支撑矢量预选取的中心距离比值法

焦李成,张 莉,周伟达
(西安电子科技大学雷达信号处理重点实验室,西安 710071)

摘 要: 支撑矢量机为小样本模式识别提供了一新的途径,但其支撑矢量的选择相当困难,也成为其应用的瓶颈问题. 对此,本文提出了一种能够预先选取支撑矢量的方法——中心距离比值法. 该方法在不影响支撑矢量机的分类能力情况下,大大地减少了训练样本,提高了支撑矢量机的训练速度. 文中给出的仿真实验结果也验证了该方法的有效性和可行性. 类似的结果在国内外还未见报导.

关键词: 支撑矢量机; 中心距离比值法; 边界矢量; 支撑矢量
中图分类号: TP391.4 **文献标识码**: A **文章编号**: 0372-2112 (2001) 03-0383-04

Pre-extracting Support Vectors for Support Vector Machine

JIAO Li-cheng, ZHANG Li, ZHOU Wei-da
(National Key Lab. for Radar Signal Processing, Xidian Univ., Xi'an 710071, China)

Abstract: A new method called center distance ratio which is able to extract support vectors from given training examples is presened for support vector algorithm. The method greatly reduces the training samples and so improves the speed of support vector machine, while the ability of support vector machine to classification is unaffected. Our experimental results show remarkable improvement of speed support vector to support our idea.

Key words: support vector machine; center distance ratio method; margin vector; support vector

1 引言

70 年代后期以来,Vapnik 等一直致力于统计学习理论的研究,他们针对传统模式识别存在的困难,提出了一类新的学习算法,即支撑矢量机[7]. 由于支撑矢量机是一种小样本学习方法,而且其推广能力强,所以支撑矢量机受到了越来越多的重视. 它已应用于模式识别[1],回归估计[2~4]等领域,从而成为一种通用的学习机.

在目前的支撑矢量机算法[1~5]中,都是在优化计算后才能得到支撑矢量,即优化过程中不仅包含了对支撑矢量的优化,也包括了对非支撑矢量的优化,这无疑大大增加了不必要的计算量. 如果能够预先选取支撑矢量,且优化计算只针对这些支撑矢量来进行,这将大大减少了计算量. 对此国内外已有文献还未曾考虑这一途径. 本文针对这一问题,提出了一种支撑矢量预选取的方法,称之为中心距离比值法. 该方法仅在训练样本中提取出一个包含了所有支撑矢量的最小的边界矢量集合. 在不影响支撑矢量机的分类能力情况下,本文方法大大地减少了训练样本,同时提高了支撑矢量机的训练速度. 文中给出的仿真实验也验证了该方法的有效性和可行性.

2 支撑矢量机

2.1 线性支撑矢量机

假设已知训练样本$\{(x_1, y_1), \cdots, (x_l, y_l)\}$,其中 $x \in R^N$, $y \in [-1, 1]$, R^N 表示输入模式的特征空间. 学习的目的就是寻找最优的判决函数 $f(x)$,使得对任意 x,有 $f(x) = y$. 假设训练样本是线性可分的,则存在一个 N 维矢量 w 和标量 b,使得下面的不等式约束成立:

$$w \cdot x_i + b \geq +1; \quad y_i = +1 \tag{1}$$

$$w \cdot x_i + b \leq -1; \quad y_i = -1 \tag{2}$$

支撑矢量算法的主要思想是:构造一个能够使得间隔(margin)最大的超平面,这里的间隔是指该超平面到最近的样本的距离,即 $2/\|w\|$. 对线性不可分的情况,Vapnik 和 Cortes(1995)引入了松弛变量

$$\xi_i \geq 0, \quad i = 1, \cdots, l \tag{3}$$

从而有下面的的凸优化问题

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + C \cdot \sum_{i=1}^{l} \xi_i \\ s.t. \quad & y_i((w \cdot x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, \cdots, l \end{aligned} \tag{4}$$

其中 $C > 0$ 是用户自定义的惩罚因子. 利用 Lagrange 乘子法,可以把式(4)变成其对偶形式,有

收稿日期:1999-12-27;修回日期:2000-09-13
基金项目:国家"863"计划(No. 863-306-06-06-1);国防预研基金资助项目

$$\max \quad W(\boldsymbol{\alpha}) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i y_i \alpha_j y_j (x_i \cdot x_j) \tag{5}$$

$$s.t. \quad \sum_{i=1}^{l}\alpha_i y_i = 0$$

$$\alpha_i \in [0, C], i = 1, \cdots, l$$

和

$$w = \sum_{i=1}^{l}\alpha_i y_i x_i \tag{6}$$

采用线性判决函数:

$$f(x) = \mathrm{sgn}(w \cdot x + b) \tag{7}$$

由式(6),得到:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i (x_i \cdot x) + b\right) \tag{8}$$

如果 $\alpha_i \neq 0$,则训练样本 x_i 就被称为是支撑矢量.

2.2 非线性支撑矢量机

在线性支撑矢量训练算法中,数据以点积形式$(x_i \cdot x_j)$出现.现在用非线性映射把输入空间映射到某一特征空间,记为 $\Phi: R^N \rightarrow H$. 如果存在一种核函数 K,使得

$$K(x_i, x_j) = (\Phi(x_i) \cdot \Phi(x_j)) \tag{9}$$

可以在特征空间中进行许多计算,而且并不需要知道具体的映射 Φ. 事实上一个函数只要满足 Mercer 条件[3],就可以被分解成式(9).

现在用核函数代替线性支撑矢量机中的点积形式,对偶规划即可为:

$$\max \quad W(\boldsymbol{\alpha}) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i y_i \alpha_j y_j K(x_i \cdot x_j) \tag{10}$$

$$\text{s.t.} \quad \sum_{i=1}^{l}\alpha_i y_i = 0$$

$$\alpha_i \in [0, C], i = 1, \cdots, l$$

非线性支撑矢量机的判决函数为:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(x_i, x) + b\right) \tag{11}$$

上面提到,当 $\alpha_i > 0$ 时 x_i 是支撑矢量. 但是 α_i 值要在进行二次规划后才能得到,也就是说支撑矢量不能事先知道. 显然有大量的时间浪费在对非支撑矢量的计算上(支撑矢量机训练过程中庞大的计算量已成为其应用的瓶颈问题). 如果能够先验地选取支撑矢量,那么就可在不降低分类能力的基础上,进一步减少训练样本个数,从而大大提高支撑矢量机的训练速度.

3 支撑矢量预选取方法 ——中心距离比值法

支撑矢量从物理意义上来说,就是在样本线性可分的空间中两类的交遇区中,那些靠得最近但是又属于不同类的样本[6]. 与支撑矢量机相对应,这里我们给出的支撑矢量预选取方法也分为线性的和非线性两种形式,分别用于线性和非线性支撑矢量机.

3.1 线性中心距离比值法

首先给出本文涉及到的一些定义.

定义 1 某一类样本的平均特征称为该类样本的中心 m,已知样本向量组$\{x_1, x_2, \cdots, x_n\}$,那么其中心为:

$$m = \frac{1}{n}\sum_{i=1}^{n}x_i \tag{12}$$

定义 2 两个样本之间的特征差异称为样本距离,已知两个 N 维样本向量 x_1, x_2,其样本距离为:

$$d(x_1, x_2) = \| x_1 - x_2 \|_2 = \sqrt{\sum_{i=1}^{N}(x_1^i - x_2^i)^2} \tag{13}$$

定义 3 中心距离指的是各样本到中心的距离. 假设有两类模式的 N 维训练样本矢量分别为$\{x_1, x_2, \cdots, x_n\}$和$\{x_1', x_2', \cdots, x_n'\}$,其中心分别为 m_x 和 $m_{x'}$,则中心距离有四个,分别是 x 到 m_x 的距离 d_{xx},x 到 $m_{x'}$ 的距离 $d_{xx'}$,x' 到 m_x 的距离 $d_{x'x}$,以及 x' 到 $m_{x'}$ 的距离 $d_{x'x'}$;其中 d_{xx}和 $d_{x'x'}$称为是自中心距离,$d_{xx'}$ 和 $d_{x'x}$称为是互中心距离.

$$d_{xx}(x, m_x) = \sqrt{\sum_{i=1}^{N}(x^i - m_x^i)^2} \tag{14}$$

$$d_{xx'}(x, m_{x'}) = \sqrt{\sum_{i=1}^{N}(x^i - m_{x'}^i)^2} \tag{15}$$

$$d_{x'x}(x', m_x) = \sqrt{\sum_{i=1}^{N}((x')^i - m_x^i)^2} \tag{16}$$

$$d_{x'x'}(x', m_{x'}) = \sqrt{\sum_{i=1}^{N}((x')^i - m_{x'}^i)^2} \tag{17}$$

定义 4 中心距离比值(Ratio):已知两类模式,某一类模式的自中心距离到其互中心距离的比值称为该类的中心距离比值:

$$Ratio_x = d_{xx}/d_{xx'} \tag{18}$$

$$Ratio_{x'} = d_{x'x'}/d_{x'x} \tag{19}$$

定义 5 边界矢量是某一类模式中,位于其边界上的那些矢量.

定理 1 已知两类模式$\{x_1, x_2, \cdots, x_n\}$和$\{x_1', x_2', \cdots, x_n'\}$及它们的中心距离比值. 设定两个阈值 r_x 和 $r_{x'}$,则集合$\{x_i \mid Ratio_x(i) > r_x, i = 1, \cdots, n\}$就是模式$\{x_1, x_2, \cdots, x_n\}$的边界矢量;集合$\{x_i' \mid Ratio_{x'}(i) > r_{x'}, i = 1, \cdots, n\}$就是另一类的边界矢量.

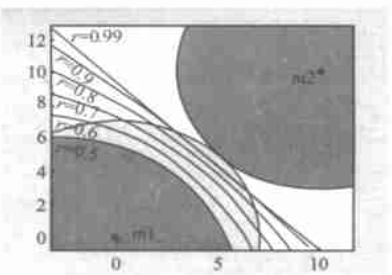

图 1 两类模式的自中心距离相差不大的情况下,边界矢量的分布

下面举一个简单的例子来证明定理 1. 如图 1 和 2,m_1 和 m_2 分别是两类模式 C_1 和 C_2 的中心. 假设 C_1 和 C_2 的样本分别分布在半径为 L_1 和 L_2 的圆内. L_1(或 L_2)也可以看作是最大的自中心距离. 其中 r 是阈值,它的取值决定了边界矢量的选取. 图中的阴影部分为原训练样本的分布区域,浅色阴影

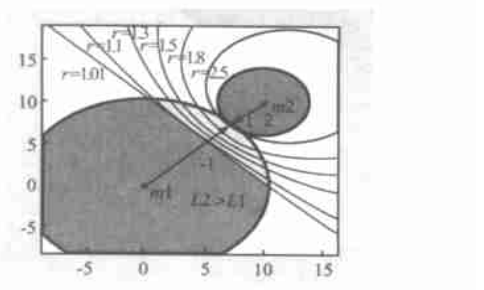

图 2　两类模式的自中心距离相差较大的情况下,边界矢量的分布

部分是 r 为不同值时选中的边界矢量分布区域. 当 L_1 和 L_2 差不多大小时, r 的取值范围一般为小于 1,如图 1 所示,当 r 值越接近于 1,边界矢量的分布区域越小. 而 L_1 和 L_2 的值相差很大时, r 的取值就有可能大于 1,如图 2 所示, r 值较小时,边界矢量的分布区域就比较大. 从这两幅图可以看出,阈值 r 的取值,对边界矢量的分布区域的大小是至关重要的. 如果阈值选定合适的话,边界矢量集合就是包含了支撑矢量集合的最小集合,更有甚者边界矢量集合就是支撑矢量集合.

3.2　非线性中心距离比值法

对非线性可分的模式,采用非线性映射 Φ 把输入空间映射到某一特征空间 H. 输入空间的两个矢量 z_1 和 z_2 之间的距离可以用 Euclidean 距离 $\| z_1 - z_2 \|_2$ 来表示,那么映射到特征空间后这两点间的距离该如何表示呢?

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

$$d^H(z_1, z_2) = \sqrt{K(z_1, z_1) - 2K(z_1, z_2) + K(z_2, z_2)} \tag{20}$$

其中 $K(\cdot, \cdot)$ 正是 2.2 节中提到的核函数. 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矢量 m_Φ 要在特征空间中求得:

$$m_\Phi = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i) \tag{21}$$

其中 n 是样本的个数. 因为不知道映射 $\Phi(x)$ 的具体表达形式,所以无法根据式(21)来求样本中心矢量. 对此本文给出如下引理.

引理 2　已知两类模式的训练样本分别为 $\{x_1, x_2, \cdots, x_n\}$ 和 $\{x'_1, x'_2, \cdots, x'_n\}$, $x_i, x'_i \in R^N$, $i = 1, \cdots, n$,经非线性映射 Φ 作用后,映射到某一特征空间 H,则在特征空间中的自中心距离为:

$$d^H_{xx}(x, m_\Phi) = \sqrt{K(x, x) - \frac{2}{n}\sum_{i=1}^{n}K(x, x_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i, x_j)} \tag{22}$$

$$d^H_{x'x'}(x', m_\Phi) = \sqrt{K(x', x') - \frac{2}{n}\sum_{i=1}^{n}K(x', x_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x'_i, x'_j)} \tag{23}$$

互中心距离为:

$$d^H_{xx}(x, m_\Phi) = \sqrt{K(x, x) - \frac{2}{n}\sum_{i=1}^{n}K(x, x_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i, x_j)} \tag{24}$$

$$d^H_{x'x'}(x', m_\Phi) = \sqrt{K(x', x') - \frac{2}{n}\sum_{i=1}^{n}K(x', x_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i, x_j)} \tag{25}$$

引理 1、2 的证明从略.

根据引理 2,可以直接求得原训练样本在特征空间中的中心距离,因此特征空间的中心距离比值为:

$$Ratio_{\Phi(x)} = d^H_{xx} / d^H_{x'x'} \tag{26}$$

$$Ratio_{\Phi(x')} = d^H_{x'x'} / d^H_{x'x'} \tag{27}$$

由于在经过 Φ 映射后,样本已是线性可分的,定理 1 在此同样成立. 由此可得定理 2.

定理 2　已知两类模式 $\{x_1, x_2, \cdots, x_n\}$ 和 $\{x'_1, x'_2, \cdots, x'_n\}$,经过 Φ 映射到 H 空间中,得到它们在 H 空间中的中心距离比值. 设定两个阈值 r_x 和 $r_{x'}$,则集合 $(x_i \mid Ratio_{\Phi(x)}(i) > r_x, i = 1, \cdots, n)$ 就是模式 $\{x_1, x_2, \cdots, x_n\}$ 的边界矢量集合. 集合 $(x'_i \mid Ratio_{\Phi(x')}(i) > r_{x'}, i = 1, \cdots, n)$ 就是另一类的边界矢量集合.

这样,我们就可以根据设定的阈值,提取出最小的包含了支撑矢量的边界矢量集合. 已知训练样本 $\{(x_1, y_1), \cdots, (x_l, y_l)\}$,用中心距离比值法选取边界矢量集合,然后把选取的边界矢量组成新的训练样本:$\{(x'_1, y'_1), \cdots, (x'_k, y'_k)\}$,此 $k < l$ 时;然后运用到支撑矢量机中.

4　仿真实验

例 1　线性可分的例子

随机产生两类线性可分的均匀分布的数据. 图 3 中"Δ"分别是两类模式的中心. 一个圈圈起来的点,就是在某一阈值下得到的边界矢量集,而两个圈圈起来的点就是支撑矢量,我们明确地可以看出 CDRM 选取的边界矢量和 SVA 得到的支撑矢量是完全一样的. 表 1 是两种算法的分类结果比较. 从表中,可以看出结合了中心距离比值法的支撑矢量机的分类效果和没有结合该算法的支撑矢量机的分类效果是完全一致的. 应该指出的是 CDRM + SVA 的运算量大大地少于 SVA 的运算量.

表 1　线性可分情况的分类结果

算法	训练样本个数	边界矢量个数	支撑矢量个数	训练样本识别率	检验样本个数	检验样本识别率
SVA	20	(未选)	3	C1 类:100 % C2 类:100 %	80	C1 类:100 % C2 类:100 %
CDRM + SVA	20	3	3	C1 类:100 % C2 类:100 %	80	C1 类:100 % C2 类:100 %

例 2　非线性可分的例子

随机产生两类非线性可分的数据. 这里采用的核函数是二次多项式核: $K(x_i, x_j) = (x_i, x_j)^2$,那么核函数的可以分解为 $(\Phi(x_i) \cdot \Phi(x_j))$. 非线性映射函数 Φ 为:

$$\Phi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2}\,x_1^2 x_2^2 \\ x_2^2 \end{pmatrix}$$

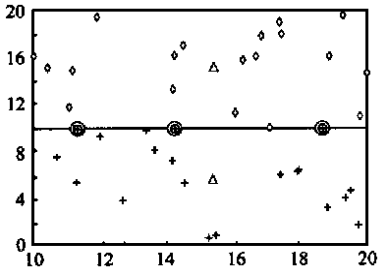

图 3 线性可分的例子

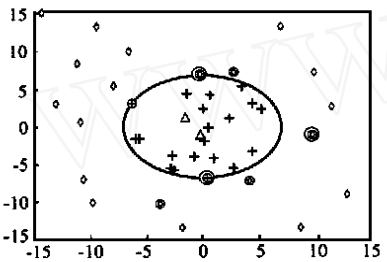

图 4 非线性可分的例子

图 4 中“Δ”符号分别表示两类模式的中心,其中被一个圈圈起来的点是边界矢量,被两个圈圈起来的点是支撑矢量. 可以看出 CDRM 得到边界矢量包括了 SV 算法得到的支撑矢量. 表 2 也证实了 CDRM + SVA 和 SVA 分类结果的一致性和本文提出算法的有效性.

表 2 非线性可分情况的分类结果

算法	训练样本个数	边界矢量个数	支撑矢量个数	训练样本识别率	检验样本个数	检验样本识别率
SVA	20	(未选)	3	C1 类:100 % C2 类:100 %	80	C1 类:100 % C2 类:98.75 %
CDRM + SVA	20	7	3	C1 类:100 % C2 类:100 %	80	C1 类:100 % C2 类:98.75 %

5 结论和讨论

本文提出了一种预先选取支撑矢量的方法 ——中心距离比值法. 这一方法能够提取靠近边界上的点(即边界矢量),在阈值适当的情况下,就能完全地提取出所需要的支撑矢量. 在此情况下,用比训练样本少得多的边界矢量来进行训练,大大地加快了支撑矢量机的训练速度,而且保证了支撑矢量机的分类能力不会受到影响. 仿真实验验证了该方法的可行性和有效性,从而为支撑矢量机的应用提供了一种有效的实用化方法.

本文虽然是围绕二分模式识别问题展开讨论的,但文中所提出的方法同样适用于多类识别问题.

参考文献:

[1] C.J.C. Burges. A tutorial on support vector machines for pattern recognition [C]. Data Mining and Knowledge Discovery, 1998, 2(2): 1-47.

[2] A. Smola. Regression estimation with support vector learning machines [D]. Master's Thesis. Tech. University of Mumchen, 1996. Available http://www.first.gmd.de/(smola

[3] A. Smola, B. Schölkopf. A tutorial on support vector regression [R]. NeuroCOLT, Rep. 19, 1998. Available http://svm.first.gmd.de

[4] B. Schölkopf, A.J. Smola, R. Willianson, P. Bartlett. New support vector algorithms [R]. NeuroCOLT2 Technical Report Series, 1998. Available http://www.neurocolt.com

[5] B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, V. Vapnik. Comparing support vector machines with Gaussian kernels to radial basis function classifiers [J]. IEEE Transactions on Signal Processing, 1997, 45(11).

[6] 阎平凡. 对多层前向神经网络研究的进一步看法[J],电子学报, 1999, 27(5): 82-85.

[7] V. Vapnik. The Nature of Statistical Learning Theory [M]. New York: Springer. 1995.

作者简介:

焦李成 1959 年出生. 1984 年和 1990 年在西安交通大学研究生院分别获硕士学位和博士学位,现为西安电子科技大学教授,博士生导师. 主要研究领域包括:非线性理论、神经网络、智能信息处理、数据挖掘以及现代通讯.

张 莉 1975 年出生. 西安电子科技大学在读博士生. 主要研究方向有数据挖掘、模式识别和神经网络.