第7卷第2期　　空军工程大学学报(自然科学版)　　Vol.7 No.2
2006年4月　　JOURNAL OF AIR FORCE ENGINEERING UNIVERSITY(NATURAL SCIENCE EDITION)　　Apr.2006

改进的SVM决策树分类算法

史朝辉，王晓丹，赵士敏，杨建勋
(空军工程大学 导弹学院，陕西 三原 713800)

摘　要:为解决多类分类问题,在分析SVM决策树分类器及存在问题的基础上,通过引入类间可分离性测度,并将其扩展到核空间,提出一种改进的SVM决策树分类器。实验表明了该分类算法对提高分类正确率的有效性。

关键词:支持向量机;SVM决策树;可分离性测度;核空间
中图分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1009-3516(2006)02-0032-04

SVM本质上是2值分类[1-2],而实际的目标分类通常是多值分类问题。为了构造有效的SVM多值分类器,除采用一对一、一对多方法及其改进外,将SVM和二叉决策树结合起来,构成SVM决策树[3],也是一种有效的方法。本文分析了SVM决策树分类器及存在问题,基于待分类的分布,定义了一种基于类分布的类间分离性测度,并将其扩展到核空间从而给出了一种改进的SVM决策树算法。

1　SVM决策树

与通常的方法相比,SVM决策树方法对于一个N值分类问题,需要寻找$N-1$个最优分类面。随着训练的进行,需要的训练样本数逐渐减少,因此,在训练阶段,随着训练的进行,生成最优分类面所需要的训练时间逐渐减少。在分类阶段,该方法并不像通常的方法需要计算所有分类决策函数的值,它仅需要根据决策树的结构,计算所需要的分类决策函数值。

该方法的缺点是如果在某个结点上发生分类错误,则会把分类错误延续到该结点的后续下一级结点上。因此,分类错误在越靠近树根的地方发生,分类性能就越差。为构造性能良好的决策树结构,可以考虑:将容易分(不易产生错分)的类先分割出来,然后再分不容易分的类。这样,就能够使可能出现的错分尽可能地远离树根。

2　类间分离性测度

根据训练数据估计各类间易分性,通常的做法是用类间的Euclidean距离作为分离性测度[4],但这种方法的缺点在于类中心间的距离远近并不能够代表类间的分离度,如图1所示,图1(a)、图1(b)中所代表的类间的距离均相等,但是很明显图1(b)中的两类要比图1(a)中的两类容易分。

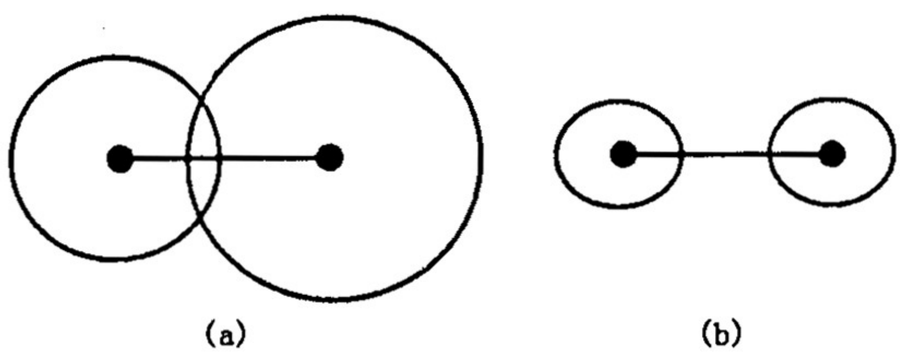

图1　类中心间距离相等类间易分析的比较

基于上述分析,定义了一种基于类分布的类间分离性测度。

收稿日期:2005-04-13
基金项目:陕西省自然科学基金资助项目(2004F36)
作者简介:史朝辉(1974-),男,河北博野人,讲师,硕士生,主要从事智能信息处理、模式识别、支持向量机研究;
王晓丹(1966-),女,陕西汉中人,教授,博士生导师,主要从事智能信息处理、目标识别、支持向量机研究.

假设要进行 k 类分类,训练样本集由类 $X_i, i=1,2,\cdots,k$ 组成。

定义1　类 i 和类 j 间的分离性测度 sm_{ij} 为

$$\mathrm{sm}_{ij} = \frac{d_{ij}}{(\sigma_i + \sigma_j)} \tag{1}$$

其中,$d_{ij}(i,j = 1,2,\cdots,k)$ 表示类 i 和类 j 中心间的距离:

$$d_{ij} = \| \boldsymbol{c}_i - \boldsymbol{c}_j \| \tag{2}$$

$\boldsymbol{c}_i$ 是根据训练样本计算出的类中心:

$$\boldsymbol{c}_i = \frac{1}{n_i}\sum_{x \in X_i} \boldsymbol{x} \quad , \quad i = 1,2,\cdots,k \tag{3}$$

记 n_i 为类 X_i 中的样本个数; σ_i 是类方差,表明了待分类的分布:

$$\sigma_i = \frac{1}{n_i - 1}\sum_{x \in X_i} \| \boldsymbol{x}_i - \boldsymbol{c}_i \| \quad , \quad i = 1,2,\cdots,k \tag{4}$$

若 $\mathrm{sm}_{ij} \geqslant 1$,则类 i 和类 j 间无交叠;若 $\mathrm{sm}_{ij} < 1$,则类间有交叠。sm_{ij} 的值越大,则类 i 和类 j 间的分离性越好。

定义2　类 i 的分离性测度:

$$\mathrm{sm}_i = \min_{\substack{j=1,2,\cdots,k \\ j \neq i}} (\mathrm{sm}_{ij}) \tag{5}$$

类 i 的分离性测度表明了类 i 与其余类间的分离性,将类 i 与其余各类间的最小分离性测度作为该类的分离性测度。

定义3　最易分的类:

$$s = \arg \max_{i=1,\cdots,k} (\mathrm{sm}_i) \quad , \quad i = 1,2,\cdots,k \tag{6}$$

即分离性测度值最大的类是最易分的类。

基于上述定义,在线性情况下,定义中的距离采用 Euclidean 距离即可。受文献[5]启发,在非线性情况下,采用非线性映射 Φ 把输入空间映射到某一特征空间 H 后这两点间的距离可采用如下表示:

引理1　已知两个向量 z_1 和 z_2,经非线性映射 Φ 作用,映射到特征空间 H,则这两个向量在特征空间的 Euclidean 距离为

$$d^{\mathrm{H}}(z_1,z_2) = \sqrt{K(z_1,z_1) - 2K(z_1,z_2) + K(z_2,z_2)} \tag{7}$$

其中 $K(\cdot,\cdot)$ 是核函数。这里需要指出的是输入空间样本的中心经映射后得到的值不再是特征空间中样本的中心。特征空间样本的中心向量 $\boldsymbol{m}_\Phi$ 要在特征空间中求得:

$$\boldsymbol{m}_\Phi = \frac{1}{n}\sum_{i=1}^{n} \Phi(\boldsymbol{x}_i) \tag{8}$$

其中 n 是样本的个数。因为不知道映射 $\Phi(\boldsymbol{x}_i)$ 的具体表达形式,所以无法根据式(8)求类中心向量。

引理2　已知两类模式的训练样本分别为 $\{x_1,x_2,\cdots,x_{n_1}\}$ 和 $\{x'_1,x'_2,\cdots,x'_{n_1}\}$,设经非线性映射 Φ 作用,映射到特征空间 H 后,类中心分别为 $\boldsymbol{m}_\Phi$ 和 $\boldsymbol{m}'_\Phi$,则在特征空间中 $\boldsymbol{m}_\Phi$ 和 $\boldsymbol{m}'_\Phi$ 间的距离为

$$d^{\mathrm{H}}(\boldsymbol{m}_\Phi,\boldsymbol{m}'_\Phi) = \sqrt{\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} K(\boldsymbol{x}_i,\boldsymbol{x}_j) - \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K(\boldsymbol{x}_i,\boldsymbol{x}'_j) + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} K(\boldsymbol{x}'_i,\boldsymbol{x}'_j)} \tag{9}$$

引理3　已知模式的训练样本为 $\{x_1,x_2,\cdots,x_n\}$,经非线性映射 Φ 作用后,在特征空间 H 中,训练样本 x 到类中心 $\boldsymbol{m}_\Phi$ 的距离为

$$d^{\mathrm{H}}(\boldsymbol{x},\boldsymbol{m}'_\Phi) = \sqrt{K(\boldsymbol{x},\boldsymbol{x}_i) - \frac{2}{n}\sum_{i=1}^{n_1} K(\boldsymbol{x},\boldsymbol{x}_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K(\boldsymbol{x}_i,\boldsymbol{x}_j)} \tag{10}$$

进而求得特征空间中的类方差:

$$\sigma^{\mathrm{H}} = \frac{1}{n-1}\sum_{i=1}^{n} d^{\mathrm{H}}(\boldsymbol{x}_i,\boldsymbol{m}_\Phi) \tag{11}$$

由此,对非线性的训练样本集,经非线性映射 Φ 作用后,在特征空间 H 中类 i 和类 j 间的分离性测度定义如下:

$$\mathrm{sm}_{ij}^{\mathrm{H}} = \frac{d^{\mathrm{H}}(\boldsymbol{m}_\Phi^i,\boldsymbol{m}_\Phi^j)}{(\sigma_i^{\mathrm{H}} + \sigma_j^{\mathrm{H}}} \tag{12}$$

3 算法描述

基于上述分析和定义,可对传统的 SVM 决策树进行改进,从而构造出性能良好的决策树结构,描述如下:

设:在各树结点生成的最优分类面是将一类和其余类分开。

假设要进行 k 类分类,训练样本集 X 由类 $X_i, i=1,2,\cdots,k$ 组成。

Step1 由式(12)计算特征空间中的分离性测度 $\mathrm{sm}_{ij}^{\mathrm{H}}$,组成分离性测度矩阵 $\mathbf{SM}^{\mathrm{H}}$:

$$\mathbf{SM}^{\mathrm{H}}=\begin{bmatrix} \mathrm{Inf} & \mathrm{sm}_{12}^{\mathrm{H}} & \cdots & \mathrm{sm}_{1,k}^{\mathrm{H}} \\ \mathrm{sm}_{21}^{\mathrm{H}} & \mathrm{Inf} & \cdots & \mathrm{sm}_{2,k}^{\mathrm{H}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{sm}_{k,1}^{\mathrm{H}} & \mathrm{sm}_{k,2}^{\mathrm{H}} & \cdots & \mathrm{Inf} \end{bmatrix}$$

为了编程方便,将同类间的分离性测度设为无穷大 Inf。设计数器 $t=k$;

Step2 选择当前 X 中最易分的类,假设类号为 i,为了后续训练的需要,将分离性测度矩阵中 i 行、i 列设为 Inf;

Step3 将当前最易分的类与其余各类进行分类训练,得到最优分类面,构成树结点;

Step4 $X=X-X_i, t=t-1$;

Step5 若 $t>1$,转 Step2;否则,结束。

4 实验结果与分析

为验证改进后的 SVM 决策树算法的有效性,使用三螺旋线和 UCI 数据库中的 wine 数据集,对改进 SVM 决策树与传统 SVM 决策树进行了比较实验。

4.1 三螺旋线

三螺旋线问题是一个 3 类划分问题,是公认的检验多类学习算法能力的"试金石"。该问题的分类要求是把 $x-y$ 坐标平面上 3 条不同螺旋线上的点正确地分开。螺旋线的平面坐标形式可用参数方程表示如下:

$$x=(k\theta+\alpha)\cos\theta \tag{13}$$

$$y=(k\theta+\alpha)\sin\theta \tag{14}$$

其中 k 和 α 都是常量,分别代表速度和起始距离。θ 是以弧度为单位的相角。三螺旋线有 6 个参数 $k_1,k_2,k_3,\ \alpha_1,\alpha_2,\alpha_3$ 都是待设定的。本实验中,令 $k_1=k_2=k_3,\alpha_1=1,\alpha_2=16,\alpha_3=32$,这是 3 条速度相同,起始位置不同的螺旋线。在式(13)和式(14)中取 $\theta\in[0,6\pi]$,即取三螺旋线的 3 个周期的点,采用平均采样,每条线采样 240 点,共计 720 个样本点(如图 2 所示)。

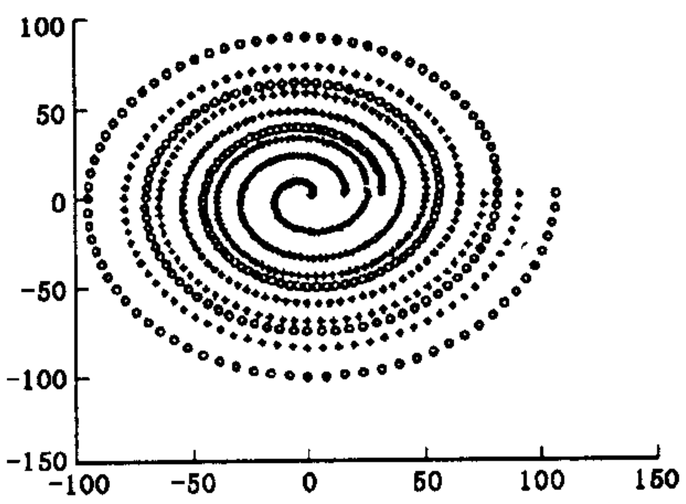

图 2 3 个周期内的三螺旋线

实验分 3 次进行,每次随机选取每条线的 160 个点,共计 480 个作为训练样本;每条线的其余 80 个点,共计 240 个作为测试样本。两种方法对比时,训练 SVM 在相同条件下进行,分类正确率取 3 次实验的平均值。表 1 给出了两种方法对三螺旋线分类的平均正确分类率比较结果。

表 1 三螺旋线实验比较结果

C	参数	传统 SVM 决策树分类正确率			改进 SVM 决策树分类正确率		
		线 1	线 2	线 3	线 1	线 2	线 3
1 000	$\sigma=15$	97.93%	94.20%	95.87%	98.77%	95.03%	96.7%
1 000	$\sigma=10$	98.37%	94.20%	92.93%	98.37%	94.20%	95.43%
1 000	$\sigma=5$	95.83%	92.93%	91.30%	96.27%	93.37%	92.93%

4.2 wine 数据集

wine数据集的样本特征维数为13,总数178,分3类(类1样本数59,类2样本数71,类3样本数48)。

实验分3次进行,每次随机选取每类样本的2/3作为训练样本;选取每类样本的1/3,作为测试样本。两种方法对比时,训练SVM在相同条件下进行,分类精度取3次实验的平均值。表2给出了两种方法对wine数据集分类的平均正确分类率比较结果。

表2　wine数据集实验比较结果

核函数	C	参数	传统SVM决策树分类正确率			改进SVM决策树分类正确率		
			线1	线2	线3	线1	线2	线3
RBF	100	$\sigma=5$	87.72%	72.46%	81.25%	89.47%	73.91%	83.33%
RBF	100	$\sigma=40$	89.47%	81.16%	68.75%	89.47%	82.61%	72.92%
RBF	100	$\sigma=90$	89.47%	88.40%	89.58%	91.23%	88.40%	93.75%

上述两组实验结果表明了改进后的SVM决策树算法对平均正确分类率的改善。

5　结束语

通过引入类间可分离性测度,并将其扩展到核空间,改进的SVM决策树算法可最大程度的减少积累误差,提高推广能力。该方法中选择更好的类间可分离性测度是一个有待深入研究的问题。

参考文献:

[1] Vapnik V. The Nature of Statistical Learning Theory[M]. New York:Springer - Verlag, 1995.

[2] 王晓丹,王积勤. 支持向量机研究与应用[J]. 空军工程大学学报(自然科学版),2004,5(3):49 - 55.

[3] Bennett K P, Blue J A. A Support Vector Machine Approach to Decision Trees[A]. Proceedings of IJCNN98[C]. Anchorage Alaska:IEEE Press,1998. 2396 - 2401.

[4] Fumitake Takahashi,Shigeo Abe. Decision - Tree - Based Multi - Class Support Vector Machines[A]. Proceeding of ICONIP02[C],Singapore:IEEE Press,2002. 1419 - 1422.

[5] 焦李成,张　莉,周伟达. 支撑矢量预选取的中心距离比值法[J]. 电子学报,2001,29(3):383 - 386.

(编辑:田新华)

An Improved Algorithm for SVM Decision Tree

SHI Zhao - hui, WANG Xiao - dan, ZHAO Shi - min, YANG Jian - xun

(The Missile Institute, Air Force Engineering University, Sanyuan, Shaanxi 713800, China)

Abstract: For the multi - class classification with Support Vector Machines (SVMs), a decision tree architecture has been proposed for computational efficiency. But by SVM decision tree, the generalization ability depends on the tree structure. In this paper, to improve the generalization ability of SVM decision tree, a novel separability measure is defined based on the distribution of the training samples in the kernel space, and an improved SVM decision tree is provided. The theoretical analysis and experimental results show that this algorithm has higher generalization ability.

Key words: support vector machine; SVM decision tree; separability measure; the kernel space