文章编号:1000-582X(2012)06-125-04

超球体支持向量机的不完全二叉树多类分类算法

黄扬帆[1],张慧敏[2],徐子航[1],曹鹏程[1]
(1.重庆大学 通信工程学院,重庆 400044;2.重庆电子工程职业学院 通信系,重庆 401331)

摘 要:针对现有的支持向量机多类分类方法的不足之处,提出了一种基于超球体支持向量机的不完全二叉树多类分类算法。该算法首先采用超球体 SVM 算法,计算各类样本群的分布范围。再利用距离公式,计算各类样本间的距离,基于将最容易分离出来的类最先分割出来的原则,设计二叉树结构,从而提高分类精度。通过仿真实验,分析比较各种方法的性能,从而验证了该算法的有效性。

关键词:支持向量机;多类分类;超球体;二叉树
中图分类号:TP18 **文献标志码**:A

An incomplete binary tree SVM multi-class ciassification algorithm based on hypersphere

HUANG Yang-fan[1], *ZHANG Hui-min*[2], *XU Zi-hang*[1], *CAO Peng-cheng*[1]
(1. College of Communication Engineering, Chongqing University, Chongqing 400044, P. R. China;
2. Department of Communication, Chongqing College of Electronic Engineering, Chongqing 401331, P. R. China)

Abstract: On the base of current researches on multiclass classification with support vector machine, an incomplete binary tree SVM multi-class classification algorithm based on hypersphere is proposed. The algorithm adopts hypersphere SVM algorithm to calculate the distribution of each sample groups. Then, the distance formula is used to calculate the distance among the sample classes. According to the principle that the class which can be separated easiest must be split first, the algorithm designs binary tree to improve the classification accuracy. Compared with many classification methods, the effectiveness of the algorithm is verified by simulation experiments.

Key words: support vector machine; multi-class classification; hypersphere; binary tree

支持向量机(SVM)在解决非线性、小样本及高维模式识别中表现出许多特有的优势,因此被广泛应用于模式识别、图像处理、故障诊断、回归估计等领域,并取得了良好的效果[1-3]。

然而支持向量机最初是为了解决二类分类问题提出的,但在实际应用中遇到的往往是多类分类的问题,比如文本分类、调制方式识别、数字识别、故障诊断等等。如何将二类分类推广到多分类问题一直是支持向量机研究的重要内容。目前,SVM 多类分类问题大致有 2 种解决方法:一是一次性求解[4],将多分类问题在一个优化公式中体现出来,然后再对这个优化式进行直接求解;二是构造多个二类分类器,再由这些二类分类器按照不同方式组合成多类分类器,按组合方式的不同,可以分为:一对一

收稿日期:2011-12-30
基金项目:国家自然科学基金资助项目(61071190);重庆市自然科学基金重点资助项目(CSTC,2009BA2021);2009 重大科技专项"信息制造业"资助项目(CSTC,2010AB2002)
作者简介:黄扬帆(1964-),男,重庆大学博士,高级工程师,主要从事信号与信息处理、模式识别等方向研究。

(OVO)[5]、一对多(OVR)[6]、有向无环图(DAG)[7]、二叉树[8]等。

考虑 k 类分类问题,一次性直接求解法在理论上看似简单可行,但在实际应用中,由于优化问题所需要优化的参数多,其计算量是很大的。特别是当训练的样本数量比较多的时候,这个计算量大的缺点就更加突出。因此在实际应用中,很少采用这种一次性求解的方法来实现 SVM 的多类分类。OVO 算法在每个 SVM 子分类器中只考虑到 2 个样本,所以训练过程简单,单个子分类器的训练时间短。但是若某个子分类器不规范,就可能导致整个分类器趋于过学习,而且还存在着推广误差无界所需构造和测试的二类分类器的数量随着类别数量增加而急剧增加、存在拒识区域等缺点。OVR 对 k 类样本只需构造 k 个子分类器,由于各个样本都要参与到每个子分类器中,故整个训练时间较长,并且存在拒识区域的缺点。DAG SVM 与 OVO 的主要区别是,对一个未知样本,DAGSVM 只需要 $k-1$ 个分类器,而 OVO 需要所有的,即 $k(k-1)/2$ 个分类器[9-10]。

二叉树方法的提出解决了存在拒识区域的问题,该方法首先将整个样本集分成 2 个子样本集,再分别将每个子样本集又各自分成 2 个次级子样本集,依此类推,直到所有的样本类别都被单独分类出来,所以不存在拒识区域[11-12]。

SVM 一般来说是基于超平面来实现的,但是基于超球面的 SVM 在分类过程中可以更好的考虑到样本的平衡问题[13-14]。因此,在现有的二叉树结构以及超球体支持向量机的基础上,提出了一种基于超球体支持向量机的不完全二叉树多类分类算法。

1 超球体 SVM

Tax 和 Duin 提出了一种基于超球体支持向量机的数据描述(SVDD)[15]方法,其主要思想是,在高维空间中以尽可能小的半径包含尽可能多的样本,并计算出包含样本的最小超球体——球心和半径,权衡超球体半径和它所覆盖的样本数,本质上就等价于类内聚类性的最大化。

将 SVDD 思想引入到多类分类问题中。对于 k 类分类问题,给定 k 个 n 维向量空间的元素集合 $X_m, m=1,2,\cdots,k$。每个集合 X_m 中包含 l_m 个 n 维训练样本 $x_{m,i}, i=1,2,\cdots,l_m$,这个集合中的所有样本都属于类 m。现在来寻找一个超球体 (a_m, R_m),其中 a_m 是球心,R_m 是半径。考虑一些落在球面附近的样本点,引入松弛变量 $\xi_{m,i}$。为了将理论应用于非线性系统中,可以将输入通过映射函数 $\varphi(x_{m,i})$ 映射到高维特征向量,再在高维特征向量中进行分类。对第 m 类来说,应该使集合 X_m 所包含的样本数尽量多,半径 R_m 尽量小。于是其优化问题可以描述如式(1)所示,采用拉格朗日乘子法求解可以求得原函数的对偶函数如式(2)所示。

$$\begin{aligned}&\min_{R,m} \quad R_m^2 + C_m \sum_{i=1}^{l_m} \xi_{m,i},\\ &\text{s.t.} \quad \begin{cases} R_m^2 - (\varphi(x_{m,i}) - a_m)(\varphi(x_{m,i}) - a_m)^{\mathrm{T}} \geqslant \\ \qquad -\xi_{m,i}, \xi_{m,i} \geqslant 0, \\ i = 1,2,\cdots,l。\end{cases}\end{aligned} \tag{1}$$

$$\begin{aligned}&\max \quad \sum_{i=1}^{l_m} \alpha_{m,i} K(x_{m,i}, x_{m,i}) - \sum_{i,j=1}^{l_m} \alpha_{m,i}\alpha_{m,j} K(x_{m,i}, x_{m,j}),\\ &\text{s.t.} \quad \begin{cases} 0 \leqslant \alpha_{m,i} \leqslant C_m, \\ \sum_{i=1}^{l} \alpha_{m,i} = 1, \\ i = 1,2,\cdots,l_m。\end{cases}\end{aligned} \tag{2}$$

求解式(2)可以得到

$$a_m = \sum_{i=1}^{l_m} \alpha_{m,i} \varphi(x_{m,i})。 \tag{3}$$

利用 QP 方法可以解出 $\alpha_{m,i}$ 的值,带入式(3)可得到 a_m 的值,再利用 KKT 条件,当 $0<\alpha_{m,i}<C_m$ 时,$\xi_{m,i}=0$,其对应的样本点在超球面上,这些样本点也就是支持向量,再根据 $R^2 = \| \varphi(x_{m,i}) - a \|$,可以求出 R_m 的值。以此类推,可以求得所有样本类别的超球体 (a_m, R_m)。

2 算法分析

二叉树的结构对分类系统的训练时间和分类精度都是有影响的,比如完全二叉树和偏二叉树,它们的训练时间与分类精度就不同。

定义 1(完全二叉树) 若设二叉树的高度为 h,除第 h 层外,其它各层 $(1\sim h-1)$ 的结点数都达到最大个数,而且第 h 层所有的节点都连续集中在最左边,这就是完全二叉树。

SVM 的训练时间 $t=cl^{\gamma}$ 与支持向量的数量成超线性关系[16],其中,c 为常数,l 为训练样本数,γ 与具体的支持向量机算法有关。SVM 多类分类器的分类速度取决于 2 方面因素:①分类时所用的二分类 SVM 个数;②各个二分类 SVM 所包含的支持向量数目。与完全二叉树结构相比,偏二叉树结构的大部分分类器的训练样本数量相对较多,因此它总的训练速度要比完全二叉树的速度稍慢。

二叉树的不同的层次结构对分类精度也是有一定影响的，并且这种影响有可能产生“误差累积”的现象，即若在某个结点上发生分类错误，后面的分类将会把错误延续下去，该结点后续下一级结点上的分类就失去意义[17]。

对基于二叉树的多类分类器来说，为了提高分类精度，在分类的时候要考虑的问题包括如何让最容易分离出来的类最先分割出来。

在超球体不完全二叉树分类中，不拘于完全二叉树的格式，与聚类思想类似，将距离最远的 2 个类首先确定出来，然后再根据类间距离选择二值分类器的正负类样本。为提高分类精度，整个算法都是基于将最容易分离出来的类最先分割出来的原则。

对于 K 类分类，定义集合 S，S_1，S_2，以及各个集合中的元素个数 N_S，N_{S1}，N_{S2}。定义距离函数

类 i 到类 j 的距离：$d_{ij}=a_i-a_j-(R_i+R_j)$。　(4)

类 i 到 S_1 的距离：$d_{iS1}=\min\{d_{ij}\}$，j 为集合 S_1 中类的标号；

类 i 到 S_2 的距离：$d_{iS2}=\min\{d_{ij}\}$，j 为集合 S_2 中类的标号；

具体算法描述如下

1)对于 K 类分类问题，选择合适的核函数，按照公式(1)建立其优化方程，求解得到各类的包裹超球体(a_i,R_i)，也就得到了各类的几何空间分布，如图 1 所示，将每个球体所代表的类标号按从小到大顺序置入一个集合 S 中。

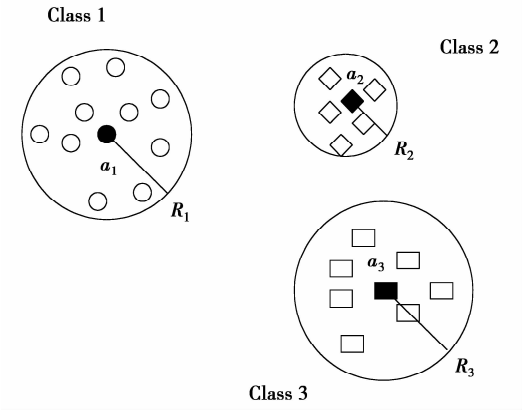

图 1　样本几何空间分布

2)利用公式(4)从集合 S 中选出 d_{ij} 值最大的 2 类 i 和 j，将它们按类标号大小分别置入集合 S_1，S_2 中，$S=S-(S_1\cup S_2)$；

3)如果 $S=\phi$，转入 5)；

4)计算 S 中各类到 S_1 与 S_2 中的最小距离 d_{iS1} 和 d_{jS2}，如果 $d_{iS1}<d_{jS2}$，则将 i 类并入 S_1 中，否则将 j 类并入 S_2 中，$S=S-(S_1\cup S_2)$，转入 3)；

5)分别将 S_1，S_2 作为二叉树的左右子树，至此，一个决策节点的左右子类已形成；

6)将左子类进一步分割为 2 个次级子类，方法是令 $S=S_1$，返回 2)，直到每个类都成为二叉树中的叶子节点。

7)同样，将右子类进一步分割为 2 个次级子类，方法同上，令 $S=S_2$，返回 2)，直到每个类都成为二叉树中的叶子节点。

8)算法结束。

3　仿真实验

为了测试基于超球体不完全二叉树多类分类算法的性能，采用 UCI 数据库[18] 中的 Optical Recognition of Handwritten Digits、Glass Identification、Letter Recognition 数据集进行实验。表 1 列出了实验所使用的各个数据集的样本总数、类个数、属性个数信息。

表 1　试验用数据集构成

数据集	样本总数	类个数	属性个数
Optdigits	5 620	10	64
Glass	214	6	10
Letter	20 000	26	16

实验开发与调试工具为 Matlab 7.4。分类器所选的核函数为高斯径向基核函数

$$K(x_i,y_i)=\exp(-\frac{x_i-y_i^2}{\sigma^2})。\qquad(5)$$

从各个类别的样本中分别提取 60%的样本数作为训练，剩余 40%的样本作为测试使用。令 C=100，$\varepsilon=0.005$，采用 5 折交叉验证方法确定核函数。

为了测试超球体不完全二叉树多类分类算法的性能，拿它与 OVR、OVO、以及文献[6]所介绍的完全二叉树算法做比较。表 2、表 3 分别列出了各种算法的分类精度及分类时间。

表 2　不同数据集下 4 种算法的分类精度　%

数据集	OVR	OVO	完全二叉树	研究算法
Optdigits	94.32	94.24	94.22	94.52
Glass	67.10	64.82	65.52	68.28
Letter	97.77	97.78	95.12	98.02

表3 不同数据集下4种算法的分类时间 s

数据集	时间	OVR	OVO	完全二叉树	研究算法
Optdigits	训练	40.4	37.6	25.9	32.7
	测试	20.0	21.3	17.6	19.3
Glass	训练	0.254	0.296	0.132	0.141
	测试	0.101	0.140	0.042	0.065
Letter	训练	207.6	160.4	129.3	142.4
	测试	58.4	57.4	32.9	38.0

从表2和表3可以得出相同条件下，采用不同的多类分类方法时，SVM分类器的性能，包括分类精度与分类时间。由于采用了基于将最容易分离出来的类最先分割出来的原则，所提出的不完全超球体二叉树算法的分类精度在这几种算法中是最高的，但训练时间比完全二叉树算法要稍长些。

4 结 论

在现有支持向量机多类分类方法研究的基础上，重点研究了二叉树分类方法。二叉树的结构对分类性能，包括分类时间与分类精度都有影响。基于将最容易分离出来的类最先分割出来的原则，文中提出一种基于超球体支持向量机的不完全二叉树多类分类算法，并进行了实验仿真。通过分析比较各种分类方法的性能，验证了该算法的有效性。

参考文献：

[1] ZHANG W B, CAI Q, WANG H J, et al. Application of harmonic wavelet package to feature extraction in impulsive signal [C]//Proceedings of the 2nd International Congress on Image and Signal Processing, Oct. 17-19, 2009. Tianjin, China: IEEE, 2009: 1-3.

[2] 朱凤明，樊明龙. 混沌粒子群算法对支持向量机模型参数的优化[J]. 计算机仿真，2010，27(11)：183-186.
ZHU FENG-MING, FAN MING-LONG. Chaos particle swarm optimization algorithm for optimizing the parameter of SVM [J]. Computer Simulation, 2010, 27(11): 183-186.

[3] DOSHI R A, KING R L, LAWRENCE G W. Wavelet-SOM in feature extraction of hyperspectral data for classification of nematode species [C]// Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, July 23-28, 2007. Barcelona, Spain: IEEE, 2007: 2818-2821.

[4] 唐浩，屈梁生. 基于支持向量机的发动机故障诊断[J]. 西安交通大学学报，2007，41(9)：1124-1126.
TANG HAO, QU LIANG-SHENG. Fault diagnosis of engine based on support vector machine [J]. Journal of Xi'an Jiaotong University, 2007, 41(9): 1124-1126.

[5] 付阳，李昆仑. 支持向量机模型参数选择方法综述[J]. 电脑知识与技术，2010，6(28)：8081-8082.
FU YANG, LI QUN-LUN. A survey of model parameters selection method for support vector machines [J]. Computer Knowledge and Technology, 2010, 6(28): 8081-8082.

[6] RIFKIN R M, KLAUTAU A. In defense of one-vs-all classification [J]. Journal of Machine Learning Research, 2004, 5: 101-141.

[7] 王晓锋. 一种改进的有向无环图支持向量机分类算法[J]. 重庆交通大学学报：自然科学版，2009，28(5)：973-975.
WANG XIAO-FENG. An improved SVM multiclass classification algorithm based on DAG [J]. Journal of Chongqing Jiaotong University: Natural Science, 2009, 28(5): 973-975.

[8] 孟媛媛，刘希玉. 一种新的基于二叉树的SVM多类分类方法[J]. 计算机应用，2005，25(11)：2653-2657.
MENG YUAN-YUAN, LIU XI-YU. A new SVM multiclass classification based on binary tree [J]. Journal of Computer Applications, 2005, 25(11): 2653-2657.

[9] 王艳，陈欢欢，沈毅. 有向无环图的多类支持向量机分类算法[J]. 电机与控制学报，2011，15(4)：85-89.
WANG YAN, CHEN HUAN-HUAN, SHEN YI. Multi-class support vector machine based on directed acyclic graph [J]. Electric Machines and Control, 2011, 15(4): 85-89.

[10] HUANG R Q, XI L F, LI X L, et al. Residual life predictions for ball bearings based on self-organizing map and back propagation neural network methods[J]. Mechanical Systems and Signal Processing, 2007, 21(1): 193-207.

[11] JOSHI A J, PORIKLI F, PAPANIKOLOPOULOS N. Multiclass active learning for image classification[C]// Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009. Miami, FL, USA: IEEE, 2009: 2372-2379.

[12] HOI S C H, JIN R, LYU M R. Batch mode active learning with applications to text categorization and image retrieval [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1233-1248.

（下转第140页）

Transactions on Industrial Electronics, 2006, 53(2): 631-639.

[7] RUBI J, RUBIO A, AVELLO A. Swing-up control problem for a self-erecting double inverted pendulum [J]. IEEE Proceedings of Control Theory and Applications, 2002, 149(2): 169-175.

[8] DEVASIA S, CHEN D G, PADEN B. Nonlinear inversion-based output tracking[J]. IEEE Transactions on Automatic Control, 1996, 41(7): 930-942.

[9] ZHONG W, ROCK H. Energy and passivity based control of the double inverted pendulum on a cart[C]// Proceedings of the 2001 IEEE International Conference on Control Application. Sept 5-7, 2001. Mexico: IEEE Press, 2001: 896-901.

[10] 李祖枢，王育新，张华，等. 小车二级摆系统的摆起倒立控制与实践[C]. 杭州：第5届全球智能控制与自动化大会，2004：2360-2364.

[11] LI Z S, DAN Y H, WEN Y L, et al. Swinging-up and handstand control of cart- triple-pendulum system based on human simulated intelligent control theory [J]. Journal of Huazhong University of Science and Technology, 2004(S1): 1-6.

[12] YAMAKITA M, IWASHIRO M, SUGAHARA Y, et al. Robust swing-up control of double pendulum[C]//Proceedings of the American Control Conference. June 21-23, 1995, Seattle, WA, USA. [S.l.]: IEEE Press, 1995, 1: 290-295.

[13] 李祖枢，但远宏，张小川，等. 双摆机器人摆杆平衡态任意转换运动控制的实现[J]. 自动化学报，2010，36(12)：1720-1731.
LI ZU-SHU, DAN YUAN-HONG, ZHANG XIAO-CHUAN, et al. Fulfillment of arbitrary transfer movement control between equilibrium statuses for a double pendulum robot[J]. Acta Automatica Sinica, 2010, 36(12): 1720-1731.

[14] 李祖枢，张华，古建功，等. 3关节单杠体操机器人的动力学参数辨识[J]. 控制理论与应用，2008，25(2)：242-246，252.
LI ZU-SHU, ZHANG HUA, GU JIAN-GONG, et al. Dynamic parameter identification of three-link acrobot on horizontal bar [J]. Control Theory & Applications, 2008, 25(2): 242-246, 252.

[15] 涂序彦. 人工智能：回顾与展望[M]. 北京：科学出版社，2006：174-207.

（编辑 侯 湘）

（上接第128页）

[13] SONG F X, ZHANG D P, YANG J Y, et al. A multiple maximum scatter difference discriminant criterion for facial feature extraction [J]. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2007, 37(6), 1599-1606.

[14] 谢志强，高丽，杨静. 基于球结构的完全二叉树SVM多类分类算法[J]. 计算机应用研究，2008，25(11)：3268-3274.
XIE ZHI-QIANG, GAO LI, YANG JING. SVM multi-class classification algorithm based on full-binary tree of sphere-structured [J]. Application Research of Computers, 2008, 25(11): 3268-3274.

[15] TAX D M J, DUIN R P W. Support vector domain description [J]. Pattern Recognition Letters, 1999, 20(11/13): 1191-1199.

[16] 陈荣，曹永锋，孙洪. 基于主动学习和半监督学习的多类图像分类[J]. 自动化学报，2011，37(8).
CHEN RONG, CAO YONG-FENG, SUN HONG. Multi-class image classification with active learning and semi-supervised learning [J]. Acta Automatica Sinica, 2011, 37(8): 954.

[17] 孙永奎，陈光禹，李辉. 支持向量机在模拟电路故障诊断中应用[J]. 电子测量与仪器学报，2008，22(2)：72-75.
SUN YONG-KUI, CHEN GUANG-JU, LI HUI. Support vector machine for analog circuit fault diagnosis [J]. Journal of Electronic Measurement and Instrument, 2008, 22(2): 72-75.

[18] UCI Machine Learning Repository [DB/OL]. http://archive.ics.uci.edu/ml.

（编辑 侯 湘）