DOI: 10.16383/j.aas.2000.01.005
第26卷 第1期 自 动 化 学 报 Vol. 26, No. 1
2000年1月 ACTA AUTOMATICA SINICA Jan., 2000

综述与评论

关于统计学习理论与支持向量机[1)]

张学工
(清华大学自动化系,智能技术与系统国家重点实验室 北京 100084)

摘 要 模式识别、函数拟合及概率密度估计等都属于基于数据学习的问题,现有方法的重要基础是传统的统计学,前提是有足够多样本,当样本数目有限时难以取得理想的效果.统计学习理论(SLT)是由Vapnik等人提出的一种小样本统计理论,着重研究在小样本情况下的统计规律及学习方法性质. SLT为机器学习问题建立了一个较好的理论框架,也发展了一种新的通用学习算法——支持向量机(SVM),能够较好的解决小样本学习问题.目前,SLT和SVM已成为国际上机器学习领域新的研究热点.本文是一篇综述,旨在介绍SLT和SVM的基本思想、特点和研究发展现状,以引起国内学者的进一步关注.

关键词 统计学习理论,支持向量机,机器学习,模式识别.

INTRODUCTION TO STATISTICAL LEARNING THEORY AND SUPPORT VECTOR MACHINES

ZHANG Xuegong
(*Dept. of Automation, Tsinghua University, Beijing* 100084)
(*State Key Laboratory of Intelligent Technology and Systems of China*)

Abstract Data-based machine learning covers a wide range of topics from pattern recognition to function regression and density estimation. Most of the existing methods are based on traditional statistics, which provides conclusion only for the situation where sample size is tending to infinity. So they may not work in practical cases of limited samples. Statistical Learning Theory or SLT is a small-sample statistics by Vapnik *et al.*, which concerns mainly the statistic principles when samples are limited, especially the properties of learning procedure in such cases. SLT provides us a new framework for the general learning problem, and a novel powerful learning method called Support Vector Machine or SVM, which can solve small-sample learning problems better. It is believed that the study of SLT and SVM is becoming a new hot area in the field of machine learning. This review introduces the basic ideas of SLT and SVM, their major characteristics and some current research

1)本文受到国家自然科学基金赞助,项目编号为69885004.
收稿日期 1998-08-24 收修改稿日期 1999-04-27

trends.

Key words Statistical learning theory, support vector machine, machine learning, pattern recognition.

1 引言

基于数据的机器学习是现代智能技术中的重要方面,研究从观测数据(样本)出发寻找规律,利用这些规律对未来数据或无法观测的数据进行预测.包括模式识别、神经网络等在内,现有机器学习方法共同的重要理论基础之一是统计学.传统统计学研究的是样本数目趋于无穷大时的渐近理论,现有学习方法也多是基于此假设.但在实际问题中,样本数往往是有限的,因此一些理论上很优秀的学习方法实际中表现却可能不尽人意.

与传统统计学相比,统计学习理论(Statistical Learning Theory或 SLT)是一种专门研究小样本情况下机器学习规律的理论. V. Vapnik等人从六、七十年代开始致力于此方面研究[1],到九十年代中期,随着其理论的不断发展和成熟,也由于神经网络等学习方法在理论上缺乏实质性进展,统计学习理论开始受到越来越广泛的重视[2,3].

统计学习理论是建立在一套较坚实的理论基础之上的,为解决有限样本学习问题提供了一个统一的框架.它能将很多现有方法纳入其中,有望帮助解决许多原来难以解决的问题(比如神经网络结构选择问题、局部极小点问题等);同时,在这一理论基础上发展了一种新的通用学习方法——支持向量机(Support Vector Machine或 SVM),它已初步表现出很多优于已有方法的性能.一些学者认为,SLT和 SVM正在成为继神经网络研究之后新的研究热点,并将有力地推动机器学习理论和技术的发展[3].

我国早在八十年代末就有学者注意到统计学习理论的基础成果[4],但之后较少研究,目前只有少部分学者认识到这个重要的研究方向.本文旨在向国内介绍统计学习理论和支持向量机方法的基本思想和特点,以使更多的学者能够看到它们的优势从而积极进行研究.文章第二节给出机器学习问题的一般表示,并简要讨论现有方法的一些问题;第三节介绍 SLT的基本思想和最有影响的结论;第四节介绍 SVM方法的原理、应用及由此发展出的其它方法;第五节是讨论.

2 机器学习的基本问题

2.1 问题的表示

机器学习的目的是根据给定的训练样本求对某系统输入输出之间依赖关系的估计,使它能够对未知输出作出尽可能准确的预测.可以一般地表示为:变量 y与 x存在一定的未知依赖关系,即遵循某一未知的联合概率 $F(x,y)$,(x和 y之间的确定性关系可以看作是其特例),机器学习问题就是根据 n个独立同分布观测样本

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n), \tag{1}$$

在一组函数 $\{f(x,w)\}$中求一个最优的函数 $f(x,w_0)$对依赖关系进行估计,使期望风险

$$R(w)=\int L(y,f(x,w))\,\mathrm{d}F(x,y) \tag{2}$$

最小.其中,$\{f(x,w)\}$称作预测函数集,w为函数的广义参数,$\{f(x,w)\}$可以表示任何函数集;$L(y,f(x,w))$为由于用 $f(x,w)$对 y进行预测而造成的损失,不同类型的学习问题有不同形式的损失函数.预测函数也称作学习函数、学习模型或学习机器.

有三类基本的机器学习问题,即模式识别、函数逼近和概率密度估计.对模式识别问题,输出 y是类别标号[1],两类情况下 $y=\{0,1\}$或$\{1,-1\}$,预测函数称作指示函数,损失函数可以定义为

$$L(y,f(\boldsymbol{x},w))=\begin{cases}0, & \text{if } y=f(\boldsymbol{x},w),\\ 1, & \text{if } y\neq f(\boldsymbol{x},w),\end{cases} \tag{3}$$

使风险最小就是 Bayes决策中使错误率最小.在函数逼近问题中,y是连续变量(这里假设为单值函数),损失函数可定义为

$$L(y,f(\boldsymbol{x},w))=(y-f(\boldsymbol{x},w))^2, \tag{4}$$

即采用最小平方误差准则.而对概率密度估计问题,学习的目的是根据训练样本确定 x的概率密度.记估计的密度函数为 $p(\boldsymbol{x},w)$,则损失函数可以定义为

$$L(p(\boldsymbol{x},w))=-\log p(\boldsymbol{x},w). \tag{5}$$

2.2 经验风险最小化

在上面的问题表述中,学习的目标在于使期望风险最小化,但是,由于我们可以利用的信息只有样本(1),(2)式的期望风险并无法计算,因此传统的学习方法中采用了所谓经验风险最小化(ERM)准则,即用样本定义经验风险

$$R_{\mathrm{emp}}(w)=\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(\boldsymbol{x}_i,w)), \tag{6}$$

作为对(2)式的估计,设计学习算法使它最小化.对损失函数(3),经验风险就是训练样本错误率;对(4)式的损失函数,经验风险就是平方训练误差;而采用(5)式损失函数的 ERM准则就等价于最大似然方法.

事实上,用 ERM准则代替期望风险最小化并没有经过充分的理论论证,只是直观上合理的想当然做法,但这种思想却在多年的机器学习方法研究中占据了主要地位.人们多年来将大部分注意力集中到如何更好地最小化经验风险上,而实际上,即使可以假定当 n趋向于无穷大时(6)式趋近于(2)式,在很多问题中的样本数目也离无穷大相去甚远.那么在有限样本下 ERM准则得到的结果能使真实风险也较小吗?

2.3 复杂性与推广能力

ERM准则不成功的一个例子是神经网络的过学习问题.开始,很多注意力都集中在如何使 $R_{\mathrm{emp}}(w)$更小,但很快就发现,训练误差小并不总能导致好的预测效果.某些情况下,训练误差过小反而会导致推广能力的下降,即真实风险的增加,这就是过学习问题.

之所以出现过学习现象,一是因为样本不充分,二是学习机器设计不合理,这两个问题是互相关联的.设想一个简单的例子,假设有一组实数样本$\{x,y\}$,y取值在$[0,1]$之间,那么不论样本是依据什么模型产生的,只要用函数 $f(x,\Gamma)=\sin(\Gamma x)$去拟合它们($\Gamma$是待定参数),总能够找到一个 Γ使训练误差为零,但显然得到的"最优"函数并不能正确代表

1)这里暂时没有讨论非监督模式识别问题.实际上,如何在非监督模式识别问题中应用统计学习理论正是当前值得研究的课题之一.

真实的函数模型.究其原因,是试图用一个十分复杂的模型去拟合有限的样本,导致丧失了推广能力.在神经网络中,若对有限的样本来说网络学习能力过强,足以记住每个样本,此时经验风险很快就可以收敛到很小甚至零,但却根本无法保证它对未来样本能给出好的预测.学习机器的复杂性与推广性之间的这种矛盾同样可以在其它学习方法中看到.

文献[3]给出了一个实验例子,在有噪声条件下用模型 $y= x^2$产生 10个样本,分别用一个一次函数和一个二次函数根据 ERM原则去拟合,结果显示,虽然真实模型是二次,但由于样本数有限且受噪声的影响,用一次函数预测的结果更好.同样的实验进行了 100次,71%的结果是一次拟合好于二次拟合.

由此可看出,有限样本情况下,1)经验风险最小并不一定意味着期望风险最小; 2)学习机器的复杂性不但应与所研究的系统有关,而且要和有限数目的样本相适应.我们需要一种能够指导我们在小样本情况下建立有效的学习和推广方法的理论.

3 统计学习理论的核心内容

统计学习理论就是研究小样本统计估计和预测的理论,主要内容包括四个方面[2]:

1) 经验风险最小化准则下统计学习一致性的条件;

2) 在这些条件下关于统计学习方法推广性的界的结论;

3) 在这些界的基础上建立的小样本归纳推理准则;

4) 实现新的准则的实际方法(算法).

其中,最有指导性的理论结果是推广性的界,与此相关的一个核心概念是 VC维.

3.1 VC维

为了研究学习过程一致收敛的速度和推广性,统计学习理论定义了一系列有关函数集学习性能的指标,其中最重要的是 VC维(Vapnik-Chervonenkis Dimension).模式识别方法中 VC维的直观定义是: 对一个指示函数集,如果存在 h 个样本能够被函数集中的函数按所有可能的 2^h 种形式分开,则称函数集能够把 h 个样本打散;函数集的 VC维就是它能打散的最大样本数目 h.若对任意数目的样本都有函数能将它们打散,则函数集的 VC维是无穷大.有界实函数的 VC维可以通过用一定的阈值将它转化成指示函数来定义.

VC维反映了函数集的学习能力,VC维越大则学习机器越复杂(容量越大).遗憾的是,目前尚没有通用的关于任意函数集 VC维计算的理论,只对一些特殊的函数集知道其 VC维.比如在 n 维实数空间中线性分类器和线性实函数的 VC维是 $n+1$,而上一节例子中 $f(x,\alpha)= \sin(\alpha x)$的 VC维则为无穷大.对于一些比较复杂的学习机器(如神经网络),其 VC维除了与函数集(神经网结构)有关外,还受学习算法等的影响,其确定更加困难.对于给定的学习函数集,如何(用理论或实验的方法)计算其 VC维是当前统计学习理论中有待研究的一个问题[5].

3.2 推广性的界

统计学习理论系统地研究了对于各种类型的函数集,经验风险和实际风险之间的关系,即推广性的界[2].关于两类分类问题,结论是: 对指示函数集中的所有函数(包括使经验风险最小的函数),经验风险 $R_{emp}(w)$和实际风险 $R(w)$之间以至少 $1-\eta$的概率满足如下关系[6]:

$$R(w) \leqslant R_{\mathrm{emp}}(w) + \sqrt{\frac{h(\ln(2n/h)+1) - \ln(Z/4)}{n}}, \tag{7}$$

其中 h 是函数集的 VC维,n 是样本数.

这一结论从理论上说明了学习机器的实际风险是由两部分组成的: 一是经验风险 (训练误差),另一部分称作置信范围,它和学习机器的 VC维及训练样本数有关.可以简单地表示为

$$R(w) \leqslant R_{\mathrm{emp}}(w) + H(h/n). \tag{8}$$

它表明,在有限训练样本下,学习机器的 VC维越高 (复杂性越高)则置信范围越大,导致真实风险与经验风险之间可能的差别越大.这就是为什么会出现过学习现象的原因.机器学习过程不但要使经验风险最小,还要使 VC维尽量小以缩小置信范围,才能取得较小的实际风险,即对未来样本有较好的推广性.

需要指出,推广性的界是对于最坏情况的结论,在很多情况下是较松的,尤其当 VC维较高时更是如此 (文献 [6]指出当 $h/n > 0.37$ 时这个界肯定是松弛的,当 VC维无穷大时这个界就不再成立[1]).而且,这种界只在对同一类学习函数进行比较时有效,可以指导我们从函数集中选择最优的函数,在不同函数集之间比较却不一定成立. Vapnik指出[2],寻找更好地反映学习机器能力的参数和得到更紧的界是学习理论今后的研究方向之一.

3.3 结构风险最小化

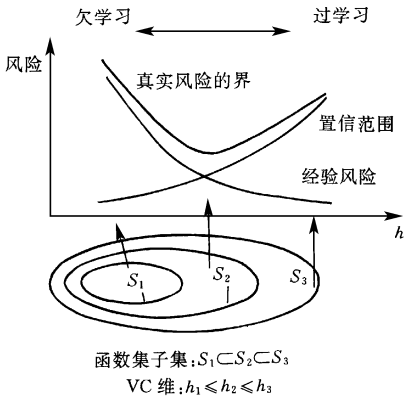

图 1 有序风险最小化示意图

从上面的结论看到, ERM原则在样本有限时是不合理的,我们需要同时最小化经验风险和置信范围.其实,在传统方法中,选择学习模型和算法的过程就是调整置信范围的过程,如果模型比较适合现有的训练样本 (相当于 h/n 值适当),则可以取得比较好的效果.但因为缺乏理论指导,这种选择只能依赖先验知识和经验,造成了如神经网络等方法对使用者"技巧"的过分依赖.

统计学习理论提出了一种新的策略,即把函数集构造为一个函数子集序列,使各个子集按照 VC维的大小 (亦即 H的大小)排列;在每个子集中寻找最小经验风险,在子集间折衷考虑经验风险和置信范围,取得实际风险的最小,如图 1所示.这种思想称作结构风险最小化 (Structural Risk Minimization或译有序风险最小化[4])即 SRM准则.统计学习理论还给出了合理的函数子集结构应满足的条件及在 SRM准则下实际风险收敛的性质[2].

实现 SRM原则可以有两种思路,一是在每个子集中求最小经验风险,然后选择使最小经验风险和置信范围之和最小的子集.显然这种方法比较费时,当子集数目很大甚至是无穷时不可行.因此有第二种思路,即设计函数集的某种结构使每个子集中都能取得最小的经验风险 (如使训练误差为 0),然后只需选择选择适当的子集使置信范围最小,则这个

1)比如最近邻法,可以很简单地证明它的 VC维为无穷大,但它却在很多情况下有较好的推广性,说明统计学习理论中关于推广性界的理论并不能解决机器学习中的所有问题,很多问题仍值得深入研究.

子集中使经验风险最小的函数就是最优函数.支持向量机方法实际上就是这种思想的具体实现.文献[3]中讨论了一些函数子集结构的例子[1)]和如何根据SRM准则对某些传统方法进行改进的问题.

4 支持向量机

支持向量机简称SVM,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[7,8,9,2],目前仍处在不断发展阶段.

4.1 广义最优分类面

SVM是从线性可分情况下的最优分类面发展而来的,基本思想可用图2的两维情况说明.图中,实心点和空心点代表两类样本,H为分类线,H_1, H_2分别为过各类中离分类线最近的样本且平行于分类线的直线,它们之间的距离叫做分类间隔(margin).所谓最优分类线就是要求分类线不但能将两类正确分开(训练错误率为0),而且使分类间隔最大.分类线方程为$\boldsymbol{x}\cdot\boldsymbol{w}+b=0$,我们可以对它进行归一化,使得对线性可分的样本集$(\boldsymbol{x}_i, y_i)$, $i=1,\cdots,n$, $\boldsymbol{x}\in R^d$, $y\in\{+1,-1\}$,满足

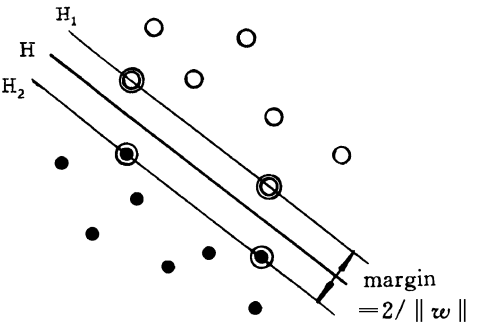

图2　线性可分情况下的最优分类线

$$y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]-1\geqslant 0,\quad i=1,\cdots,n. \tag{9}$$

此时分类间隔等于$2/\|\boldsymbol{w}\|$,使间隔最大等价于使$\|\boldsymbol{w}\|^2$最小.满足条件(9)且使$\frac{1}{2}\|\boldsymbol{w}\|^2$最小的分类面就叫做最优分类面, H_1, H_2上的训练样本点就称作支持向量.

使分类间隔最大实际上就是对推广能力的控制,这是SVM的核心思想之一.统计学习理论指出[2],在N维空间中,设样本分布在一个半径为R的超球范围内,则满足条件$\|\boldsymbol{w}\|\leqslant A$的正则超平面构成的指示函数集$f(\boldsymbol{x},\boldsymbol{w},b)=\mathrm{sgn}\{(\boldsymbol{w}\cdot\boldsymbol{x})+b\}$(sgn(　)为符号函数)的VC维满足下面的界

$$h\leqslant \min([R^2A^2],N)+1. \tag{10}$$

因此使$\|\boldsymbol{w}\|^2$最小就是使VC维的上界最小,从而实现SRM准则中对函数复杂性的选择.

利用Lagrange优化方法可以把上述最优分类面问题转化为其对偶问题[6],即在约束条件

$$\sum_{i=1}^{n} y_i\alpha_i=0 \tag{11a}$$

和

$$\alpha_i\geqslant 0,\quad i=1,\cdots,n \tag{11b}$$

下对α_i求解下列函数的最大值

1)应当指出,SRM准则并没有指出如何选择函数子集结构,而且由于尚没有一般地关于VC维计算的理论和方法,构造函数子集的一般方法也是一个需要进一步研究的问题.

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j),\tag{12}$$

α_i为与每个样本对应的 Lagrange乘子.这是一个不等式约束下二次函数寻优的问题,存在唯一解.容易证明,解中将只有一部分(通常是少部分)α_i不为零,对应的样本就是支持向量.解上述问题后得到的最优分类函数是

$$f(\boldsymbol{x})=\operatorname{sgn}\{(\boldsymbol{w}\cdot\boldsymbol{x})+b\}=\operatorname{sgn}\left\{\sum_{i=1}^{n}\alpha_i^{*}y_i(\boldsymbol{x}_i\cdot\boldsymbol{x})+b^{*}\right\},\tag{13}$$

式中的求和实际上只对支持向量进行.b^{*}是分类阈值,可以用任一个支持向量(满足(9)式中的等号)求得,或通过两类中任意一对支持向量取中值求得.

在线性不可分的情况下,可以在条件(9)中增加一个松弛项 $\xi_i\geqslant 0$,成为

$$y_i[(w\cdot x_i)+b]-1+\xi_i\geqslant 0,\quad i=1,\cdots,n,\tag{14}$$

将目标改为求$(w,\xi)=\frac{1}{2}\|w\|^2+C\left(\sum_{i=1}^{n}\xi_i\right)$最小,即折衷考虑最少错分样本和最大分类间隔,就得到广义最优分类面.其中,$C>0$是一个常数,它控制对错分样本惩罚的程度.广义最优分类面的对偶问题与线性可分情况下几乎完全相同,只是条件(11b)变为

$$0\leqslant\alpha_i\leqslant C,\quad i=1,\cdots,n.\tag{15}$$

4.2 支持向量机

对于 N 维空间中的线性函数,其 VC维为 $N+1$,但根据式(10)的结论,在$\|\boldsymbol{w}\|\leqslant A$的约束下其 VC维可能大大减小,即使在十分高维的空间中也可以得到较小 VC维的函数集(比如文[2]中介绍了在10^{3}维空间中取得 VC维在10^{3}左右的分类面的例子),以保证有较好的推广性.同时我们看到,通过把原问题转化为对偶问题,计算的复杂度不再取决于空间维数,而是取决于样本数,尤其是样本中的支持向量数.这些特点使有效地对付高维问题成为可能.

对非线性问题,可以通过非线性变换转化为某个高维空间中的线性问题,在变换空间求最优分类面.这种变换可能比较复杂,因此这种思路在一般情况下不易实现.但是注意到,在上面的对偶问题中,不论是寻优函数(12)还是分类函数(13)都只涉及训练样本之间的内积运算$(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)$,这样,在高维空间实际上只需进行内积运算,而这种内积运算是可以用原空间中的函数实现的,我们甚至没有必要知道变换的形式.根据泛函的有关理论,只要一种核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$满足 Mercer条件,它就对应某一变换空间中的内积[2].

因此,在最优分类面中采用适当的内积函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$就可以实现某一非线性变换后的线性分类,而计算复杂度却没有增加,此时目标函数(12)变为

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j),\tag{16}$$

而相应的分类函数也变为

$$f(\boldsymbol{x})=\operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i^{*}y_iK(x_i,x)+b^{*}\right),\tag{17}$$

这就是支持向量机.

概括地说,支持向量机就是首先通过用内积函数定义的非线性变换将输入空间变换到一个高维空间,在这个空间中求(广义)最优分类面. SVM分类函数形式上类似于一个神经网络,输出是中间节点的线性组合,每个中间节点对应一个支持向量,如图3所示.

4.3 核函数

SVM 中不同的内积核函数将形成不同的算法,目前研究最多的核函数主要有三类,一是多项式核函数

$$K(\boldsymbol{x},\boldsymbol{x}_i) = [(\boldsymbol{x}\cdot\boldsymbol{x}_i)+1]^q, \quad (18)$$

所得到的是 q 阶多项式分类器;二是径向基函数(RBF)

$$K(x,x_i) = \exp\left\{-\frac{|x-x_i|^2}{\sigma^2}\right\}, \quad (19)$$

所得分类器与传统 RBF方法的重要区别是,这里每个基函数中心对应一个支持向量,它们及输出权值都是由算法自动确定的.也可以采用 Sigmoid函数作为内积,即

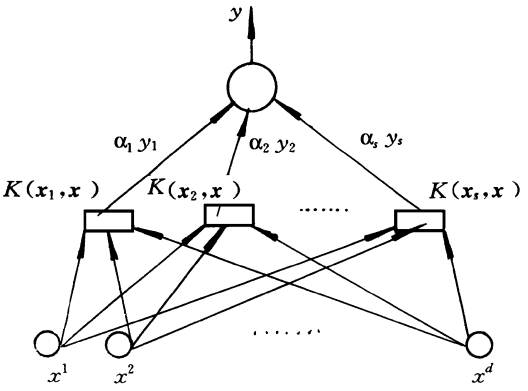

图 3　支持向量机示意图

$$K(x,x_i) = \tanh(v(x\cdot x_i)+c), \quad (20)$$

这时 SVM 实现的就是包含一个隐层的多层感知器,隐层节点数是由算法自动确定的,而且算法不存在困扰神经网络方法的局部极小点问题.

4.4 用于函数拟合的 SVM

SVM 方法也可以很好地应用于函数拟合问题中[10-12],其思路与在模式识别中十分相似.首先考虑用线性回归函数 $f(\boldsymbol{x})=\boldsymbol{w}\cdot\boldsymbol{x}+b$ 拟合数据 $\{\boldsymbol{x}_i,y_i\}$, $i=1,\cdots,n$, $\boldsymbol{x}\in R^d$, $y_i\in R$ 的问题,并假设所有训练数据都可以在精度 ε 下无误差地用线性函数拟合,即

$$\begin{cases} y_i-\boldsymbol{w}\cdot\boldsymbol{x}_i-b\leqslant\varepsilon, \\ \boldsymbol{w}\cdot\boldsymbol{x}_i+b-y_i\leqslant\varepsilon, \end{cases} \quad i=1,\cdots,n. \quad (21)$$

与最优分类面中最大化分类间隔相似,这里控制函数集复杂性的方法是使回归函数最平坦,它等价于最小化 $\frac{1}{2}\|\boldsymbol{w}\|^2$.考虑到允许拟合误差的情况,引入松弛因子 $\xi_i\geqslant 0$ 和 $\xi_i^*\geqslant 0$,则条件(21)变成

$$\begin{cases} y_i-\boldsymbol{w}\cdot\boldsymbol{x}_i-b\leqslant\varepsilon+\xi_i, \\ \boldsymbol{w}\cdot\boldsymbol{x}_i+b-y_i\leqslant\varepsilon+\xi_i^*, \end{cases} \quad i=1,\cdots,n. \quad (22)$$

优化目标变成最小化 $\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*)$,常数 $C>0$ 控制对超出误差 ε 的样本的惩罚程度.采用同样的优化方法可以得到其对偶问题.在条件

$$\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0, \quad 0\leqslant\alpha_i,\alpha_i^*\leqslant C, \quad i=1,\cdots,n \quad (23)$$

下,对 Lagrange因子 α_i,α_i^* 最大化目标函数

$$W(\alpha,\alpha^*)=-\varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{n}y_i(\alpha_i^*-\alpha_i)-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)(\boldsymbol{x}_i\cdot\boldsymbol{x}_j), \quad (24)$$

得回归函数为

$$f(x)=(w\cdot x)+b=\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)(x_i\cdot x)+b^*. \quad (25)$$

与模式识别中的 SVM 方法一样,这里 α_i, α_i^* 也将只有小部分不为 0,它们对应的样本就是支持向量,一般是在函数变化比较剧烈的位置上的样本;而且这里也是只涉及内积运算,只要用核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$替代(24),(25)中的内积运算就可以实现非线性函数拟合.

4.5 核函数主成分分析

SVM 方法中一个重要启示是用内积运算实现某种非线性变换,这种思想也可以在其它问题中得到的应用,比较成功的例子就是用核函数实现非线性主成分分析[13, 14],它是传统主成分分析(PCA)方法的推广.

对于样本集 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$,主成分方向是矩阵 $C = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}$ 的特征向量.对 $\boldsymbol{x}$ 进行非线性变换 $\Phi(\boldsymbol{x})$,可得 $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} \Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_i)^{\mathrm{T}}$,其特征向量 v 就是原样本集的非线性主成分方向,满足 $\lambda v = \bar{C}v$.将每个样本与该式内积,得

$$\lambda(\Phi(\boldsymbol{x}_k) \cdot v) = \Phi(\boldsymbol{x}_k) \cdot \bar{C}v, \quad k = 1, \cdots, n. \tag{26}$$

可以证明,特征向量 v 可以写成 $v = \sum_{i=1}^{n} \alpha_i \Phi(\boldsymbol{x}_i)$,将它代入(26)式中,并定义矩阵

$$K = \{K_{ij}\} = \{(\Phi(x_i) \cdot \Phi(x_j))\} = \{K(x_i, x_j)\}, \tag{27}$$

(K_{ij} 为矩阵的第 i 行第 j 列个元素),可以得到

$$n\lambda a = Ka, \tag{28}$$

其中 $a = [\alpha_1, \cdots, \alpha_n]^{\mathrm{T}}$.从矩阵 K 的特征向量 a 即可求出 $\bar{C}$ 的特征向量 v,即 $\Phi(\boldsymbol{x})$空间的主成分方向.对于原空间中的任意向量 $\boldsymbol{x}$,它在变换空间中的主成分是 $\Phi(\boldsymbol{x})$在主成分方向 v 上的投影,即

$$v \cdot \Phi(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}). \tag{29}$$

显然,与 SVM 算法中类似,这里得到的非线性主成分方法只需在原空间中计算用作内积的核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$,而无需真正计算对应的非线性变换,因此称作核函数主成分分析.

4.6 应用研究

比较遗憾的是,虽然 SVM 方法在理论上具有很突出的优势,但与其理论研究相比,应用研究尚相对比较滞后,目前只有较有限的实验研究报道,且多属仿真和对比实验[15]. SVM 的应用应该是一个大有作为的方向.

在模式识别方面最突出的应用研究是贝尔实验室对美国邮政手写数字库进行的实验[2, 8],这是一个可识别性较差的数据库,人工识别平均错误率是 2.5%,用决策树方法识别错误率是 16.2%,两层神经网络中错误率最小的是 5.9%,专门针对该特定问题设计的五层神经网络错误率为 5.1%(其中利用了大量先验知识),而用三种 SVM 方法(采用式(18)~(20)的核函数)得到的错误率分别为 4.0%、4.1% 和 4.2%,且其中直接采用了 16×16的字符点阵作为 SVM 的输入,并没有进行专门的特征提取.实验一方面说明了 SVM 方法较传统方法有明显的优势,同时也得到了不同的 SVM 方法可以得到性能相近的结果(不像神经网络那样依赖于模型的选择).实验还观察到,三种 SVM 求出的支持向量中有 80% 以上是重合的,它们都只是总样本中很少的一部分,说明支持向量本身对不同方法具有一定的不敏感性(遗憾的是这些结论仅仅是有限的实验中观察到的现象,如果能得到证明,将会使 SVM 的理论和应用有更大的突破).围绕这一字符识别实验,还提出了一些

对 SVM的改进,比如引入关于不变性的知识[16,17]、识别和去除样本集中的野值[18]、通过样本集预处理提高识别速度[19]等,相关的应用还包括 SVM与神经网络相结合对笔迹进行在线适应[20].除此之外, MIT用 SVM进行的人脸检测实验也取得了较好的效果,可以较好地学会在图像中找出可能的人脸位置[21~23].其它有报道的实验领域还包括文本识别[23,24]、人脸识别[25]、三维物体识别[26]、遥感图像分析[27]等.

在函数拟合方面,主要实验尚属于原理性研究,包括函数逼近、时间序列预测及数据压缩[10~12,28,29]等.

其它有关研究还有对 SVM中优化算法实现的研究[30,31,23]、改进的 SVM方法[32,33]甚至硬件实现研究[34]等.

5 讨论

由于统计学习理论和支持向量机建立了一套较好的有限样本下机器学习的理论框架和通用方法,既有严格的理论基础,又能较好地解决小样本、非线性、高维数和局部极小点等实际问题,因此成为九十年代末发展最快的研究方向之一,其核心思想就是学习机器要与有限的训练样本相适应.本文对它们的基本思想、方法及研究方向进行了介绍,希望使读者对这一领域有一个基本的了解.统计学习理论虽然已经提出多年,但从它自身趋向成熟和被广泛重视到现在毕竟才只有几年的时间,其中还有很多尚未解决或尚未充分解决的问题,在应用方面的研究更是刚刚开始.我们认为,这是一个十分值得大力研究的领域.

参 考 文 献

1 Vapnik V N. Estimation of Dependencies Based on Empirical Data. Berlin: Springer-Verlag, 1982

2 Vapnik V N. The Nature of Statistical Learning Theory, NY: Springer-Verlag, 1995
张学工译. 统计学习理论的本质.北京: 清华大学出版社, 1999(待出版)

3 Cherkassky V, Mulier F. Learning from Data: Concepts, Theory and Methods. NY: John Viley & Sons, 1997

4 边肇祺等.模式识别.北京: 清华大学出版社, 1988

5 Vapnik V, Levin E, Le Cun Y. Measuring the VC-dimension of a learning machine. *Neural Computation*, 1994, **6**: 851~876.

6 Burges C J C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 1998 **2**(2)

7 Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers, Fifth Annual Workshop on Computational Learning Theory. Pittsburgh: ACM Press, 1992

8 Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, **20**: 273~297

9 Scholkopf B, Burges C, Vapnik V. Extracting support data for a given task. In: Fayyad U M, Uthurusamy R (eds.). Proc. of First Intl. Conf. on Knowledge Discovery & Data Mining, AAAI Press, 1995, 262~267

10 Vapnik V, Golowich S, Smola A, Support vector method for function approximation, regression estimation, and signal processing. In: Mozer M, Jordan M, Petsche T (eds). Neural Information Processing Systems, MIT Press, 1997, 9

11 Müller K-R, Smola A J, Ratsch G *et al*. Predicting time series with support vector machines. In: Proc. of ICANN 97, Springer Lecture Notes in Computer Science, 1997, 999~1005

12 Drucker H, Burges C, Kaufman L *et al*. Support vector regression machines. In: Mozer M, Jordan M, Petsche T (eds). Neural Information Processing Systems, MIT Press, 1997, 9

13 Scholkopf B, Smola A, Müller K-R. Kernel principal component analysis. In Proc. of ICANN'97, 1997, 583~589

14 Scholkopf B, Smola A, Müller K-R. Nonlinear component analysis as kernel eigenvalue problem. *Neural Computation*, 1998, **10**(5): 1299~1319

15 Scholkopf B, Sung K-K, Burges C *et al*. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. on Signal Processing*, 1997, **45**(11): 2758~2765

16 Scholkopf B, Burges C, Vapnik V. Incorporating invariances in support vector learning machines, In von der Malsburg C, von Seelen W, Vorbrüggen J C *et al*(eds). Artificial Neural Networks-ICANN'96, Spingers Lecture Notes in Computer Science, Berlin, 1996, **1112**: 47~52

17 Scholkopf B, Simard P, Smola A *et al*. Prior knowledge in support vector kernels. NIPS'97, 1997

18 Guyon I, Matic N, Vapnik V. Discovering informative patterns and data cleaning. In Fayyad U M, Piatetsky-Shapiro G, Smyth P *et al*(eds). Advances in Knowledge Discovery & Data Mining, MIT Press, 1996, 181~203

19 Burges C, Scholkopf B. Improving the accuracy and speed of support vector machines. In Mozer M, Jordan M, Petsche T(eds). Neural Information Processing Systems, MIT Press, 1997, 9

20 Matic N, Guyon I, Denker J *et al*. Writer adaptation for on-line handwritten character recognition. In 2nd Intl. Conf. on Pattern Recognition and Document Analysis, 1993, 187~191

21 Oren M, Papageorgiou C, Sinha P *et al*. Pedestrian detection using wavelet templates, In Proc. of CVPR'97, Puerto Rico, 1997

22 Hearst M A, Scholkopf B, Dumais S *et al*. Trends and controversies-support vector machines, *IEEE Intelligent Systems*, 1998, **13**(4): 18~28

23 Osuna E, Freund R, Girosi F. Training support vector machines: an application to face detection. In Proc. of CVPR'97, Puerto Rico, 1997

24 卢增祥,李衍达. 交互 SVM学习算法及其在文本信息过滤中的应用. 清华大学学报, 1999(待发表)

25 Lu Chunyu, Yan Pingfan, Zhang Changshui, Zhou Jie. Face recognition using support vector machine. In Proc. of ICNNB'98, Beijing, 1998, 652~655

26 Blanz V, Scholkopf B, Bülthoff H. *et al*. Comparison of view-based object recognition algorithms using realistic 3D models, In: von der Malsburg C, von Seelen W, Vorbrüggen J C *et al*(eds). Artificial Neural Networks——ICANN'96, Spingers Lecture Notes in Computer Science, Berlin, 1996, **1112**: 251~256

27 Brown M, Lewis H G, Gunn S R. Linear spectral mixture models and support vector machines for remote sensing, (submitted to)*IEEE Trans. Geoscience and Remote Sensing*, 1998

28 Mukherjee S, Osuna E, Girosi F. Nonlinear prediction of chaotic time series using a support vector machine. In Proc. of NNSP'97, 1997

29 Kwok J T-Y, Support vector mixture for classification and regression problems. ICPR'98, 1998

30 Bennett K, Mangasarian O. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1992, **1**: 23~34

31 Osuna E, Freund R, Girosi F. An improved training algorithm for support vector machines. In Proc. of NNSP'97, 1997

32 Bennett K P, Demiriz A. Semi-supervised support vector machines. In Proc. of NIPS'98, 1998

33 Zhang Xuegong. Using class-center vectors to build support vector machines. In Proc. of NNSP'99, 1999: 3~11

34 Anguita D, Ridella S, Rovetta S. Circuital implementation of support vector machines. *Electronics Letters*, 1998, **34**(16): 1596~1597

张学工 1965年生,1994年于清华大学获博士学位,现为清华大学自动化系副研究员,主要研究方向为模式识别的理论与方法、智能信息处理与融合技术及其在地球物理勘探等领域中的应用.