基于几何分析的支持向量机快速训练与分类算法

胡正平[1] 吴 燕[2] 张 晔[1]

[1](哈尔滨工业大学 图象信息处理研究所, 哈尔滨 150001) [2](燕山大学通信电子工程系, 秦皇岛 066004)

摘 要 当支持向量机中存在相互混叠的海量训练样本时,不但支持向量求取困难,且支持向量数目巨大,这两个问题已成为限制其应用的瓶颈问题。该文通过对支持向量几何意义的分析,首先研究了支持向量的分布特性,并提出了基于几何分析的支持向量机快速算法,该算法首先从训练样本中选择出部分近邻向量,然后在进行混叠度分析的基础上,选择真实的边界向量样本子空间用来代替全部训练集,这样既大大减少了训练样本数目,同时去除了混叠严重的奇异样本的影响,并大大减少了支持向量的数目。实验结果表明:该算法在不影响分类性能的条件下,可以加快支持向量机的训练速度和分类速度。

关键词 支持向量机 近邻向量 边界向量

中图法分类号: TP301 **文献标识码**: A **文章编号**: 1006-8961(2007)01-0082-05

A Novel Fast Support Vector Machine Based on Support Vector Geometry Analysis

HU Zheng-ping[1], WU Yan[2], ZHANG Ye[1]

[1](*Institute of Image Information Processing Harbin Institute of Technology, Harbin* 150001)

[2](*Department of Communication and Electronic Engineering& Yanshan University, Qinhuangdao* 066004)

Abstract Support vector machine, a research hotspot of the pattern recognition in recent years, performs successfully in solving the nonlinear and high dimensional problems. However, training a support vector machine is equivalent to solving a linearly constrained quadratic programming problem in a number of variables equal to the number of data points. This optimization problem is known to be challenging when existing large number of training data points. Also, it is well known that the number of support vector plays an important role in the classification speed of SVM. So the method of pre-analysis efficient support vectors are used to train classifier becomes a novel task in SVM fields. In this paper, on the basis of a deep investigation into the geometry principle of support vectors and its distribution, we firstly pick out some neighbor vectors by nearest interclass distance analysis, and then select the margin vector by computing its intermixed factor of the neighbor vectors. So this method speeds up the SVM training and classifying synchronously by reducing the number of training samples and trimming the intermixed samples, while the ability of SVM remains unchanged.

Keywords support vector machine, neighbor vector, margin vector

1 引 言

支持向量机(support vector machines, SVM)因为其优越的性能而成为近年研究的热点。支持向量机是建立在统计学习理论的 VC 维概念以及结构风险最小原理的基础上,并根据有限样本信息在模型复杂度与经验风险之间进行折衷,以获得最好的推广能力[1],而支持向量机的训练就是求解线性凸约束的二次规划问题。当训练样本数目巨大,且相互混叠严重时,一方面由于训练时间花费巨大(甚至不可求解);另一方面由于所求支持向量数目巨大,

基金项目:国家自然科学基金项目(60272073);河北省科学技术研究与发展项目(2005315)

收稿日期:2005-05-12;**改回日期**:2005-11-13

第一作者简介:胡正平(1970~),男。哈尔滨工业大学信号信息处理专业博士研究生,燕山大学通信电子工程系教师。目前的研究方向为统计学习理论、图像分析与处理。E-mail: tnpochw@263.net

从而导致模式分类需要花费大量的时间,这已成为制约支持向量机实际应用的主要问题。针对此问题,许多学者提出了不少解决该问题的思路。Hu 等提出了加速分解的鲁棒支持向量机算法[2],即通过引入归一化中心距离这一新的松弛变量来建立一种新的支持向量目标函数,并以训练数据点到该类中心的距离来计算自适应类间间隔,再通过对松弛变量参数进行的控制,就可以达到从不同范围选取训练样本的目的,进而减少训练样本数量,以加快训练速度,但其缺点是最佳控制参数选择困难。文献[3,4]提出了迭代修剪的方法用于解决大规模支持向量机的学习问题,该方法通过修剪训练样本的混叠样本来达到提高 SVM 分类器的训练速度和分类精度的目的。文献[5]提出将选择到的训练样本点到另一类中心距离作为支持向量的选取准则,这种方法适合于处理类内数据聚合度大的球形分布训练样本,而对于复杂分布的情况却缺乏适应性。文献[6]提出基于向量投影的支持向量预选取思路,即通过用检测边界向量代替训练样本来大大加快 SVM 的训练速度。如何根据训练样本的分布特点来建立支持向量合理有效的选取标准,同时克服混叠和大容量样本训练复杂度的困难,以及建立高效的支持向量机分类器是本文研究工作的出发点。

通过分析支持向量机的几何意义及其分类原理可以发现,支持向量机的训练准则就是在追求最大类间间隔的同时,不能付出太多的错分代价。在支持向量机分类器中,支持向量主要位于两类样本相互接近的边界部分,同时那些混杂在另一类中的样本点无助于提高分类器的性能,反而会增加分类器训练与分类的复杂度。基于上述发现,本文通过从训练样本中选择出部分近邻向量,在进行混叠度分析的基础上,选择合理的边界向量作为训练样本子空间,用于代替原来的训练样本,再结合已有的快速算法进行训练。这样就大大减少了训练样本,同时去除了影响分类速度的混叠严重的支持向量,从而加快了支持向量机的训练速度和分类速度。对比实验结果表明,本文提出的方法不仅在训练速度、分类正确率、分类速度等方面的表现优异,而且适用于大多训练样本分布情况。

2　支持向量机

支持向量机用于分类的实质就是寻找一个最优最大间隔超平面,并通过引入核函数来巧妙解决将低维向量空间映射到高维空间的维数升高问题。

给定一组训练样本数据 $\boldsymbol{T}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{l}$,其中 $\boldsymbol{x}_i\in\mathbf{R}^n$,$y_i\in\{-1,+1\}$,$\boldsymbol{x}_i$ 表示输入模式的特征向量,训练的目的就是为了寻找结构风险最小的判决函数 $f(\boldsymbol{x},\boldsymbol{a})$。具有最大间隔的分类超平面就可以满足这个要求,假定超平面为 $\boldsymbol{w}\cdot\boldsymbol{x}+b=0$,若要得到此超平面,则需要求解下面的凸二次规划问题

$$\begin{cases}\min\Phi(\boldsymbol{w})=\dfrac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w})\\ \text{s.t. } y_i[(\boldsymbol{x}_i\cdot\boldsymbol{w})-b]\geqslant 1\quad i=1,\cdots,l\end{cases}\tag{1}$$

如果两类样本不可分,则引入松弛变量 $\xi_i\geqslant 0,i=1,\cdots,l$,上面的优化问题就转化为求解下面的凸二次规划问题

$$\begin{cases}\min\Phi(\boldsymbol{w},\boldsymbol{\xi})=\dfrac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w})+C\cdot\left(\sum\limits_{i=1}^{l}\xi_i\right)\\ \text{s.t. } y_i[(\boldsymbol{x}_i\cdot\boldsymbol{w})-b]\geqslant 1-\xi_i,i=1,\cdots,l\\ \xi_i\geqslant 0,i=1,2,\cdots,l\end{cases}\tag{2}$$

通过拉格朗日因子法,就可以将上式转化为其对偶形式

$$\begin{cases}\max L(\boldsymbol{a})=\sum\limits_{i=1}^{l}a_i-\dfrac{1}{2}\sum\limits_{i,j=1}^{l}a_ia_jy_iy_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)\\ \text{s.t. } 0\leqslant a_i\leqslant C,\ i=1,2,\cdots,l\\ \sum\limits_{i=1}^{l}a_iy_i=0\end{cases}\tag{3}$$

求解上式,即得到以下的线性分类器

$$f(\boldsymbol{x})=\operatorname{sgn}\left(\sum_{i=1}^{l}a_iy_i(\boldsymbol{x}\cdot\boldsymbol{x}_i)+b\right)\tag{4}$$

这里,因子 a_i 是二次规划(quadratic programming, QP)问题的解,它的范围约束在 $[0,C]$ 内,且每一个训练样本对应一个 a_i,同时许多系数 a_i 严格等于 0,只有那些非 0 系数的样本才影响分类结果,由于分类超平面只与这些样本有关,因此将对应系数 a_i 不等于 0 的样本称为支持向量。

如果输入模式在原始空间线性不可分,则通过引入非线性映射函数 $\varphi(\boldsymbol{x})$,即可以将低维空间线性不可分问题映射到高维特征空间而变成线性可分的问题,这样式(2)的二次规划问题就转化为

$$\begin{cases}\min\Phi(\boldsymbol{w},\boldsymbol{\xi})=\dfrac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w})+C\cdot\left(\sum\limits_{i=1}^{l}\xi_i\right)\\ \text{s.t. }\ y_i[(\varphi(\boldsymbol{x}_i)\cdot\boldsymbol{w})-b]\geqslant 1-\xi_i,i=1,\cdots,l\\ \xi_i\geqslant 0,i=1,2,\cdots,l\end{cases}\tag{5}$$

通过拉格朗日因子法,则可以将上式转化为其对偶形式

$$\begin{cases} \max L(\boldsymbol{a}) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j (\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j)) \\ \text{s.t.} \quad 0 \leqslant a_i \leqslant C,\ i = 1,2,\cdots,l \\ \sum_{i=1}^{l} a_i y_i = 0 \end{cases} \tag{6}$$

令核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j)$，则上面的优化问题就简化为

$$\begin{cases} \max L(\boldsymbol{a}) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ \text{s.t.} \quad 0 \leqslant a_i \leqslant C,\ i = 1,2,\cdots,l \\ \sum_{i=1}^{l} a_i y_i = 0 \end{cases} \tag{7}$$

进而得到的非线性超平面为

$$\sum_i a_i y_i K(\boldsymbol{x},\boldsymbol{x}_i) + b = 0 \tag{8}$$

相应的非线性分类器为

$$f(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_i a_i y_i K(\boldsymbol{x},\boldsymbol{x}_i) + b\right) \tag{9}$$

3 支持向量影响因子分析

3.1 支持向量几何意义

对于 SVM 分类器而言，训练的过程就是找到处于类边界上的样本，因为它们最可能为支持向量。由于支持向量决定了最优超平面的形式，即决定了分类函数的形式，因此传统的 SVM 方法总是包含了全部支持向量，特别是当两类样本混叠严重时，支持向量的数目非常巨大，这一方面使分类器的性能没有提高，另一方面使模式分类时严重影响分类速度。从几何上直观地看，支持向量虽最靠近两类类边界，但是应该去掉那些交互区域内混杂在另一类中的样本点，因为它们不但无助于提高分类性能，反而导致分类器形式更加复杂。由此可见，如何利用支持向量的几何分布来建立支持向量的预选取标准是解决问题的关键。

3.2 支持向量选取准则

定义 1 训练样本原始空间的类间最近邻距离定义为

$$d^{(1)}(\boldsymbol{x}_i) = d(\boldsymbol{x}_i,\boldsymbol{z}_j) = \min_{\boldsymbol{z}_j\in\{-1\text{类}\}} \|\boldsymbol{x}_i - \boldsymbol{z}_j\|_2$$

$$= \min_{\boldsymbol{z}_j\in\{-1\text{类}\}} \sqrt{\sum_{k=1}^{n} (\xi_k^{(i)} - \varepsilon_k^{(j)})^2} \tag{10}$$

$$\boldsymbol{x}_i = (\xi_1^{(i)},\cdots,\xi_n^{(i)}),\boldsymbol{z}_i = (\varepsilon_1^{(j)},\cdots,\varepsilon_n^{(j)})$$

同样地，也可以得到非线性映射后的特征空间的最近邻距离为

$$D^{(1)}(\boldsymbol{x}_i) = D(\boldsymbol{x}_i,\boldsymbol{z}_j) = \min_{\boldsymbol{z}_j\in\{-1\text{类}\}} \|\varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{z}_j)\|_2 \tag{11}$$

引入核函数后，则最近邻距离表示为

$$D^{(1)}(\boldsymbol{x}_i) = \min_{\boldsymbol{z}_j\in\{-1\text{类}\}} (K(\boldsymbol{x}_i,\boldsymbol{x}_i) + K(\boldsymbol{z}_j,\boldsymbol{z}_j) - 2K(\boldsymbol{x}_i,\boldsymbol{z}_j))^{1/2} \tag{12}$$

这里 $\boldsymbol{x}_i$ 属于正样本集合（+1 类），$\boldsymbol{z}_j$ 属于负样本集合（-1 类）。同样的道理也可以得到任意负样本的最近邻类间距离。最近邻类间距离示意图如图 1 所示。

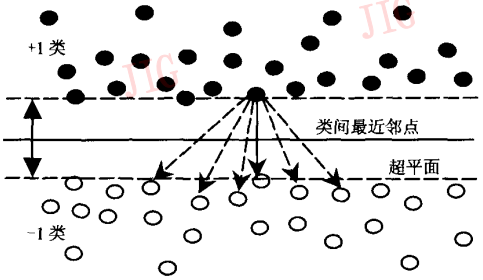

图 1 最近邻类间距离示意图

Fig. 1 Schematic illustration of interclass nearest distance

定义 2（近邻向量） 已知某一样本类间最近邻距离 d_1，如果 $d_1 \leqslant T_{\text{dis}}$，则该样本为近邻向量。这里 T_{dis} 为一距离门限值（可通过分析类间距离分布来确定，也可以采用排序操作来确定，这里相对距离最重要）。支持向量分布区域示意图如图 2 所示。

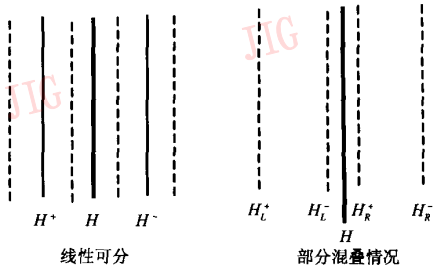

图 2 支持向量分布范围示意图

Fig. 2 Schematic illustration of support vectors distribution

这里的目的是为了利用预选取的部分边界样本代替海量原始训练样本来进行训练，并通过降低训练样本数目和消除奇异点，以便在加快训练速度的同时，提高所得分类器的性能。下面本文将建立向量的影响因子度量准则作为支持向量的预选取标准。从几何上直观地看，支持向量最靠近两类边界，

但是不包括那些混叠严重的样本点。

定义 3　训练样本原始空间的类内最近邻距离定义为

$$d^{(2)}(\boldsymbol{x}_i)=d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\min_{\boldsymbol{x}_j\in\{+1\text{类}\}}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2$$

$$=\min_{\boldsymbol{x}_j\in\{+1\text{类}\}}\sqrt{\sum_{k=1}^{n}(\xi_k^{(i)}-\varepsilon_k^{(j)})^2}\qquad(13)$$

$$\boldsymbol{x}_i=(\xi_1^{(i)},\cdots,\xi_n^{(i)}),\ \boldsymbol{x}_j=(\varepsilon_1^{(j)},\cdots,\varepsilon_n^{(j)})$$

同样地,也可以得到非线性映射后的特征空间的最近邻距离,即

$$D^{(2)}(\boldsymbol{x}_i)=D(\boldsymbol{x}_i,\boldsymbol{x}_j)=\min_{\boldsymbol{x}_j\in\{-1\text{类}\}}\|\varphi(\boldsymbol{x}_i)-\varphi(\boldsymbol{x}_j)\|_2\qquad(14)$$

当引入核函数后,则最近邻距离可表示为

$$D^{(2)}(\boldsymbol{x}_i)=\min_{\boldsymbol{x}_j\in\{-1\text{类}\}}(K(\boldsymbol{x}_i,\boldsymbol{x}_i)+K(\boldsymbol{x}_j,\boldsymbol{x}_j)-2K(\boldsymbol{x}_i,\boldsymbol{x}_j))^{1/2}\qquad(15)$$

这里 $\boldsymbol{x}_i$ 属于正样本集合(+1 类),$\boldsymbol{x}_j$ 也属于正样本集合(+1 类)。同样的道理也可以得到任意负样本的最近邻类内距离。

考虑到简单性,近邻 $\boldsymbol{x}_i$ 的混叠度 M 定义为

$$M(\boldsymbol{x}_i)=\frac{D^{(2)}(\boldsymbol{x}_i)}{D^{(1)}(\boldsymbol{x}_i)}\qquad(16)$$

如果考虑到多近邻(k)的情况,这里假设 k 近邻类间距离分别为 $d_1^{(1)},d_2^{(1)},\cdots,d_k^{(1)}$,而 k 近邻类内距离分别为 $d_1^{(2)},d_2^{(2)},\cdots,d_k^{(2)}$,则近邻 $\boldsymbol{x}_i$ 的混叠度 M 定义为

$$M(\boldsymbol{x}_i)=\frac{\mu_2/\delta_2}{\mu_1/\delta_1}\qquad(17)$$

其中,$\mu_1,\delta_1,\mu_2,\delta_2$ 分别为 k 近邻类间距离和类内距离的均值与方差。当 $M(\boldsymbol{x}_i)$ 越大时,则说明样本 $\boldsymbol{x}_i$ 混叠严重。

定义 4(边界向量)　已知某一样本 $\boldsymbol{x}_i$ 为近邻向量,如果该近邻向量的混叠度 $M(\boldsymbol{x}_i)<T_{\mathrm{M}}$,则该样本为边界向量。这里 T_{M}(下角 M 代表 mixture)为一混叠度门限值(可通过分析混叠度的分布来确定)。

根据上面的定义,本文构建的训练算法主要包括下列步骤:

(1) 选择合适的核函数以及核参数;

(2) 根据类间近邻距离进行排序处理;

(3) 选择合适的样本子集作为近邻向量;

(4) 根据步骤(3)的结果,分析各个近邻向量的混叠度,进而选择出部分边界向量用来代替全部训练样本进行 SVM 训练。

(5)返回步骤(1),通过多次试验即可找到最佳核参数。

通过上面的步骤,就可得到边界向量的预选取方法,训练时,可利用较少的边界向量代替整个训练样本集,这样就可以大大加快支持向量机的训练速度。同时因为消除了部分混叠严重的奇异样本的影响,所以使得支持向量的数目下降明显,进而可以加快支持向量机的分类速度。

4　实验仿真

实验 1　合成数据 1

该实验是先随机产生两类均匀分布的样本,第 1 类样本坐标分布范围为 $U([0,4]\times[0,10])$,第 2 类样本坐标分布范围是 $U([6,10]\times[0,10])$,两类样本共 6 000 个,其中 4 000 个为训练样本,2 000 个为测试样本,然后分别用标准 SVM 和本文方法进行实验。考虑到混叠的影响,可人为地将 20 个训练样本进行错误标记。表 1 给出了实验比较结果,实验结果表明,本文提出的方法不但大大减少了训练样本,而且剔除了部分奇异数据的影响,并提高了分类识别性能。值得指出的是,这里边界向量选择的数目是可以控制的。从实验结果可以看出,奇异训练数据对于支持向量的数目与分类精度都有影响,可见混叠度分析是非常有必要的。

实验 2　合成数据 2

该实验与文献[6]相同,首先产生两类同心圆样本,第 1 类样本的半径是均匀分布,其分布参数范围是 $U[0,6]$,第 2 类样本的半径分布参数为 $U[5,10]$,两类样本共 4 000 个,其中 2 000 个为训练样本,2 000 个为测试样本;然后分别用标准 SVM 和本

表 1　不同方法性能对比 1

Tab. 1　Comparison results with different methods

算法	训练样本数	删除奇异数据数	边界向量数	支持向量数	训练时间(s)	检验样本数	识别率(%)
标准 SVM	4 000	无奇异数据	无	3	1 534	2 000	100
标准 SVM	4 000	无	无	73	1 791	2 000	76.8
本文方法	4 000	19	17	7	135	2 000	96.7

文提出的方法进行实验，同时分别采用高斯核函数与二次多项式核函数进行实验，并比较了不同边界向量选取方法对实验结果的影响，本文选择了不同混叠门限进行实验。实验结果如表 2 所示，由表 2 可见，方法 1 与方法 2 的差别主要体现在边界向量选择的数目不同，从实验结果可以看出，边界向量的数目对于分类结果影响不大，但是对于训练速度却有一定程度的影响。

表 2 不同方法性能对比 2

Tab. 2 Comparison results with different methods

方法	核函数	边界向量数	支持向量数	训练时间(s)	识别率(%)
SVM	高斯	无	307	1 947	89.97
SVM	多项式	无	319	1 605	90.13
本文方法 1	高斯	506	313	193	90.65
本文方法 1	多项式	506	319	197	90.34
本文方法 2	高斯	700	315	215	90.67
本文方法 2	多项式	700	321	220	90.38

实验 3 真实数据

利用 MNIST 手写数字数据库进行了仿真实验，实验时，首先从全部 MNIST 手写数字数据库的 60 000个训练样本库选择了 3 000 个训练样本("3"和"5")，然后从 MNIST 全部 10 000 个测试样本库中选择 1 000 个作为测试样本，并采用量化 PCA 投影矢量作为训练特征，对比实验结果如表 3 所示。

表 3 对比实验结果 2

Tab. 3 Comparison results with different methods

方法	核函数	支持向量数	边界向量数	分类率(%)	训练时间(s)
标准 SVM	多项式	109	无	98.4	2 011
标准 SVM	高斯	119	无	98.9	2 086
本文方法	多项式	114	397	99.2	375
本文方法	高斯	105	401	99.3	402

从上面的 3 个对比实验可以看出，本文提出的方法无论对于海量合成数据还是真实数据，在保持分类精度的情况下，通过类间最近邻距离排序、混叠度分析，可大大减少训练样本数量，进而可减少计算量，因此加快了 SVM 训练算法的计算速度。当样本容量巨大(大于6 000)，如果不采用基于最近邻类间距离的子空间选择，那么经典 SVM 方法就会因为样本数量巨大而求解困难。

5 结 论

本文根据支持向量的分布特性，提出了基于支持向量几何分析的支持向量机快速算法，该算法先根据类间最近邻距离选择出近邻向量，再依据混叠度分析去除奇异训练样本的影响，并用得到的纯洁的边界向量来代替整个海量训练样本集。该算法一方面加快了支持向量机的训练速度，使得任意海量训练样本情况下的支持向量机的训练问题可解；同时因为通过去掉混叠严重的奇异样本使得支持向量的数目明显下降，所以加快了支持向量机分类速度，这就为解决 SVM 识别的实时性提供了可能。与文献[6]提出的向量投影支持向量预选取方法相比，本文提出的方法适合于更加复杂数据分布形状(比如同类样本有多个中心的情况)。

参考文献(References)

1 Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Wiley, 1998. chapter 5.

2 Hu W J, Song Q. An accelerated decomposition algorithm for robust support vector machines[J]. IEEE Transactions on Circuits and Systems—II: Express Briefs, 2004, **51**(5): 234 ~ 240.

3 Li Hong-lian, Wang Chun-hua, Yuan Bao-zong. An improved SVM: NN-SVM[J]. Chinese Journal of Electronics, 2004, **13**(2): 321 ~ 324.

4 Li Rong, Ye Shi-wei, Shi Zhong-zhi. SVM-KNN classifier—A new method of improving the accuracy of SVM Classifier[J]. Acta Electronica Sinica, 2002, **30**(5): 746 ~ 749. [李蓉, 叶世伟, 史忠植. SVM-KNN 分类器——一种提高 SVM 分类精度的新方法[J]. 电子学报, 2002, **30**(5): 746 ~ 749.]

5 Keerthi S S, Shevade S K, Bhattacharyya C, *et al*. Fast iterative nearest point for support vector machine classifier design[J]. IEEE Transactions on Neural Networks, 2000, **11**(1): 124 ~ 136.

6 Li Qing, Jiao Li-cheng, Zhou Wei-da. Pre-extracting support vector for support vector machine based on vector projection[J]. Chinese Journal of Computers, 2005, **28**(2): 145 ~ 151. [李青, 焦李成, 周伟达. 基于向量投影的支持向量预选取[J]. 计算机学报, 2005, **28**(2): 145 ~ 151.]