一种改进的支持向量机 NN-SVM

李红莲[1)] 王春花[2)] 袁保宗[1)]

[1)](北方交通大学信息科学研究所 北京 100044)

[2)](北京三星通信技术研究所 北京 100081)

摘 要 支持向量机(SVM)是一种较新的机器学习方法,它利用靠近边界的少数向量构造一个最优分类超平面.在训练分类器时,SVM 的着眼点在于两类的交界部分,那些混杂在另一类中的点往往无助于提高分类器的性能,反而会大大增加训练器的计算负担,同时它们的存在还可能造成过学习,使泛化能力减弱.为了改善支持向量机的泛化能力,该文在其基础上提出了一种改进的 SVM——NN-SVM:它先对训练集进行修剪,根据每个样本与其最近邻类标的异同决定其取舍,然后再用 SVM 训练得到分类器.实验表明,NN-SVM 相比 SVM 在分类正确率、分类速度以及适用的样本规模上都表现出了一定的优越性.

关键词 支持向量机;最近邻;修剪

中图法分类号 TP391

An Improved SVM: NN-SVM

LI Hong-Lian[1)] WANG Chun-Hua[2)] YUAN Bao-Zong[1)]

[1)](*Institute of Information Science, Northern Jiaotong University, Beijing* 100044)

[2)](*Beijing Samsung Communication Technology Research Institute, Beijing* 100081)

Abstract A support vector machine constructs an optimal hyperplane from a small set of samples near the boundary. This makes it sensitive to these specific samples and tends to result in machines either too complex with poor generalization ability or too imprecise with high training error, depending on the kernel parameters. SVM focuses on the samples near the boundary in training time, and those samples intermixed in another class are usually no good to improve the classifier's performance, instead they may greatly increase the burden of computation and their existence may lead to overlearning and decrease the generalization ability. In order to improve the generalization ability we present an improved SVM: NN-SVM. It first prunes the training set, reserves or deletes a sample according to whether its nearest neighbor has same class label with itself or not, then trains the new set with SVM to obtain a classifier. Experiment results show that NN-SVM is better than SVM in speed and accuracy of classification.

Keywords support vector machines; nearest neighbor; pruning

1 引 言

支持向量机 SVM(Support Vector Machines)以其泛化能力强著称,并因此得到了人们的青睐.它仅仅考虑类的边界情况,用较少的向量(支持向量)来分类.一方面,它要使得两类边界之间的宽度最大(分类边界将居于两类边界的正中间);另一方面,它

收稿日期:2002-05-22;修改稿收到日期:2002-10-21.本课题得到国家自然科学重点基金(69789301)和国家"九七三"重点基础研究发展规划项目(G19980305011)资助.**李红莲**,1971 年生,博士研究生.研究方向为模式识别、语音识别与理解、语音上网及知识获取等.E-mail: honglianli@263.net.**王春花**,1971 年生,博士,研究方向为数据仓库、数据通信等.**袁保宗**,1932 年生,博士,教授,博士生导师,研究领域为信号与信息处理、多媒体视听信息处理、计算机视觉、图像信号处理、语音信号处理、计算机图形学与虚拟现实.

还要使得错分的代价不要太高.两方面权衡的结果,就是要使得下面的式子取得最小值:

$$1/h + l(e),$$

其中 h 表示两类的边界之间的宽度,$l(e)$ 表示错分带来的损失.

争取最大的边界宽度是为了保证分类器具有较强的泛化能力,同时这里的"最大"是有条件的,就是不能付出太多的错分代价,SVM 是这两种要求折衷的结果.在训练分类器时,SVM 的着眼点在于两类的交界部分,那些混杂在另一类中的点往往无助于提高分类器的性能,反而会大大增加训练器的计算负担,同时它们的存在还可能造成过学习,使泛化能力减弱.基于这种想法,本文提出了一种改进的 SVM——NN-SVM:它先对训练集进行修剪,根据每个样本与其最近邻(nearest neighbor)类标的异同决定其取舍,然后再用 SVM 训练得到分类器.实验表明,NN-SVM 相比 SVM 在分类正确率、分类速度以及适用的样本规模上都较优.

本文第 2 节介绍 SVM,第 3 节给出 NN-SVM 的算法,第 4 节给出实验结果及相关分析,最后是结论.

2 SVM

SVM 是支持向量机的简称,是统计学习理论中最年轻的内容,也是最实用的部分.其核心内容是在 1992 到 1995 年间提出的,目前仍处在不断发展阶段[1].详细内容参见文献[2~6].支持向量机可用于模式识别、回归分析、主成分分析等.下面以模式分类为例来介绍支持向量机的含义.

给定一组训练数据 $(x_1, y_1), \cdots, (x_l, y_l)$, $x_i \in R^n$, $y_i \in \{+1, -1\}$, $i = 1, 2, \cdots, l$.我们要寻找一个分类规则 $I(x)$,使它能对未知类别的新样本(新样本与训练样本独立同分布)作尽可能正确的划分.

支持向量机用于分类问题其实就是寻找一个最优分类超平面,把此平面作为分类决策面.同时它还通过引进核函数巧妙地解决了在将低维空间向量映射到高维空间向量时带来的"维数灾难"问题.

2.1 最优分类超平面

在训练集线性可分情形,SVM 就是要构造一个最优超平面

$$(\omega \cdot x) + b = 0 \tag{1}$$

这个超平面既要满足下面的约束条件

$$y_i[(\omega \cdot x_i) + b] \geq 1, \quad i = 1, 2, \cdots, l \tag{2}$$

同时还要使下面的函数取得最小值

$$\phi(\omega) = \frac{1}{2}\|\omega\|^2 = \frac{1}{2}(\omega \cdot \omega) \tag{3}$$

通过求解最优化问题可得最优超平面的形式如下

$$\sum_{SV} y_i \alpha_i^0 (x \cdot x_i) + b_0 = 0 \tag{4}$$

其中 SV 表示支持向量,α_i^0 是拉格朗日乘子.

在训练集线性不可分时,我们引进松弛因子 $\xi_i \geq 0$ 及惩罚参数 C.这时需要做的是在约束 $y_i((\omega \cdot x_i) + b) \geq 1 - \xi_i$, $i = 1, 2, \cdots, l$ 下最小化函数 $\phi(\xi) = \frac{1}{2} \cdot \|\omega\|^2 + C\sum_{i=1}^{l} \xi_i$.类似可得最优超平面.有了最优超平面,分类规则或分类函数只要取 $I(x) = \mathrm{sgn}\left(\sum_{SV} y_i \alpha_i^0 (x \cdot x_i) + b_0\right)$ 即可.

2.2 核函数

支持向量机特点之一在于核函数的引入.我们知道,低维空间向量集往往难于划分.因此,自然想把它们映射到高维空间,但随之而来的是计算复杂度的大大增加,核函数巧妙地解决了这个问题.

若函数 $K(x, y)$ 满足 Mercer 条件[2],则 $K(x, y) = \phi(x) \cdot \phi(y)$,其中 ϕ 表示某个映射(未必知其具体表达式).这样,只要适当选取核函数我们就可以得到对应高维空间的分类函数

$$I(x) = \mathrm{sgn}\left(\sum_{SV} y_i \alpha_i^0 K(x, x_i) + b_0\right) \tag{5}$$

其中,$\phi(x)$,$\phi(y)$ 是比 x,y 更高维的向量(注意我们不必知道 ϕ 的具体形式),由于 $K(x, y)$ 只涉及 x,y,因此计算没有涉及高维运算.

我们将分类函数(决策函数)类型为式(5)的学习机称为支持向量机.

3 NN-SVM

尽管支持向量机追求的目标是较强的泛化能力,但相对于具体的样本集,也可能出现过学习的问题.如两类样本集混叠较严重时,SVM 的决策面可能由于过分复杂反而降低了其泛化能力.

文献[7]提出了一种 ESVM (Editing Supporting Vector Machines),其基本做法是:

首先用 SVM 对训练集学习得到决策边界,去掉决策边界附近一定区域内的样本以及错分的样本,然后再对新训练样本集重新用 SVM 学习得到新的决策边界.在必要的情况下,对最初的训练样本集用新决策边界编辑,去掉错分的样本,得到另一个新的训练集,再对它训练得到更新的决策边界.

文献[3]的做法较为复杂,需要反复使用 SVM 训练.本文提出了另一种改进的 SVM——NN-SVM:它先对训练集进行修剪,根据每个样本与其最近邻(nearest neighbor)类标的异同决定其取舍,然后再用 SVM 训练得到分类器.相比文献[7]的做法,我们的做法非常简捷,且实验表明,与 SVM 相比,NN-SVM 不单在分类正确率上有了较大提高,而且分类速度更快,并能适用更大规模的训练样本集.

我们采取下面的策略对训练集进行修剪:

首先找出每一个点的最近邻,然后对每一个点,如果该点与其最近邻属于同类,则保留此点;如果该点与其最近邻属于异类,将该点删除.

采用欧氏距离作为两个向量之间的距离,即

设

$$x_i=(x_i^1,x_i^2,\cdots,x_i^n),x_j=(x_j^1,x_j^2,\cdots,x_j^n),$$

则 x_i与 x_j之间的距离定义为

$$D(x_i,x_j)=\sqrt{\sum_{k=1}^{n}(x_i^k-x_j^k)^2},$$

一个样本的最近邻就是在上述定义下与其距离最近的样本.

下面我们给出上述方法的实现算法.

给定一个训练集$(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)$,$x_i\in R^n$,$y_i\in\{1,-1\}$,$i=1,2,\cdots,m$.将训练集表示为矩阵

$$\boldsymbol{TR}_{m\times(n+1)}=(\boldsymbol{XY}),\text{其中 }\boldsymbol{X}=\begin{pmatrix}x_1\\ \vdots\\ x_m\end{pmatrix},\boldsymbol{Y}=\begin{pmatrix}y_1\\ \vdots\\ y_m\end{pmatrix}.$$

修剪算法如下:

1. 找出每一个向量的最近邻;

(1)求出每个点与其它各点的距离,与自身的距离定义为∞

for $p=1$ to m

{ $\boldsymbol{Z}_{1\times m}=(z_{ij}),z_{ij}=\infty,i=1,j=1,2,\cdots,m$;

for $q=1$ to m

{if $q\neq p$, $z_{1q}=D(x_p,x_q)$;}

}

(2)找出最短距离及相应点(最近邻)

$\boldsymbol{NN}_{m\times1}=(nn_{ij})$, $nn_{ij}=1,i=1,2,\cdots,m$, $j=1$

$t=1$; $value=z_{11}$;

for $q=1$ to m

{ if $z_{1q}<value$ {$value=z_{1q}$; $t=q$;}

$nn_{p1}=t$; }

2. 判断每个向量的类标与其最近邻是否一致,分别标记为 1 与−1

$\boldsymbol{L}_{m\times1}=(l_{ij}),l_{ij}=1$, $i=1,2,\cdots,m,j=1$;

for $p=1$ to m

{if $y_p\neq y_{nn_p}$, $l_{p1}=-1$; }

3. 删除与最近邻类标不一致的向量

$i=0$;

for $p=1$ to m

{ if $l_{(p-i)1}=-1$

{ 删除矩阵 $\boldsymbol{TR}$ 及 $\boldsymbol{L}$ 的第 $p-i$ 行,新矩阵仍设为 $\boldsymbol{TR}$ 及 $\boldsymbol{L}$; $i=i+1$; }

}

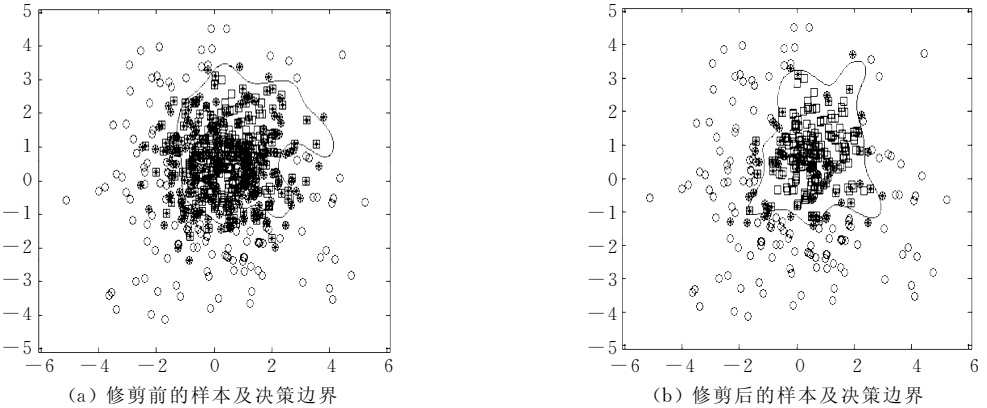
经过上述 3 步后就可得到修剪后的训练集 $\boldsymbol{TR}$.

我们把上述先利用最近邻(nearest neighbor)对训练集进行修剪,然后再用 SVM 训练得到分类器的方法称为 NN-SVM.

相对于 SVM,NN-SVM 有以下优点:

(1)分类正确率有望提高

由于修剪了训练集,NN-SVM 的分类边界相比 SVM 的过于复杂的分类边界有所简化(见图 1),因而其泛化能力可能更强,分类正确率可能更高.第 4 节的实验结果证实了这种想法.可见 NN-SVM 是解决由于两类混叠严重而造成分类器过学习和泛化能力减弱问题的有效途径.



(a) 修剪前的样本及决策边界　　(b) 修剪后的样本及决策边界

图 1　修剪前后样本的的决策边界(修剪后,样本数目有了大幅度缩减,同时分类边界有所改变.后面的实验数据表明,这样的改变提高了分类精度)

(2)分类所用时间更短

训练集经过修剪后,分类器的支持向量大大减少,而分类所用时间与支持向量的个数是成正比的(见式(5)),因此大大节省了分类时间.

(3)可用于更大训练集

由于修剪过程使较大的训练集变小,因此在同样的硬件条件下,NN-SVM 可适用于更大的训练集.

当然,我们应该指出,上述优点的获得也是付出了一定的代价的:那就是修剪过程需要额外的时间,但是这一点代价与上述任何一种收益相比都是微不足道的.因为更高的正确率、更快的分类速度以及训练更大的样本集是我们追求的首要目标.

4 实验结果

(1)实验环境

我们在 PC 机(奔腾 1.4G,256M 内存)上,利用 http://svm.first.gmd.de/提供的 Matlab SVM 软件工具包以及我们所编制的数据修剪程序进行实验.所采用的测试数据如下:

数据 1,ringnorm,该数据集是 Leo Breiman 生成的用于两类划分的样本集,每一类都是取自一个 20 维的多变量正态分布.类 1 的均值为 0,方差为单位元的四倍;类 2 的均值为$(a,a,\cdots,a)$,方差为单位元,其中 $a=2/\mathrm{sqrt}(20)$.在实验中通过截取某 k 维得到相应 k 维正态分布的样本集.

数据 2,letter,此数据集是 26 个大写字母打印体的 16 维特征向量集.在实验中把某一个字母作为一类,其它的 25 个字母作为另一类.

数据 3,SVM 工具包自带的数据集 *iris*,它是三种植物的 4 维特征向量集.

以上数据都可以从网站 http://svm.first.gmd.de/上得到.

(2)实验结果及分析

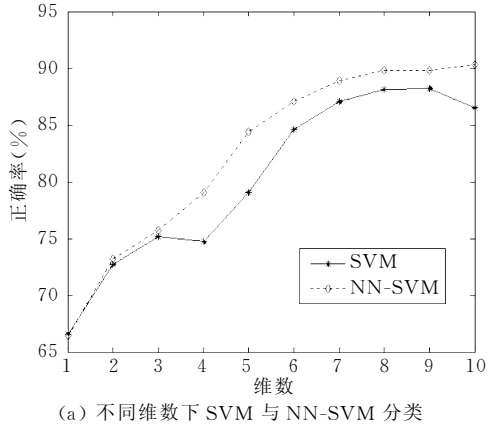
实验中,核函数使用高斯核,其中 $\sigma=0.5$,惩罚参数 $C=100$.我们重点对正态分布样本集进行了训练,实验结果如图 2～图 5.这些图都是相对于数据集 1 实验的结果,数据集 1 为 7400 个 20 维向量的集合,每个向量还附带一个类标.我们通常截取 20 维向量的前 K 个分量构成 K 维向量,取前 N 个样本作为训练集(图中横坐标对应 K 或 N 的具体取值),后 3400 个样本作为测试集.由于数据集 1 本身是随机排列的,向量的各分量之间也是独立的,因此这样选取是合理的,并且我们尝试了不同的选取方法,实验结果类似.图 2 表明,NN-SVM 分类正确率有了明显的提高.图 3 表明,NN-SVM 的支持向量数大大减少了,从而分类时间也大大减少了(图 4).图 5 所示修剪后样本集有较大程度的减小,表明在样本集太大(受硬件条件限制)无法直接进行训练的情况下,我们有可能通过修剪样本集使训练能够进行.

对其它数据实验,得到了类似的结果.

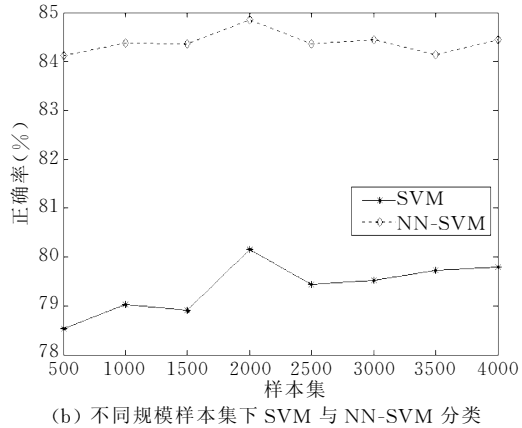
下面指出需要注意的几个问题:

(1) NN-SVM 特别适用于混叠较为严重的数据集,即两类的交叉区域较大不易划分的情形.对于混叠很轻、容易划分的情形,NN-SVM 虽然仍具备上述优点,但优势不明显.

(2) 我们知道,当样本集足够多的时候,分类器的正确率将逼近一个极限,这时候 SVM 与 NN-SVM 的分类正确率将同时达到极限,因此这时再比较孰优孰劣显然已没有意义.而在样本集严重不足的情况下,显然保留现有的样本集更为明智,因此这时候再对样本集进行修剪显然不合时宜,在这种情况下,



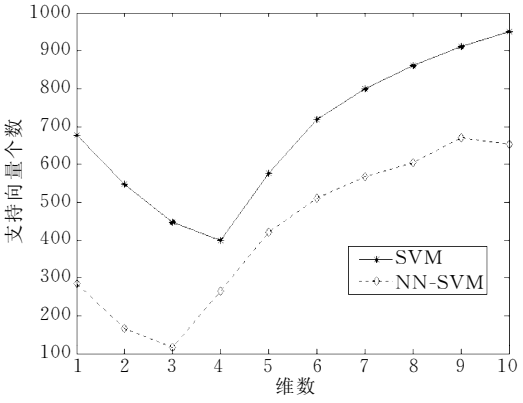
(a) 不同维数下 SVM 与 NN-SVM 分类正确率的比较(样本数1000)



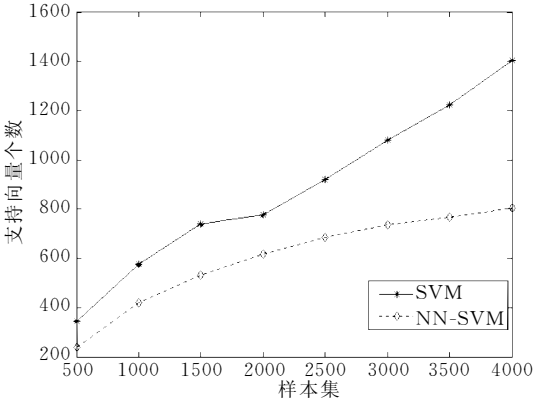
(b) 不同规模样本集下 SVM 与 NN-SVM 分类正确率的比较(维数5)

图 2 正确率比较

没有必要使用 NN-SVM. 也就是说,当样本集的规模处于某一范围之内(既不是充分多也不是特别少)时,NN-SVM 会表现出较为明显的优势.

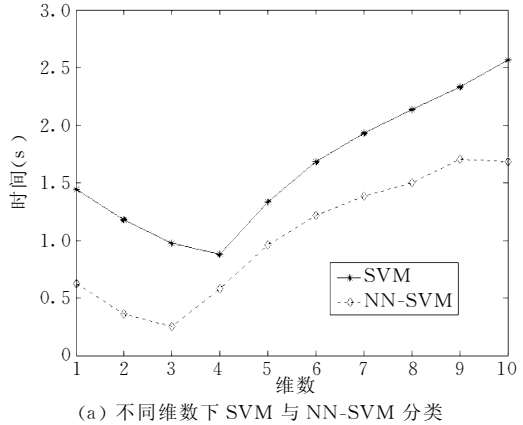


(a) 不同维数下 SVM 与 NN-SVM 支持向量个数的比较(样本数1000)

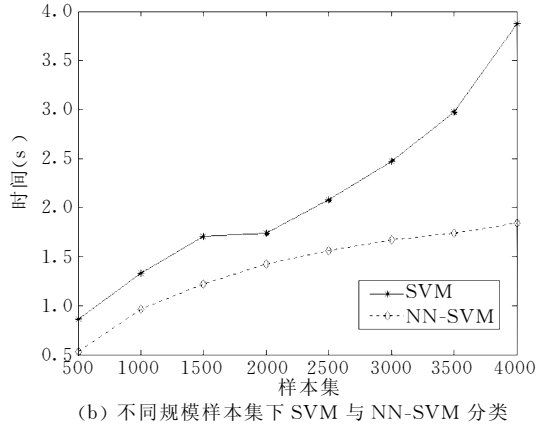


(b) 不同规模样本集下 SVM 与 NN-SVM 支持向量个数的比较(维数5)

图 3　支持向量个数的比较

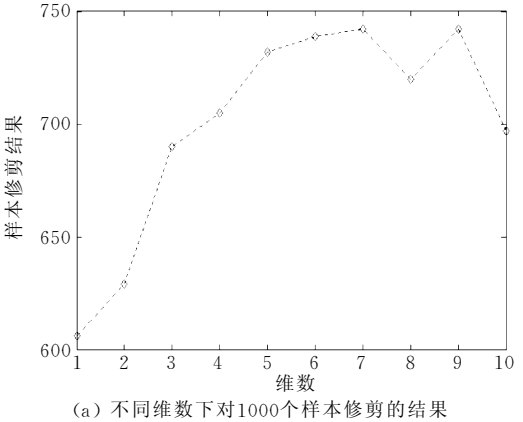


(a) 不同维数下 SVM 与 NN-SVM 分类所用时间的比较(样本数1000)

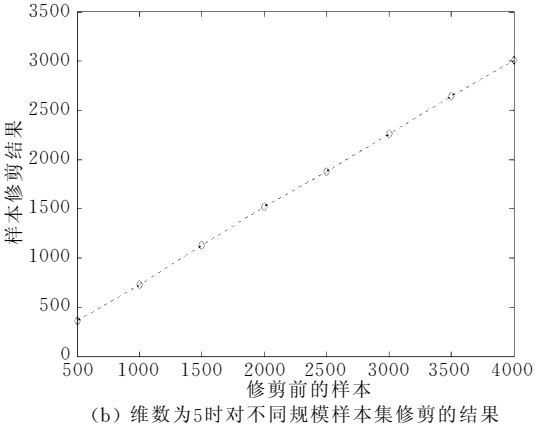


(b) 不同规模样本集下 SVM 与 NN-SVM 分类所用时间的比较(维数5)

图 4　分类时间的比较



(a) 不同维数下对1000个样本修剪的结果



(b) 维数为5时对不同规模样本集修剪的结果

图 5　样本修剪结果

5　结束语

本文给出了一种改进的支持向量机——NN-SVM:它首先对训练集进行修剪,根据每个样本与其最近邻类标的异同决定其取舍;然后用 SVM 训练分类器. 实验表明,相比 SVM,NN-SVM 的分类正确率更高、分类速度更快、能训练的样本集更大. 可见 NN-SVM 是解决由于两类混叠严重而造成分类器过学习和泛化能力减弱问题的有效途径. 基于本文的方法还有

许多需要研究的问题,例如样本混叠程度的数量化描述,样本的规模在什么样的范围内时,NN-SVM相比SVM的优势更明显,有没有更快捷的修剪算法等.另外,文献[8]讨论了近邻法参考样本集的最优选择问题,或许对高效修剪有帮助,这将是需要进一步研究的问题.

参 考 文 献

1 Zhang Xue-Gong. Introduction to statistical learning theory and support vector machines. Acta Automatica Sinica, 2000,26 (1): 32～42(in Chinese)
(张学工. 关于统计学习理论与支持向量机. 自动化学报,2000, 26(1):32～42)

2 Vapnik V N. An overview of statistical learning theory. IEEE Transactions on Neural Networks, 1999, 10 (5): 988～999

3 Vapnik V N. Statistical Learning Theory. 2nd ed. New York: Springer-Verlag: 1999

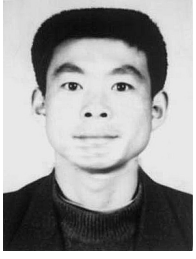
4 Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, and Bernhard Schölkopf. An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks, 2001, 12 (2): 181～201

5 Burges C J C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2 (2): 121～167

6 Hearst M A, Dumais S T, Osman E, Platt J, Scholkopf B. Support Vector Machines. IEEE Intelligent Systems, 1998, 13 (4): 18～28

7 Ke Hai-Xin, Zhang Xue-Gong. Editing support vector machines. In: Proceedings of International Joint Conference on Neural Networks, Washington, USA, 2001, 2: 1464～1467

8 Zhang Hong-Bin, Sun Guang-Yu. Optimal selection of reference subset for nearest neighbor classification. Acta Electronica Sinica, 2000, 28 (11): 16～21(in Chinese)
(张鸿宾,孙广煜. 近邻法参考样本集的最优选择. 电子学报, 2000,28(11):16～21)



LI Hong-Lian, born in 1971, Ph. D. candidate. His main research interests include machine learning and speech recognition and understanding.

WANG Chun-Hua, born in 1971, Ph. D.. Her main research interests include machine learning, data mining and data communication.

YUAN Bao-Zong, born in 1932, Ph. D., professor and Ph. D. supervisor. His research interests include digital signal processing, speech signal processing, image processing, computer vision, computer graphics, multimedia information processing and data communication.