基于向量投影的支持向量机增量算法*)

张健沛　赵　莹　杨　静
(哈尔滨工程大学　哈尔滨 150001)

摘　要　提出一种新的基于向量投影的支持向量机增量式学习算法。该算法根据支持向量的几何分布特点，采用向量投影的方法对初始样本及增量样本在有效地避免预选取失效情况下进行预选取。选取最有可能成为支持向量的样本形成边界向量集，并在其上进行支持向量机训练。通过对初始样本是否满足新增样本集 KKT 条件的判断，解决非支持向量向支持向量转化的问题，有效地处理历史数据。实验表明，基于向量投影的支持向量机增量算法可以有效地减少训练样本数，积累历史信息，提高训练的速度，从而具有更好的推广能力。

关键词　增量算法，支持向量机，向量投影，预选取失效

Incremental Learning Algorithm of Support Vector Machine Based on Vector Projection

ZHANG Jian-Pei　ZHAO Ying　YANG Jing
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001)

Abstract　A new incremental learning algorithm of support vector machine based on vector projection is proposed. The geometric character of the support vectors is used to extract the simples which can be the support vectors most possibly from the original simples and the incremental simples, avoiding the invalid per-extracting. The simples pre-extracted are used to generate the bound support vector set, on which the support vector machine training. It is used to solve the transferring between the support vectors and no-support vectors whether the original simples satisfy the KKT condition of the incremental simples, which can deal with the history data efficiently. The experiment shows that the algorithm can greatly reduce the number of training data in incremental SVM training, speeding up the training process, which has the better generalization performance.

Keywords　Incremental algorithm, Support vector machine, Vector projection, Pre-extracting

支持向量机(Support Vector Machines, SVM)是 Vapnik[1]等在统计学习理论基础上发展起来的针对小样本的机器学习方法。传统的支持向量机训练算法中，当有新增样本加入时，需要对所有训练样本重新进行训练，这无疑需要消耗大量的运算资源。因此，对支持向量机分类算法中的增量算法的研究具有重要的理论意义和实用价值。本文的研究基于这样一个出发点：如果能将原有的分类信息保存起来与新加入的样本一起进行训练，则既能继承之前所学习的知识，又能减少由于新样本的加入而重新学习的时间。文[2]提出一种使用多支持向量机进行增量学习的算法，该方法解决了在大样本情况下的 SVM 学习问题，但由于初始训练是在全部的初始训练集上进行，训练速度的提高并不明显。文[3]提出了一种 α-SVM 支持向量机增量学习算法，训练集主要从支持向量、误分数据中有选择地淘汰一些样本来获得，该算法需要选择多个参数，遗憾的是目前还没有确定这些参数的一个有效方法。本文结合文[4]提出的向量投影的方法，深入分析预选取的有效性，并将该算法与增量学习思想相结合，提出一种基于向量投影的支持向量机增量学习算法，该算法可以在避免预选取失效模式发生的同时，实现支持向量机的增量学习。

1　支持向量机

在两类模式识别问题中，给定训练样本 $G=\{(x_i,y_i)\}_{i=1}^n$，其中 $x_i\in R^d$，$y_i\in\{+1,-1\}$，通过训练样本构造分类决策函数 $f(x,\alpha)$，使训练样本以最大间隔分开，即使分类器具有较好的泛化能力[5]。

在求解决策函数的过程中，需要构造如下的优化问题[1]：

$$\min \frac{1}{2}\|\omega\|^2+C\sum_{i=1}^n\xi_i$$
$$\text{s.t. } y_i[(\omega\cdot x_i)+b]\geqslant 1-\xi_i \quad (1)$$
$$\xi_i\geqslant 0, i=1,2,\cdots,n$$

式(1)的 Wolf 对偶问题为：

$$\max \sum_{i=1}^n\alpha_i-\frac{1}{2}\sum_{i,j=1}^n\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)$$
$$\text{s.t. } \sum_{i=1}^n y_i\alpha_i=0 \quad (2)$$
$$0\leqslant\alpha_i\leqslant C$$

最后求出的决策函数为：

$$f(x)=\text{sgn}(\sum_{i=1}^n y_i\alpha_i K(x_i\cdot x)+b) \quad (3)$$

式(3)中的 $K(\cdot)$是满足 Mercer 条件的核函数，它可将原输入空间的线性不可分问题转化为高维甚至无穷维 Hibert 空间的线性可分问题，然后在高维或无穷维空间中求解最优化问题(2)。其中，称 Lagrange 乘子 α_i 不为零所对应的样本为支持向量。在式(1)中，对偶问题的最优解为 $\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_n]$，该最优解使得每一个样本 X 满足优化问题的 KKT 条件[6]：

$$\alpha_i=0\Rightarrow|f(x_i)|\geqslant 1 \quad (4)$$

*)本课题得到国家自然科学基金(60673131)和黑龙江省自然科学基金(F2005-02)的资助。张健沛　博士，教授，主要研究方向：数据挖掘。

$$0<\alpha_i<C\Rightarrow|f(x_i)|=1 \tag{5}$$

$$\alpha_i=C\Rightarrow|f(x_i)|\leqslant 1 \tag{6}$$

其中，$\alpha_i=0$ 的样本为非支持向量，对于训练没有贡献。在支持向量机的训练中，当有新样本加入时，设其 Lagrange 乘子 $\alpha_i=0$。根据式(4)，如果 $y_i f(x_i)\geqslant 1$，则说明新增样本满足 KKT 条件，否则称该样本为违背 KKT 条件的样本。

由于新增样本中有违反 KKT 条件的样本，它们的加入将会改变支持向量集。这种改变要有两种可能：那些违反 KKT 条件的样本可能由非支持向量转变成为支持向量；原来的支持向量转化为非支持向量。这种支持向量与非支持向量之间相互转化的情况如图 1 所示，其中实心和空心代表初始样本和新增样本，三角和圆圈代表正类和负类。$f(x)=0$ 为初始样本的训练结果，$g(x)=0$ 为加入新增训练样本后，新增样本与初始样本共同训练的结果。不难看出，由于新增样本的加入，使得初始样本集中的非支持向量 $N1,N2,N3,N4$ 转化为支持向量，形成了新的支持向量集。

2 支持向量预选取

由式(3)可知，支持向量机的分类决策函数仅由那些拉格朗日乘子 α_i 不为零的样本决定，即由支持向量决定。从图 1 可知，支持向量具有显著的几何特征，它们为那些位于本类边缘、最靠近分类超平面的样本。利用支持向量的几何特性，在进行支持向量机训练前，选取最有可能成为支持向量的样本形成边界向量集，可以大大缩小进行支持向量训练的样本数。通过这种方法可以有效地舍弃一些无用的历史数据，提高支持向量机的训练速度。

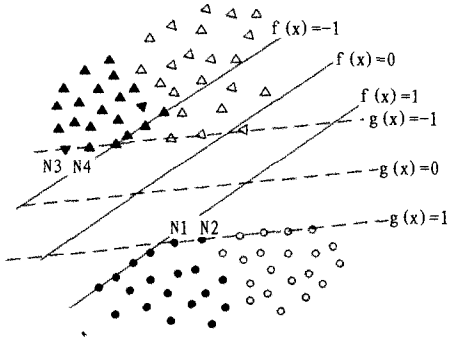

图 1 新增样本对支持向量集的影响

2.1 向量投影

首先，给出本算法中用到的一些定义。

定义 1(样本中心 m_i) 定义第 i 类样本的平均特征为该类样本的中心 m_i，即：

$$m_i=\frac{1}{n}\sum_{i=1}^{n}X_i$$

其中，X_i 为 d 维向量。

定义 2(特征方向) 定义样本中心到 m_i 的 m_j 方向 $\overrightarrow{m_i m_j}$ 为样本集第 i 类到样本集第 j 类的特征方向。

定义 3(特征距离 S) 样本点 X_i 在基于特征方向上的投影，称为 X_i 的特征距离。

$$S(X_i)=\frac{(\overrightarrow{X_i m_i}\cdot\overrightarrow{m_i m_j})}{\|m_i-m_j\|_2}$$

2.2 边界向量模式

定义 4(边界向量) 已知样本中心距离 $d(m_i,m_j)=\|m_i-m_j\|_2$，分别计算特征距离 $S(X_i)$ 和 $S(Y_i)$，以两分类问题为例，令

$$r_1=\max_{X_i\in\{1类\}}(S(X_i))$$

$$r_2=\max_{Y_i\in\{2类\}}(S(X_i))$$

引入非负修正因子 $\delta_i\geqslant 0$

(1)当 $r_1+r_2<d$ 时，若样本的特征距离 $S(X_i)$ 和 $S(Y_i)$ 满足

$$r_1-\delta_1\leqslant S(X_i)\leqslant r_1$$

$$r_2-\delta_2\leqslant S(Y_i)\leqslant r_2$$

则定义该模式为边界向量。

(2)当 $r_1+r_2>d$ 时，若样本的特征距离 $S(X_i)$ 和 $S(Y_i)$ 满足

$$d-r_2-\delta_1\leqslant S(X_i)\leqslant r_1+\delta_1$$

$$d-r_1-\delta_2\leqslant S(X_i)\leqslant r_2+\delta_2$$

则定义该模式为边界向量。

2.3 预选取失效模式

预选取的目的就是在进行支持向量训练前有效地压缩训练样本集。如果预选取中相关参数选择不当，就会使得预选取的样本集过小或过大。对于前一种情况，由于预选取的样本集不能包含样本集全部的分类信息，将会严重影响训练的精度；而后者，在预选取时加入了过多冗余的样本，这些样本的加入对于分类决策函数不但没有贡献，反而增加了训练复杂度。因此，本文引入一个预选取失效模式的定义，以此来判断预选取的效果。

定义 5(预选取失效模式)

(1)当 $r_1+r_2<d$ 时，若样本的特征距离的边界值满足

$$r_i-\delta_i\leqslant 0$$

则定义该模式为预选取失效模式。

(2)当 $r_1+r_2>d$ 时，若样本的特征距离的边界值满足

$$d-r_j-\delta_i\leqslant 0 \tag{7}$$

则定义该模式为预选取失效模式。

其中，修正因子 $\delta_i\geqslant 0$ 的确定方法为

$$\delta_i=\mu\cdot r_i+\frac{1}{(N_i+N_j)/D}$$

其中参数 μ 为覆盖度因子，D 为噪声平衡因子。

文[4]提出的修正因子 $\delta\geqslant 0$ 的确定方法为

$$\delta=\mu\cdot center\_dis+\frac{1}{(N_1+N_2)/D} \tag{8}$$

$\mu\in[0,0.2]$，$D\in[0,10]$

这里 *center-dis* 为两类样本中心距离 d。其它参数的定义与本文相同。本文之所以没有选择这个修正因子的定义方式，是由于当两类样本数量不均衡时，这个修正因子将有可能导致预选取失效的情况。这里以第一类样本为例进行分析。

当 $r_1+r_2<d$ 时，$S(X_i)$ 的下界为 $r_1-\delta$。此时，若

$$r_1<\frac{\mu}{1-\mu}r_2$$

由于 $r_1+r_2<d$，于是有

$$r_1+\mu d<0$$

即得，下界 $r_1-\delta<0$：

由式(7)可知，此模式为预选取失效模式。

同理，当 $r_1+r_2>d$ 时，$S(X_i)$ 的下界为 $d-r_2-\delta$。此时，若

$$r_1<\frac{\mu}{1-\mu}r_2$$

由于 $r_1+r_2>d$，于是有

$$(1-\mu)\cdot d-r_2<0$$

即得，下界：$d-\delta-r_2<0$

由式(8)可知此模式为预选取失效模式。

当出现预选取失效模式时，预选取的样本数接近于训练样本数，预选取没有起到缩小训练样本集，减少训练时间及空间的作用，反而增加了整体的训练时间。特别是，当样本数较大时，预选取所消耗的时间将会严重地制约训练的速度，所以应该采用避免这种情况发生的预选取策略。

3 基于向量投影的SVM增量学习算法

在进行增量学习的时候，既要考虑如何继承已学习的知识，将初始样本集的分类信息与增量样本的分类信息相融合，同时也要考虑如何完善初始样本已学到的知识。由于初始样本并不能包含新增样本的所有分类信息，当进行增量学习的时候，初始样本中的一些非支持向量会转化为支持向量。若没有充分考虑这些样本的重要性，势必会影响分类的精度。基于以上考虑，本文采取在初始样本的边界向量集中筛选违背新增样本KKT条件的训练样本，将这些样本与增量样本中的边界向量一起进行支持向量机的训练。

3.1 增量学习算法的主要步骤

算法的问题描述如下：

前提：初始样本集 X_0，增量训练数据集 X_1，并且设 $X_1 \cap X_0=\varnothing$。算法的目标是寻找基于 $X_1 \cup X_0$ 的SVM分类器 φ 和对应的支持向量集 SV_1。算法具体步骤如下：

(1)分别针对数据集 X_0^+ 和 X_0^- 求出其边界向量集 N_0，在 N_0 上得到 SVM_0 和支持向量集 SV_0；

(2)加入新增训练样本 X_1，求出其上的边界向量集 N_1，在 N_1 上进行训练得 SVM_1，支持向量集 SV_1；

(3)检验 N_0 中的样本是否有违背 SVM_1 的KKT条件的样本，如果没有样本违背KKT条件的样本，则算法停止，SV_1 为增量学习的结果；否则，将 N_0 分为 N_0^d 和 N_0^s 两部分，N_0^d 为满足KKT条件的样本集，N_0^s 为违背KKT条件的样本集；

(4)更新 SV_1，$SV'_1=SV_1+N_0^s+SV_0$，$SV_1=SV'_1$；

(5)重复步骤(2)、(3)、(4)，即可实现连续对多批新增样本进行增量学习。

4 实验与分析

基于以上的研究，本文使用基于向量投影的支持向量机增量学习算法在两组数据上进行实验：一组为UCI数据库上的tic-tac-toe数据集，另一组为人造数据。所有算法均在Matlab7.0和LibSvm-mat-2.83工具包[7]的基础上实现。

实验1

采用tic-tac-toe数据集，数据集共958个样本，样本维数样本数为280，新增量样本数为：170、120、100，测试样本数为288。选择RBF核函数 $K(x_i,x_j)=\exp(-\|x_i-x_j\|^2/\gamma)$ 选取的参数如下：$D=10$，$\mu=0.2$，$\gamma=128$，增量训练的对比结果如表1所示。

实验1的结果显示，通过基于向量投影的方法提取边界向量，可以有效地减少训练样本数，尤其当支持向量数占训练样本数的比例较小时。从训练时间上看，基于向量投影的支持向量机增量算法所用的时间要少于标准支持向量机训练算法所用的时间，而且，随着样本数的增加，这种优势更加明显。

表1 标准SVM学习结果与增量学习结果比较

训练集	标准SVM			基于向量投影SVM增量算法			
	训练时间(s)	训练精度(%)	支持向量数	训练时间(s)	边界向量数	支持向量数	训练精度(%)
初始样本	73	88.6	48	73	91	53	88.64
增量1	118	89.7	31	44	63	37	89.28
增量2	163	90.8	22	70	48	24	90.67
增量3	91	91.8	17	120	40	19	91.55

实验2

取两类正态随机分布的样本，类概率分别为 $P(\omega_1)=0.6$，$P(\omega_2)=0.4$；类1和类2的样本均值和协方差矩阵分别为 $\mu_1=(0,0)^T$，$\mu_2=(12,12)^T$，$\Sigma_1=\begin{bmatrix}2&0\\0&2\end{bmatrix}$ $\Sigma_2=\begin{bmatrix}8&0\\0&8\end{bmatrix}$。训练集包括两类共600个样本。其中选取400个作为训练样本，200作为测试样本。选取的参数如下：$D=10$，$\mu=0.2$，$\gamma=128$。表2给出了实验的对比结果。

表2 改进的基于向量投影的SVM性能比较

训练算法	训练时间(s)	边界向量数	分类精度(%)
标准SVM	124	无	98.32
改进的基于向量投影的SVM	89	43	98.21

实验2所选用的数据为样本分布不均衡的情况，即有可能出现预选取失效的情况。本文提出的基于向量投影的SVM增量训练算法由于有效地避免了预选取失效，从而提高了训练速度。

结论 本文通过对支持向量几何特征的分析，对新增样本进行了基于向量投影的预选取。在预取的过程中定义边界向量集，有效地避免了预选取失效模式的发生。选择那些最有可能成为支持向量的样本形成边界向量集，在其上进行支持向量机的训练。预选取策略与增量学习思想相结合，在增量学习的基础上合理有效地压缩了训练样本集的规模。实验证明，这种方法可以在增量学习提高分类速度的基础上进一步提高分类的速度，适合大规模样本集的操作。进一步的研究工作包括：如何将增量学习的方法运用到多分类问题中。

参考文献

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

2 萧嵘，王继成，孙正兴，张福炎. 一种SVM增量学习算法α-ISVM. 软件学报，2001，12(12)：1818～1823

3 李凯，黄厚宽. 支持向量机增量学习算法研究. 北方交通大学学报，2003，27(5)：34～37

4 李青，焦李成，周伟达. 基于向量投影的支持向量预选取. 计算机学报，2005，28(2)：145～152

5 Christopher J, Burges C. A Tutorial on Support Vector Machines for Pattern Recognition. Boston: Kluwer Academic Publishers, 1998

6 周伟达，张莉，焦李成. 支撑矢量机推广能力分析. 电子学报，2001，29(5)：590～594

7 Chang Chih-Chung, Lin Chih-Jen. LIBSVM: a library for support vector machines. 2001. http://www.csie.ntu.edu.tw/～cjlin/libsvm