一种基于距离比值的支持向量机增量训练算法

徐海龙， 王晓丹， 史朝辉， 华继学， 权 文
（空军工程大学 导弹学院，陕西 三原 713800）

摘 要：由于支持向量机具有较好地学习性能和泛化能力，目前已经得到了广泛的应用。如何使支持向量机进行有效的增量学习是目前支持向量机应用中需要解决的问题。深入研究了支持向量分布特点，提出了一种新的支持向量机增量训练淘汰机制——距离比值算法。该算法根据遗忘规则，设定一个合适的参数，按距离比值法中的定义计算各个样本中心距离与其到最优分类面距离的比值，舍弃对后续训练影响不大的样本，即可对训练数据进行有效的淘汰。对标准数据集的实验结果表明，使用该方法进行增量训练在保证分类精度的同时，能有效地提高训练速度。

关键词：支持向量机；增量训练；淘汰机制；边界矢量；距离比值算法
中图分类号：TP391.4 **文献标识码**：A **文章编号**：1009-3516(2008)04-0029-05

支持向量机(Support Vector Machines,SVM)[1]基于坚实、严谨的统计学习理论(Statistics Learning Theory, SLT)，比传统学习方法具有较好地学习性能和泛化能力，因而得到了广泛的应用[2]。但是在支持向量机增量训练中，现有的增量训练算法[3-4]在不同程度上存在一些问题，比如有些算法由于缺乏对训练数据有选择的淘汰机制，在很大程度上影响了分类精度，有些方法虽具备有效的淘汰机制，却需选择很多参数。

受文献[5-7]的启发，本文基于支持向量在样本空间的分布特性，提出了一种新的基于距离比值的增量训练算法。

1 支持向量分布规律研究

从几何理论上分析，通常样本空间呈聚集分布，或经一合适的非线性映射后在特征空间呈聚集分布。支持向量集只是样本集的一小部分，那么这部分样本在这个聚集区域是怎样分布的呢？为了说明问题，首先给出涉及到的一些定义。

1.1 线性中心距离

定义1 某一类样本的平均特征称为该类样本的中心 $\boldsymbol{m}$，已知样本向量组 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，那么其中心为

$$\boldsymbol{m}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \tag{1}$$

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 $\boldsymbol{x}_1,\boldsymbol{x}_2$，其样本距离为

$$d(\boldsymbol{x}_1,\boldsymbol{x}_2)=\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2=\sqrt{\sum_{i=1}^{N}(x_1^i-x_2^i)^2} \tag{2}$$

定义3 中心距离指的是各样本到该类中心的距离。假设有一类模式的 N 维训练样本向量为($\boldsymbol{x}_1,\boldsymbol{x}_2$,

收稿日期：2007-09-04
基金项目：国家自然科学基金资助项目(50505051)；陕西省自然科学研究计划项目(2007F19)
作者简介：徐海龙(1981-)，男，陕西韩城人，博士生，主要从事智能信息处理，模式识别，支持向量机研究；
E-mail:xhl-81329@163.com
王晓丹(1966-)，女，陕西汉中人，教授，博士生导师，博士(后)，主要从事智能信息处理，雷达目标识别，支持向量机研究.

$\cdots, x_n$),其中心为 $\boldsymbol{m}$,则中心距离为

$$d(\boldsymbol{x}_i, \boldsymbol{m}) = \| \boldsymbol{x}_1 - \boldsymbol{m} \|_2 = \sqrt{\sum_{i=1}^{N} (x_1^j - m^j)^2} \tag{3}$$

定义 4 最优分类面距离指的是样本与最优分类面的距离。计算式为

$$d_H = \frac{g(\boldsymbol{x})}{\| \omega^* \|} = \frac{(\omega^* \cdot \boldsymbol{x}) + b^*}{\| \omega^* \|} = \frac{\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot \boldsymbol{x}) + b^*}{\sum_{i=1}^{n} \alpha_i^* y_i x_i} \tag{4}$$

定义 5 中心距离最优分类面距离比值(R) 指的是某一样本的中心距离与最优分类面距离的比值:

$$R_{X_i} = \frac{d(\boldsymbol{x}_i, \boldsymbol{m})}{d_H} \tag{5}$$

定义 6 边界向量是某一类模式中,位于其边界上的那些向量。已知模式 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ 及其中心距离与最优分类面距离比值(设定阈值 r_x),则集合 $\{\boldsymbol{x}_i \mid R_{x_i} > r_x, i = 1, 2, \cdots, n\}$ 就是该模式的边界向量。

1.2 非线性中心距离

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。输入空间的两个向量 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[7] 为

$$d^H(z_1, z_2) = \sqrt{K(z_1, z_1) - 2K(z_1, z_2) + K(z_2, z_2)} \tag{6}$$

其中 $K(\cdot, \cdot)$ 是核函数。这里需要指出的是输入空间中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:

$$\boldsymbol{m}_\Phi = \frac{1}{n} \sum_{i=1}^{n} \Phi(\boldsymbol{x}_i) \tag{7}$$

式中 n 是样本的个数。因为不知道映射 $\Phi(x_i)$ 的具体表达形式,所以无法根据此式来求样本中心向量,为此本文给出如下引理[7-9]。

引理 2 已知模式的训练样本为 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$,经非线性映射 Φ 作用后,映射到某一特征空间 H,则在特征空间中的中心距离为

$$d^H(\boldsymbol{x}, \boldsymbol{m}_\Phi) = \sqrt{K(\boldsymbol{x}, \boldsymbol{x}) - \frac{2}{n} \sum_{i=1}^{n} K(\boldsymbol{x}, \boldsymbol{x}_i) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K(\boldsymbol{x}_i, \boldsymbol{x}_i)} \tag{8}$$

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,最优分类面距离,中心距离与最优分类面距离比值。

最优分类面距离为

$$d_H^H = \frac{g(\boldsymbol{x})}{\| \omega^* \|} = \frac{(\omega^* \cdot \boldsymbol{x}) + b^*}{\| \omega^* \|} = \frac{\sum_{i=1}^{n} \alpha_i^* y_i K(\boldsymbol{x}_i \cdot \boldsymbol{x}) + b^*}{\sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i} \tag{9}$$

中心距离与最优分类面距离比值为

$$R_{\Phi(x_i)} = \frac{d^H}{d_H^H} \tag{10}$$

定义 7 已知模式的训练样本为 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$,经非线性映射 Φ 作用后,映射到某一特征空间 H,得到它们在空间 H 中的中心距离与最优分类面距离比值。设定阈值 r_x,则集合:$\{\boldsymbol{x}_i \mid R_{\Phi(x_i)} > r_x, i = 1, 2, \cdots, n\}$ 就是模式的边界向量集合。

1.3 支持向量分布规律

通过对大量的分类模式样本的研究,认为支持向量集是模式的边界向量集合的子集。如图 1 所示,支持向量分布在最优分类面附近,由给定的一个阈值 r_x 确定的两条曲线(面)所夹的一个区域内。阈值 r_x 的取

值,对边界向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,边界向量集合就是包含了支持向量集合的最小集合,甚至边界向量集合就是支持向量集合。

由于满足广义 KKT 条件的样本集合中非边界向量对后继训练影响甚微,可以将其遗忘[10],所以通过只保留边界向量就可以大大提高增量学习速度。本文将定义的中心距离与最优分类面比值引入到增量学习的遗忘规则中,从而得出了一种基于距离比值的增量学习算法。

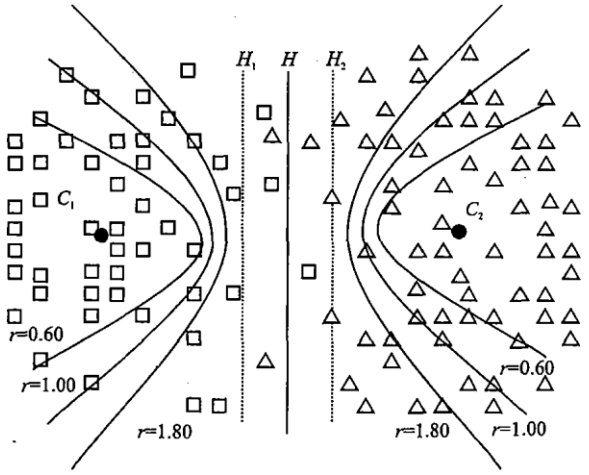

图 1　最优分类面附近向量的分布

Fig. 1　Vectors distribution of optimization hyperplane of Classification

2　算法描述

2.1　符号意义

在给出算法过程之前,首先对算法过程中所用到的符号及其意义进行说明。

$\boldsymbol{X}_o^K$ 表示第 k 次训练用原始样本集;

$\boldsymbol{\Omega}_o^k$ 表示由第 k 次训练用原始样本集得到的 SVM 分类器;

$\boldsymbol{X}_o^{sv}$ 表示 $\boldsymbol{\Omega}_o^k$ 的支持向量集;

$\boldsymbol{X}_o^{s_k}$ 表示第 k 次训练原始样本集中满足 $\boldsymbol{\Omega}_o^k$ 的广义 KKT 条件的样本集;

$\boldsymbol{X}_o^{v_k}$ 表示第 k 次训练原始样本集中违背 $\boldsymbol{\Omega}_o^k$ 的广义 KKT 条件的样本集;

$\boldsymbol{X}_l^k$ 表示第 k 次取出的新增样本($k=0$ 时表示初始样本集);

$\boldsymbol{X}_l^{s_k}$ 表示新增样本中满足 $\boldsymbol{\Omega}_o^k$ 的广义 KKT 条件的样本集;

$\boldsymbol{X}_l^{v_k}$ 表示新增样本中违背 $\boldsymbol{\Omega}_o^k$ 的广义 KKT 条件的样本集;

$\boldsymbol{X}_r^{s_k}$ 表示进行遗忘计算后,剩余的满足 $\boldsymbol{\Omega}_o^k$ 的广义 KKT 条件的样本集;

$\boldsymbol{X}_u^{s_k}$ 表示 $\boldsymbol{X}_o^{s_k}$ 和 $\boldsymbol{X}_l^{s_k}$ 的并集。

2.2　运算过程

步骤 1　判断训练样本集是否是空集,是空集则训练结束,否则转步骤 2;

步骤2　从训练样本集中取出新增样本 $\boldsymbol{X}_l^k$。若 $k=0$,则由 $\boldsymbol{X}_l^0$ 训练得到原始分类器 $\boldsymbol{\Omega}_o^0$,$\boldsymbol{X}_o^1=\boldsymbol{X}_o^1$,$k=k+1$,转步骤 1　若 $k\neq 0$,则步骤 3;

步骤 3　检验 $\boldsymbol{X}_l^k$ 中的样本是否违背 $\boldsymbol{\Omega}_o^k$ 的广义 KKT 条件。根据检验结果,$\boldsymbol{X}_l^k$ 被分为 $\boldsymbol{X}_l^{s_k}$ 和 $\boldsymbol{X}_l^{v_k}$;

步骤 4　将 $\boldsymbol{X}_o^{s_k}$、$\boldsymbol{X}_l^{s_k}$ 合并得 $\boldsymbol{X}_u^{s_k}$,根据标示符,将集合分为正例样本集 $\boldsymbol{A}^+$ 和负例样本集 $\boldsymbol{A}^-$,并分别根据遗忘规则进行处理,舍弃对后继训练影响不大的样本,得到剩余正例样本集 $\boldsymbol{A}_r^+$ 和剩余负例样本集 $\boldsymbol{A}_r^-$,合并二者得 $\boldsymbol{X}_r^{s_k}$;

步骤 5　将 $\boldsymbol{X}_r^{s_k}$、$\boldsymbol{X}_l^{v_k}$、$\boldsymbol{X}_o^{v_k}$ 合并得到 $\boldsymbol{X}_o^{k+1}$,对其训练得到新的分类器 $\boldsymbol{\Omega}_0^{k+1}$,并生成 $\boldsymbol{X}_o^{s_{k+1}}$ 和 $\boldsymbol{X}_o^{v_{k+1}}$,$k=k+1$,转步骤 1。

2.3　遗忘规则

设要处理的样本集为 $\boldsymbol{A}$,设定一合适的阈值 r,根据距离比值法中的定义计算各个样本的中心距离最优分类面距离比值 R_{xi},舍弃 $R_{xi}\leqslant r$ 的样本,保留 $R_{xi}>r$ 的样本(即保留边界向量),得到剩余样本集 $\boldsymbol{A}_r$。

3　实验结果与分析

为验证算法的有效性,使用 UCI 数据库中的 Balance 和 Westontoynonliner 两组数据集进行了实验。Bal-

ance 数据集,共有 627 个样本,样本维数为 4,样本分为 R、B、L 3 类,选取 B 类对其余两类训练;Westontoynonliner 数据集,共有 10 000 个样本,样本维数为 52,样本类别数为 2。SVM 使用的是 Steve Gunn SVM Toolbox,$C=100$,核函数为多项式函数,参数值为 2。

表 1、表 2 分别给出了 Balance 数据集和 Westontoynonliner 数据集的增量学习结果。

表 1 Balance 数据集增量学习结果

Tab. 1 Incremental learning results of Balance data set

训练方法	训练样本数	测试样本数	初始样本数	增量样本/个	阈值 r	支持向量个数	训练时间/s	正确率(%)
			10	10	15	25	7.5	92.04
距离比值法	200	427	20	10	30	27	4.9	92.27
			50	50	30	31	4.1	83.14
标准 SVM	200	427	–	–	–	33	6.0	82.90
			50	50	30	62	8.8	92.05
距离比值法	300	327	100	50	50	47	9.4	92.66
			30	30	60	36	4.7	92.35
标准 SVM	300	327	–	–	–	77	16.9	92.05

表 2 Westontoynonliner 数据集增量学习结果

Tab. 2 Incremental learning results of Westontoynonliner data set

训练方法	训练样本数	测试样本数	初始样本数	增量样本/个	阈值 r	支持向量个数	训练时间/s	正确率(%)
			100	100	500	84	57.7	96.78
距离比值法	1 000	5 000	100	10	500	84	285.4	96.94
			100	10	600	85	190.3	96.68
标准 SVM	1 000	5 000	–	–	–	83	706.8	97.10
			100	50	600	173	470.3	96.68
距离比值法	2 000	5 000	100	100	500	174	387.1	96.60
			500	500	500	174	258.5	96.66
标准 SVM	2 000	5 000	–	–	–	175	4 581.7	96.94

从实验结果可以看出,当训练样本数较大时,距离比值法显著地提高了训练速度,而训练精度却降低甚微;当训练样本数较小时,距离比值法在训练速度上虽无明显优势,但训练精度却无明显降低,甚至其推广能力反而增强。

本算法中阈值 r 的取值对训练精度有重要影响,如 r 过小,因为保留历史数据过多,则达不到提高训练速度的目的;如 r 过大,则丢失大量的重要信息,从而降低训练精度。如何根据训练样本集确定阈值 r 的范围是一个有待研究的问题。

4 结束语

本文研究了支持向量在样本空间的分布特性,在此基础上提出了一种基于距离比值的支持向量机增量训练算法。该算法只需设定一个参数,即可对训练数据进行有效的遗忘淘汰。对标准数据集的实验结果表明,使用该方法进行增量训练在保证训练精度的同时,能有效地提高训练速度。

参考文献:

[1] Vapnik N. The Nature of Statistical Learning Theory[M]. New York: Springer Press,2000.

[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.
WANG Xiaodan,WANG Jiqin. Research and Application of Support Vecor Machine[J]. Journal of Air Force Engineering University:Natural Science Edition, 2004,5(3):49-55. (in Chinese)

[3] Fung G,Mangasarian O L. Incremental Support Vector Machine Classification[R]. Data Mining Institute Technical Report 01

-08. Virginia:2001.

[4] 萧　嵘,王继成,孙正兴,等. 一种SVM增量学习算法α-ISVM[J]. 软件学报,2001 ,12(12):1818-1824.
XIAO Rong ,WANG Jicheng,SUN Zhengxing,et al. An Incremental SVM Learning Algorithm α-ISVM[J]. Journal of Software,2001,12(12):1818-1824. (in Chinese)

[5] 史朝辉,王晓丹, 赵士敏,等. 改进的SVM决策树分类算法[J]. 空军工程大学学报:自然科学版, 2006,7(2):32-35.
SHI Zhaohui, WANG Xiaodan,ZHAO Shimin,et al. An Improved Algorithm for SVM Decision Tree[J]. Journal of Air Force Engineering University:Natural Science Edition, 2006,7(2):32-35. (in Chinese)

[6] 史朝辉,王晓丹,杨建勋. 一种SVM增量训练淘汰算法[J]. 计算机工程与应用,2005,41(23):187-189.
SHI Zhaohui,WANG Xiaodan,YANG Jianxun. A Removing Algorithm for Incremental SVM Training[J]. Computer Engineering and Applications,2005,41(23):187-189. (in Chinese)

[7] 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.
JIAO Licheng,ZHANG Li,ZHOU Weida. Pre-extracting Support Vectors for Support Vector Machine[J]. Acta Electronic Sinic, 2001, 29(3):383-386. (in Chinese)

[8] Hyunsoo Kim,Haesun Park. Incremental and Decremental Least Squares Support Vector Machine and Its Application to Drug Design[C]//Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference[S. L]:2004: 2194-2195.

[9] Stefan R ping. Incremental Learning with Support Vector Machines[C]//Proceedings of the 2001 IEEE International Conference on Data Mining[S. L]:2001:1119-1120.

[10] 王晓丹,郑春颖,吴崇明,等. 一种新的SVM对等增量学习算法[J]. 计算机应用,2006,26(10):2440-2443.
WANG Xiaodan,ZHENG Chunying,WU Chongming,et al. New Algorithm for SVM-Based Incremental Learning [J]. Computer Applications,2006,26(10):2440-2443. (in Chinese)

(编辑:田新华)

An Incremental Training Algorithm of SVM Based on the Distance Ratio

XU Hai-long, WANG Xiao-dan,SHI Zhao-hui, HUA JI-xue,Quan Wen
(Missile Institute, Air Force Engineering University, Sanyuan 713800, Shaanxi, China)

Abstract:Due to the good learning and generalization performance, the SVM (support vector machine) has been widely used in practice. But, how to make the SVM more effectively perform incremental learning is a problem that needs to be solved in the present application of the SVM. The distribution characteristics of Support vectors are studied and a novel improved incremental SVM learning algorithm - distance ratio algorithm is proposed. According to the removing rules of the proposed method, an appropriate parameter is set and samples that have less effect on later training are abandoned. According to the definition in distance ratio algorithm, the ratio between the center distance of each sample and the distance of each to the optimum classification surface is calculated. In this way, the training data sets can be effectively reduced. Experiment on standard data sets shows that by using this method the classification accuracy can be guaranteed and the training speed can be effectively improved.

Key words:Support Vector Machine; Incremental Training; Removing Method; Margin vector; Distance Ratio Algorithm