果和分析

表 2 列出了 5 种方法的试验比较结果, 其中只列出最优参数 (C, γ) 及对应的测试准确率. 由表 2 可看出, 不同的问题、不同的分类算法, 最优参数 (C, γ) 所在的区域也会不同, 这也是要测试多个 (C, γ) 参数对的原因. 从实验结果可以看到, 本文算法虽然采用不同的二叉树生成方法, 使得每个类的分割顺序不一样, 但最优参数下的测试正确率却非常接近, 且与性能表现最好的 one-against-one 方法也比较接近. One-against-rest 应该是本次实验中推广性能稍差的一种 SVM 多类分类算法.

表 3 列出了实验的训练时间、测试时间以及支持向量数. 支持向量数的一列中 # SVs 表示的是各个二值分类器的支持向量总和, u-SVs (unique SVs) 表示的是不相同的支持向量个数, 因为一个训练样本可能会是多个二值分类器中的支持向量. 在测试阶段, 对于相同的支持向量 x_i 与测试样本 x 的核函数值 $K(x_i, x)$ 只需计算一次, 从而可以减少测试时间. 训练时间和测试时间是以程序运行的 CPU 时间为准, 单位为 s. 本文算法的训练时间包括了类样本体积的计算和排序.

在 3 个实验中, 本文算法的测试时间比其他两种算法少, 因为二叉树法测试样本时并不需要计算所有的二值分类器. one-against-rest 方法的支持向量个数远远多于其他的算法, 这是因为其每个子分类器都需利用到所有的训练样本, 构造的分类面较其他算法复杂. one-against-one 和 one-against-rest 的测试时间相对长, 是因为这两种方法必须计算所有的二值 SVM 分类器判别函数, 且 one-against-rest 的总支持向量数较多, 使得其测试时间更长.

总体上, 本文提出的两种算法的测试时间和测试正确率都较好, 超球体最小类包含法的测试正确率更高一些, 这是因为以类重心为球心的球体包含更能反映类样本的真实分布范围, 而超立方体包含是以样本各属性最值来构造的, 并不一定是最小包含类样本.

5　结　　语

本文首先分析了当前使用较多的几种 SVM 多类分类算法的特点以及存在的一些问题. 在此基础上, 提出了基于二叉树的 SVM 多分类算法. 新的二叉树生成算法可以使得分布广的类别在属性空间中获得更大的划分区域, 从而提高多分类模型的推广性能. 最后, 本文通过几个实验, 对这些算法进行了比较, 实验结果表明本文算法可以明显的减少测试时间, 且分类精度也较理想.

参考文献(References)

[1] Bottou L, Cortes C, Denker J, et al. Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition [A]. *Proc of the Int Conf on Pattern Recognition* [C]. Jerusalem, 1994: 77-87.

(下转第 754 页)

取参数 $\alpha_1 = 0.6, \alpha_2 = 2\alpha_1/(1+\alpha_1) = 0.75, \alpha_3 = 0.6, k_1 = 5, k_2 = k_3 = 2$. 当初始值为 $x_r(0) = y_r(0) = 0$ m, $\theta_r(0) = 0$ rad, $x(0) = 2.5$ m, $y(0) = -2$ m, $\theta(0) = 1$ rad 时, 即 $x_e(0) = 0.332$ m, $y_e(0) = 3.184$ m, $\theta_e(0) = -1$ rad 时, 仿真结果如图 4～图 6 所示.

5 结 语

与基于非连续状态反馈的有限时间控制技术相比, 基于连续状态反馈的有限时间控制技术更适合于控制工程应用. 本文将基于连续状态反馈的有限时间控制技术应用于基于运动学模型的非完整移动机器人的跟踪控制问题, 设计出有限时间收敛的跟踪控制算法. 通过对不同状态分别采用有限时间控制技术进行分步设计, 使得对角速度为非零常数的期望轨迹, 跟踪误差模型在有限时间内收敛, 同时使得非完整移动机器人在有限时间内跟踪上需要的期望轨迹. 仿真结果表明了该方法的有效性.

参考文献(References)

[1] Kolmanovsky H, McClamroch N H. Developments in Nonholonomic Control Systems [J]. *IEEE Control System Magazine*, 1995, 15(6): 20-36.

[2] Kanayama Y, Kimura Y, Miyazaki F, et al. A Stable Tracking Control Method for an Autonomous Mobile Robot [A]. *Proc of IEEE Int Conf on Robotics and Automation* [C]. Cincinnati: IEEE Computer Society Press, 1990: 384-389.

[3] Fliess M, Levine J, Martin P, et al. Flatness and Defect of Nonlinear Systems: Introductory Theory and Examples[J]. *Int J of Control*, 1995, 61(6): 1327-1361.

[4] 董文杰, 霍伟. 受非完整约束移动机器人的跟踪控制 [J]. ***自动化学报***, 2000, 26(1): 1-6.
(Dong W J, Huo W. Tracking Control of Mobile Robots with Nonholonomic Constraint [J]. *Acta Automatica Sinica*, 2000, 26(1): 1-6.)

[5] Jiang Z P, Nijmeijer H. Tracking Control of Mobile Robots: A Case Study in Backstepping [J]. *Automatica*, 1997, 33(7): 1393-1399.

[6] 李世华, 田玉平, 非完整移动机器人的轨迹跟踪控制 [J]. ***控制与决策***, 2002, 17(3): 301-305.
(Li S H, Tian Y P. Trajectory Tracking Control of Nonholonomic Mobile Robots [J]. *Control and Decision*, 2002, 17(3): 301-305.)

[7] Bhat S P, Bernstein D S. Finite Time Stability of Homogeneous Systems [A]. *American Control Conf* [C]. Evanston: American Autom Control Council, 1997: 2513-2514.

[8] Yu X H, Man Z H. Multi-input Uncertain Linear Systems with Terminal Sliding-mode Control [J]. *Automatica*, 1998, 34(3): 389-392.

[9] Hong Y. Finite-time Stabilization and Stabilizability of a Class of Controllable Systems [J]. *Systems and Control Letters*, 2002, 46(4): 231-236.

[10] 李世华, 田玉平. 移动小车的轨迹跟踪控制[J]. ***控制与决策***, 2000, 15(5): 626-628.
(Li S H, Tian Y P. Tracking Control of Mobile Robots [J]. *Control and Decision*, 2000, 15(5): 626-628.)

[11] 李世华, 田玉平. 移动小车的有限时间轨迹跟踪控制 [J]. ***东南大学学报***, 2004, 34(1): 113-116.
(Li S H, Tian Y P. Trajectory Tracking Control of Mobile Robots in Finite Time [J]. *J of Southeast University*, 2004, 34(1): 113-116.)

[12] Dixon W E, Dawson D M, Zergeroglu E, et al. *Nonlinear Control of Wheeled Mobile Robots* [M]. London: Springer-Verlag, 2000.

(上接第 749 页)

[2] Platt J, Cristianini N, Shawe-Taylor J. Large Margin DAG's for Multiclass Classification [A]. *Advances in Neural Information Processing Systems* 12 [C]. Cambridge, MA: MIT Press, 2000: 547-553.

[3] Hsu C, Lin C. A Comparison of Methods for Multiclass Support Vector Machines [J]. *IEEE Trans on Neural Networks*, 2002, 13(2): 415-425.

[4] Takahashi F, Abe S. Decision-Tree-Based Multiclass Support Vector Machines[A]. *Proc of the 9th Int Conf on Neural Information Processing* [C]. Singapore, 2002, (3): 1418-1422.

[5] Sungmoon C, Sang H O, Soo-Young L. Support Vector Machines with Binary Tree Architecture for Multi-Class Classification [J]. *Neural Information Processing-Letters and Reviews*, 2004, 2(3): 47-51.

[6] 马笑潇, 黄席樾, 柴毅. 基于 SVM 的二叉树多类分类算法及其在故障诊断中的应用[J]. ***控制与决策***, 2003, 18(3): 272-276.
(Ma X X, Huang X Y, Chai Y. 2PTMC Classification Algorithm Based on Support Vector Machines and Its Application to Fault Diagnosis [J]. *Control and Decision*, 2003, 18(3): 272-276.)

[7] Michie D, Spiegelhalter D, Taylor C. *Machine Learning, Neural and Statistical Classification* [DB/OL]. http://www.liacc.up.pt/ML/statlog/datasets.html, 1994.