

基于中心距离比值的增量支持向量机

孔 波¹, 刘小茂¹, 张 钧²

(1. 华中科技大学 数学系, 湖北 武汉 430074;

2. 华中科技大学 图像信息处理与智能控制教育部重点实验室, 湖北 武汉 430074)

(kongbo666@163.com)

摘 要:研究了支持向量、中心距离比值、边界向量以及增量学习之间的关系,提出了基于中心距离比值的增量支持向量机。与传统方法相比,基于中心距离比值的增量支持向量机有效的利用了中心距离比值,解决了 CDRM + SVM 的阈值选取问题;且适合于增量学习;从而在保证支持向量机的分类能力没有受到影响的前提下提高了支持向量机的训练速度。

关键词:统计学习理论;支持向量机;中心距离比值;增量学习

中图分类号: TP18; TP391.4 **文献标识码:** A

Incremental support vector machine based on center distance ratio

KONG Bo¹, LIU Xiao-mao¹, Zhang Jun²

(1. Department of Math, Huazhong University of Science and Technology, Wuhan Hubei 430074, China;

2. Key Laboratory of Education Ministry For Image Processing and Intelligent Control,
Huazhong University of Science and Technology, Wuhan Hubei 430074, China)

Abstract: Although a Support Vector Machine (SVM) is applicable to a learning task with small training examples, all the training examples don't play an important role in the learning task, but a few ones called support vectors do. According to the relations of support vector, center distance ratio, margin vector and incremental learning, a new method called incremental support vector machine based on center distance ratio was presented. First of all, some support vectors were extracted by the method; then others were made up by the incremental learning method so all the support vectors were found. Compared to the CDRM + SVM, incremental support vector machine based on center distance ratio utilizes effectively center distance ratio and suits to incremental learning. So the new method improves the speed of SVM greatly, while the ability of SVM to classify is unaffected.

Key words: statistical learning theory; Support Vector Machine(SVM); center distance ratio; incremental learning

0 引言

支持向量机(Support Vector Machine, SVM)是 Vapnik 等人在统计学习理论^[1]的基础上发展起来的一种小样本学习理论,是数据挖掘^[2]中的一项新技术,是借助于最优化方法(尤其是二次规划)^[3]解决机器学习问题的新工具,与神经网络、遗传算法、人工智能等现有的机器学习方法相比,具有较好的推广能力和非线性处理能力,尤其是在处理高维数据时,有效地解决了“维数灾难”问题。现已广泛应用于模式识别和回归估计等问题中。

若是仅利用支持向量样本作为训练样本,则得到的判决函数与利用所有的训练样本得到的判决函数是一致的,且提高了训练速度。文献[4]中的 CDRM + SVM 方法正是基于这一有点提出的,该方法是在优化训练样本前利用中心距离比值方法(CDRM)预抽取出所有的支持向量样本。但是该方法在实际操作时涉及到两个阈值的选取,且阈值的选取与抽取的向量息息相关,若阈值选取过小,则抽取的边界向量过多;若阈值选取过大,则无法包括全部支持向量。因此由于阈值的选取缺乏一个良好的标准,导致该方法实用性不强。在文

献[5]中提出了一个经验公式来解决阈值的选取问题,即使这样也不能够保证把所有的支持向量样本都包含在边界向量样本中,因此阈值的确定仍是一个难点。在文献[6、7]中,提出了利用块算法和增量学习来解决初始块中未包含全部支持向量的方法。

本文受到上述中心距离比值法、块算法和增量学习算法的启发,结合支持向量占训练样本比例较少的基础上,通过分析距离比值对支持向量的影响,提出了基于中心距离比值的增量支持向量机。

1 支持向量机^[1~3]

在两类模式识别问题中,给定训练样本 $\{(x_i, y_i), x_i \in R^n, y_i \in \{-1, +1\}, i = 1, \dots, l\}$, SVM 的目标就是:通过训练样本构造一个判决函数,使训练样本以最大间隔分开,且将待判决样本尽可能地正确分类。分类规则为:

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

其中 $\text{sgn}(\cdot)$ 为符号函数。求决策函数需要构造如下优化问题(1)~(3)式:

收稿日期:2005-12-05;修订日期:2006-03-06 基金项目:国家自然科学基金资助项目(60373090);航天基金(02.1.3.jw0504)

作者简介:孔波(1980-),男,河南周口人,硕士研究生,主要研究方向:统计学习理论、模式识别; 刘小茂(1965-),女,湖南洞口人,副教授,博士,主要研究方向:统计学习理论、金融风险; 张钧(1966-),男,江西南昌人,副教授,主要研究方向:模式识别与图像处理。

$$\min_{w, \xi, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \quad (1)$$

$$s. t. \quad y_i[(w \cdot \phi(x_i)) + b] + \xi_i \geq 1 \quad (2)$$

$$\xi_i \geq 0, i = 1, \dots, l \quad (3)$$

则(1) ~ (3) 式的 Wolfe 对偶为:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^l \alpha_i \quad (4)$$

$$s. t. \quad \sum_{i=1}^l y_i \alpha_i = 0 \quad (5)$$

$$0 \leq \alpha_i \leq C, i = 1, \dots, l \quad (6)$$

设优化问题(4) ~ (6) 式有最优解 $\alpha^* = (\alpha_1^*, \dots, \alpha_l^*)$, 则由 KKT 条件可知:

$$y = \sum_{i=1}^l y_i \alpha_i^* K(x_i, x) + b^* \quad (7)$$

$$f(x) = \text{sgn}(\sum_{i=1}^l y_i \alpha_i^* K(x_i, x) + b^*) \quad (8)$$

$$b^* = y_j - \sum_{i=1}^l y_i \alpha_i^* K(x_i, x_j) \quad (9)$$

其中(9) 式中的 $j \in J = \{j | 0 < \alpha_j^* < C, i = 1, \dots, l\}$, (4) ~ (9) 式中的 $K(\cdot)$ 是满足 Mercer 条件的核函数,其作用就是将原输入空间的线性不可分问题转化为高维甚至无穷维 Hilbert 空间的线性可分问题,然后在该高维或无穷维空间求解最优化问题(4) ~ (6) 式。常用的核函数主要有:

线性核: $K(x_i, x_j) = (x_i \cdot x_j)$;

多项式核: $K(x_i, x_j) = ((x_i \cdot x_j) + 1)^d$;

径向基核: $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / \sigma^2)$;

Sigmoid 核: $K(x_i, x_j) = \tanh(\kappa(x_i \cdot x_j) + v)$ 。

在上述几个核函数中,线性核函数适合于在原输入空间线性可分问题;径向基核函数的计算复杂程度不随参数的变化而变化,在全部参数空间中都满足 Mercer 条件,是 SVM 中最常用的核函数;Sigmoid 核函数虽然不是正定核,但在某些实际应用中却非常有效。

2 基于中心距离比值的增量支持向量机

用决策函数进行分类时,并非所有的训练样本都对分类起作用,只有少量被称作支持向量的训练样本才起作用,并且这些支持向量在几何位置上分布于分划超平面的周围且包含在边界向量样本集合中。因此,若是仅用这些边界向量样本进行训练,则减少了训练样本的数量,从而提高了 SVM 的训练速度,且训练结果不受任何影响。

由于 CDRM + SVM 方法的阈值选取缺乏一个有效的方法,导致该方法的实用性不强。为了解决这个难点,我们对中心距离比值进行一次简单的排序,然后取中心距离比值较大的对应的样本点,即可抽取到包含支持向量样本在内的边界向量样本,从而解决了阈值选取的难点;最后为了避免支持向量选取不全的情形,又引入了增量学习算法,保证了训练样本包括所有的支持向量所对应的训练样本,从而在保证了支持向量机的分类能力没有受到影响的前提下大大提高了支持向量机的训练速度。

2.1 中心距离比值^[4]

在这里主要给出了算法中涉及到的一些定义,且只考虑特征空间,原空间只需使用线性核函数即可。

定义 1(距离) 给定两个训练样本 x_i, x_j , 则两样本在特征空间的距离可表示为:

$$d(\varphi(x_i), \varphi(x_j)) = \sqrt{K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j)}$$

定义 2(中心) 某一类样本的平均特征称为该类样本的中心。给定一类训练样本 $\{x_1, x_2, \dots, x_n\}$, 则其中心为 $m =$

$$\frac{1}{n} \sum_{i=1}^n \varphi(x_i)。$$

定义 3(中心距离) 中心距离就是样本到中心的距离。假设两类训练样本分别为:

$$X^+ = \{x_i | x_i \in R^n, y_i = 1, i = 1, 2, \dots, l^+\}$$

$$X^- = \{x_i | x_i \in R^n, y_i = -1, i = 1, 2, \dots, l^-\}$$

则其中心分别为 $m^+ = \sum_{i=1}^{l^+} \varphi(x_i)$ 和 $m^- = \sum_{i=1}^{l^-} \varphi(x_i)$ 。且对

每个正类样本点来说,都有两个中心距离:自中心距离 $D_{si} = d(\varphi(x_i), m^+)$ 和互中心距离 $D_{mi} = d(\varphi(x_i), m^-)$;同理,对每个负类样本点来说,也都有两个中心距离:自中心距离 $D_{si} = d(\varphi(x_i), m^-)$ 和互中心距离 $D_{mi} = d(\varphi(x_i), m^+)$ 。

定义 4(中心距离比值) 已知两类样本,某一类样本中样本点 x_i 的自中心距离和互中心距离的比值分别为 D_{si} 和 D_{mi} , 则该样本点的中心距离比值: $R_i = D_{si} / D_{mi}$ 。

2.2 边界向量样本集合

由于边界向量样本所对应的样本点是对应中心距离比值较大的样本点的原则,并且边界向量样本所对应的训练样本仅占有所有训练样本的一小部分(一般不足于 10%),且支持向量样本包含在边界向量样本集合内,因此选取了中心距离比值较大的大约 20%(该比率可以根据实际情况调节)的训练样本作为边界向量样本集合。即首先对每类样本中的所有样本点求出该样本点对应的中心距离比值,然后将该类中的中心距离比值进行排序,取中心距离比值较大的(大约 20%)所对应的训练样本点作为该类的边界向量样本集合;另一类也同样如此,即可抽取包含支持向量所对应训练样本在内的边界向量样本集合。这样既解决了阈值选取的难点,使中心距离比值得到了有效的利用。由此可得出如下定理 1。

定理 1(边界向量样本集合) 若已知正类训练样本集合 $X^+ = \{x_i | x_i \in R^n, y_i = 1, i = 1, 2, \dots, l^+\}$ 及其中心距离比值集合 $R^+ = \{R_i^+ | R_i^+ \in D_{si}^+ / D_{mi}^+, i = 1, \dots, l^+\}$, 对 R^+ 进行降序排列,并取前面大约 20%(该比率可调)的元素所对应的下标集 J^+ , 然后由下标集 J 可求出正类样本的边界向量样本集合 $X_{J^+}^+$, 同理可求出负类样本的边界向量样本集合 $X_{J^-}^-$ 。

2.3 基于中心距离比值的增量支持向量机

2.3.1 停机准则

由于 KKT 条件是最优化问题(4) ~ (6) 式最优解的充分必要条件,因此可选择 KKT 条件作为算法的一个停机准则,即:

$$\sum_{i=1}^l y_i \alpha_i^* = 0, 0 \leq \alpha_i^* \leq C, i = 1, \dots, l \quad (10)$$

$$b^* = y_j - \sum_{i=1}^l y_i \alpha_i^* K(x_i, x_j) \quad (11)$$

$$y_j(\sum_{i=1}^l y_i \alpha_i^* K(x_i, x_j) + b^*) \begin{cases} \geq 1, \{x_j | \alpha_j^* = 0\} \\ = 1, \{x_j | 0 < \alpha_j^* < C\} \\ \leq 1, \{x_j | \alpha_j^* = C\} \end{cases} \quad (12)$$

2.3.2 精度要求

已知训练样本 $T = \{(x_i, y_i), x_i \in R^n, y_i \in \{-1, +1\}\}$,

$i = 1, \dots, l$, 设训练样本点 x_j 到分划超平面的距离 $d = y_j(\sum_{i=1}^l y_i \alpha_i^* K(x_i, x_j) + b^*)$, 记 $d \geq 1$ 的训练样本点个数为 m 个, 则精度 $\varepsilon = m/l$, 表示训练样本的判决正确率。

算法: (基于中心距离比值的增量支持向量机, CDR-ISVM)

1) 给定训练样本集

$$T = \{(x_i, y_i), x_i \in R^n, y_i \in \{-1, +1\}, i = 1, \dots, l\}$$

由定理 1 预抽取出边界向量集合 X_{j+}^+, X_{j-}^- ;

2) 给定精度要求 ε , 令初始工作集 $W_0 = \{X_{j+}^+, X_{j-}^-\}$, 并记其对应样本点的下标集为 J_0 , 令 $k = 0$;

3) 对工作集 W_k 用二次规划算法求解二次规划问题:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i \in J_k} \sum_{j \in J_k} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i \in J_k} \alpha_i \\ \text{s. t.} \quad & \sum_{i \in J_k} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, i \in J_k \end{aligned}$$

得最优解 $\hat{\alpha}^{J_k}$;

4) 根据 $\hat{\alpha}^{J_k}$ 按下述方式构造 $\alpha^k = (\alpha_1^k, \dots, \alpha_l^k)^T$: 当 $j \in J_k$ 时, α_j^k 取为 $\hat{\alpha}^{J_k}$ 的相应分量; 当 $j \notin J_k$, $\alpha_j^k = 0$ 。检验 α^k 在精度 ε 内是否满足停机准则 (10) ~ (12) 式: 若 α^k 满足, 则可由 α^k 构造判决函数, 停止计算; 否则转入 (5);

5) 由 $\hat{\alpha}^{J_k}$ 确定的支持向量对应的样本点组成的集合 S_k , 在集合 $T \setminus S_k$ 中找出破坏条件:

$$y_j(\sum_{i \in J_k} y_i \alpha_i^{k+1} K(x_i, x_j) + b^*) \begin{cases} \geq 1, \{x_j \mid \alpha_j^{k+1} = 0\} \\ = 1, \{x_j \mid 0 < \alpha_j^{k+1} < C\} \\ \leq 1, \{x_j \mid \alpha_j^{k+1} = C\} \end{cases}$$

该条件的样本点集合 M_k , 用 M_k 和 S_k 中的样本点一起组成新的工作集 W_{k+1} , 其相应的下标集记为 J_{k+1} ;

6) 令 $k = k + 1$, 转 3)。

3 实验分析

鸢尾属植物数据集 (Iris data set) 是一个用来检验分类算法性能的标准数据集。该数据集共 150 个样本点, 分为三类: I: Iris-setosa, II: Iris-versicolor 和 III: Iris-virginica, 每类样本各有 50 个样本点。每个样本有四个特征: 萼片长度、萼片宽度、花瓣长度、花瓣宽度。依据我们提出的 CDR-ISVM 方法和一般的 C-SVM 方法, 我们利用上述 Iris 数据进行实验: 选用每类样本的前面 25 个为训练样本, 其余为测试样本。训练结果如表 1 (其中 Iris12 表示第一类和第二类的样本点, 余同)。

表 1 Iris 数据试验结果

样本	算法	训练时间 (秒)	训练 准确率%	测试 准确率%
Iris12	C-SVM	0.8220	100	100
	CDR-ISVM	0.0300	100	100
Iris23	C-SVM	0.8310	100	90
	CDR-ISVM	0.0800	98	88
Iris13	C-SVM	0.7980	100	100
	CDR-ISVM	0.0320	100	100

从上表中可以看出: 采用 CDR-ISVM 方法与 C-SVM 方法在对待同一个训练样本, 使用相同的核函数和惩罚参数的情况下, 由于前者 (CDR-ISVM 方法) 只使用了一小部分 (20%) 包含支持向量样本在内的边界向量样本集合作为训练样本, 而有效提高了支持向量机训练速度, 且训练和测试准确率基本没有改变。从而验证了 CDR-ISVM 方法的有效性和可行性: 即该方法有效的利用了中心距离比值, 且适合于增量学习, 从而在保证了对支持向量机的分类能力没有受到影响的前提下提高了支持向量机的训练速度。

4 结语

根据支持向量、中心距离比值、边界向量以及增量学习之间的关系, 本文提出了基于中心距离比值的增量支持向量机。该方法主要具有以下几个优点: 1) 有效的利用了中心距离比值, 解决了 CDRM + SVM 的阈值选取的难点; 2) 适合于增量学习, 添补预抽取的边界向量样本可能包括不了全部支持向量的漏洞; 3) 在不影响 SVM 的判决能力的前提下, 大大提高了 SVM 的训练速度。实验结果进一步印证了该方法的有效性和可行性。

参考文献:

- [1] VAPNIK N. 统计学习理论 [M]. 许建华, 张学工, 译. 北京: 电子工业出版社, 2004.
- [2] 邓乃扬, 田英杰. 数据挖掘中的新方法——支持向量机 [M]. 北京: 科学出版社, 2004.
- [3] 薛毅. 支持向量机与数学规划 [D]. 北京: 北京工业大学, 2003.
- [4] ZHANG L, ZHOU WD, JIAO LC. Pre-extracting Support vectors for support vector machine [A]. Proceeding of ICSP2000 [C]. IEEE, 2000. 1432 - 1435.
- [5] 张莉. 支撑向量机与核方法研究 [D]. 西安: 西安电子科技大学, 2002.
- [6] CORTES C, VAPNIK N. Support vector networks [J]. Machine Learning. 1995, 20(3), 273 - 297.
- [7] DOMENICONI C, GUNOPULOS D. Incremental Support Vector Machine Construction [J]. IEEE Trans. 2001. 589 - 593.

(上接第 1433 页)

- [5] LIN SL, TSAI YJ, LIOU CY. Conscious mental tasks and the EEG signals [J]. Medical & Biological Engineering & Computing, 1993, 31: 421 - 425.
- [6] KEERTHI S, LIN C-J. Asymptotic Behavior of Support Vector Machines with Gaussian Kernel [J]. Neural Computation, 2003, 15: 1667 - 1689.
- [7] BLANCO S, DPATTELLIS C, ISAACSON S, et al. Time-Frequency Analysis of Electroencephalograms series (II): Gabor and Wavelet Transforms [J]. Physical Review E, 1996, 54(6): 6661 - 6672.
- [8] HAMID EY, MARDIANA R, KAWASAKI ZI. Method for RMS and power measurements based on the wavelet packet transform [J]. IEEE proceedings-Science, Measurement and Technology, 2002, 149(2): 60 - 66.
- [9] 李颖洁, 朱贻盛, 陈兴时, 等. 精神分裂症脑电基于维数计算的时空复杂度测量 [J]. 中国生物医学工程学报, 2003, 22(1): 88 - 92.
- [10] 王蔚, 张胜, 宁新宝, 等. 精神分裂症患者脑电信号多重分形的异常 [J]. 中国生物医学工程学报, 2004, 23(6): 511 - 515.