支持向量机

Support Vector Machine

张浩然　韩正之　李昌刚

(上海交通大学自动化系　上海200030)

Abstract This paper gives a introduction of the basic ideas, basic theory, key techniques, and application of the support vector machine (SVM), and indicates the similarities and differences between support vector machines and neural networks.

Keywords Support vector machine, Neural networks, Statistical learning theory, Machine learning

1 前言

基于数据的机器学习是人工智能技术中的重要方面，从观测数据(样本)出发寻找数据中的模式和数据间的函数依赖规律，利用这些模式和函数依赖对未来数据或无法观测的数据进行分类、识别和预测。关于其实现方法大致可以分为三种，第一种是经典的(参数)统计估计方法，在这种方法中，参数的相关形式是已知的，训练样本用来估计参数的值。这种方法有很大的局限性，首先，它需要已知样本分布形式，其次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。第二种方法是人工神经网络(ANN)。这种方法利用已知样本建立非线性模型，克服了传统参数估计方法的困难，在过去的十几年中，神经网络受各个领域学者的广泛研究，技术上得到很大的发展，提出了许多神经网络结构，其中常用的有如多层感知器(MLP)、径向基函数网络(RBF)、Hopfield 网络等等[1]，也被成功地用来解决许多实际问题，例如模式识别、信号处理、智能控制等等。但是现在的神经网络技术研究理论基石不足，有较大的经验成分，在技术上仍存在一些不易解决的问题，例如网络结构的设计问题，学习算法中局部极小问题，学习的快速性问题等等。为了克服这些难题，Vapnik 提出了一种新的神经网络－支持向量机(SVM)，它也是所说的第三种方法－统计学习理论，SVM 是统计学习理论中最年轻的内容，也是最实用的部分[2]，它目前已经成为神经网络和机器学习的研究热点之一，并已经得到很好的研究成果[3]。越来越多的学者认为，关于支持向量机的研究，将很快出现象在80年代后期人工神经网络研究那样的飞速发展阶段[3]。

2 支持向量机的基本思想

支持向量机的基本思想是这样的：首先把训练数据集非线性地映射到一个高维特征空间(这个高维特征空间是 Hilbert 空间)，这个非线性映射的目的是把在输入空间中的线性不可分数据集映射到高维特征空间后变为是线性可分的数据集，随后在特征空间建立一个具有最大隔离距离的最优分离超平面，这也相当于在输入空间产生一个最优非线性决策边界，如图1所示。

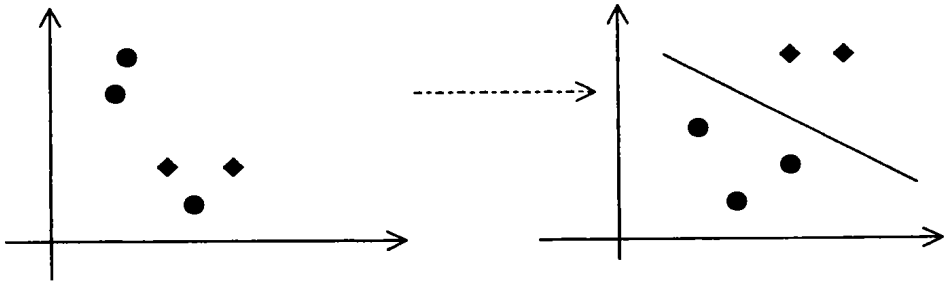

图1　支持向量机的基本思想

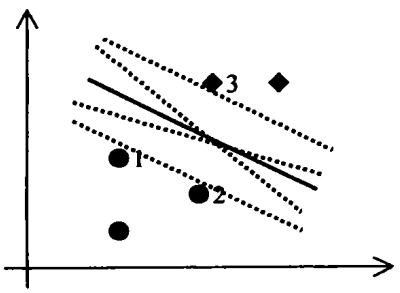

图2　最优分离超平面和非最优分离超平面

这里应该注意的是，在特征空间中支持向量机的分离超平面是最优的分离超平面，最优性可以从图2看出，几个分离超平面都可以把两个类分离开，但是只有一个是最优的，就是图中的实线所表示的，它与两个类之间最近向量的距离最大。从几何上说支持向量就是决定最优分离超平面的样本向量的最小个数，如上图中的样本1、2、3，它们就是所说的支持向量，所以这种学习机叫支持向量机。但实际上 SVM 最吸引人的地方不是支持向量思想，而是结构风险最小化思想，也就是上述的最优分离超平面不但能使学习机的经验风险很小，同时泛化误差也很小，即结构风险最小。

3 支持向量机的关键技术

要实现上述的思想，根据实际问题构造一个支持向量机，

必须解决两个关键的技术问题：(1)如何找到一个非线性映射，把输入空间中的线性不可分数据集映射到高维特征空间中的线性可分数据集；(2)在高维特征空间如何求出最优分离超平面。

对于第一个问题，使用核技术和方法来解决，已经证明，选用满足一定条件的核函数，可以把在输入空间中线性不可分问题映射到一个特征空间的线性可分问题。在1992年Boser和Vapnik发现为了在特征空间Z构造最优分类超平面，并不需要以显示形式来考虑特征空间，而只需要能够计算支持向量与特征空间中向量的内积。考虑在Hilbert空间中内积的一个一般表达：

$$(z_i \cdot z)=K(x,x_i)$$

其中z是输入空间中的向量在特征空间中的像。$K(x,x_i)$是满足Mercer条件的任意对称函数。常用的核有以下几种：

(1)线性核：$K(x,x_i)=x_i \cdot x$

(2)径向核：$K(x,x_i)=\exp(-\|x-x_i\|^2/2\sigma^2)$

(3)多项式核：$K(x,x_i)=(x_i \cdot x+1)^d, d=1,2,\cdots,N$

(4)感知器核：$K(x,x_i)=\tanh(\beta x_i \cdot x+b)$

它们分别组成线性SVM、多项式SVM、径向基函数SVM和感知器SVM。有很多的经验应用表明，径向基函数(RBF)SVM具有良好的学习能力。

对于第二个问题，即确定最优分离超平面，可以证明这是一个典型的受约束二次型规划问题，在SVM里，这个优化问题的目标函数不是经验风险。

用作分类(二分类)时SVM的输出：

$$f(x,W)=\mathrm{sgn}(\sum_{i=1}^{N} w_i K(x,x_i)+b)$$

优化目标函数：

$$J=W^TW=\|W\|^2$$

约束条件是：

$$y_j[\sum_{i=1}^{N} w_i K(x_j,x_i)+b] \geqslant 1, j=1,\cdots,N$$

其中N是样本数，W是支持向量机的输出可调参数向量，(x_i,y_i)是样本。目标函数J是为了保证分类的最优性，约束条件是为了保证分类的正确性。为了消除噪音和异常样本的影响，引入松弛变量，如下：

$$J=\frac{1}{2}W^TW+C\sum_{i=1}^{N}\zeta_i$$

$$y_j[\sum_{i=1}^{N} w_i K(x_j,x_i)+b] \geqslant 1-\zeta_j,\ j=1,\cdots,N$$

$$\zeta_j \geqslant 0, j=1,\cdots,N$$

当用作函数回归时SVM的输出：

$$f(x,W)=\sum_{i=1}^{N} w_i K(x,x_i)+b$$

引入一个Vapnik不灵敏函数，ε取值大小影响支持向量的数目：

$$|y-f(x)|_\varepsilon=\begin{cases}0, & |y-f(x)| \leqslant \varepsilon \\ |y-f(x)|-\varepsilon, & \text{其他}\end{cases}$$

优化的目标函数为：

$$J=\frac{1}{2}W^TW+C\sum_{i=1}^{N}(\zeta_i+\zeta_i^*)$$

约束条件为：

$$y_j-\sum_{i=1}^{N} w_i K(x_j,x_i)-b \leqslant \varepsilon+\zeta_j,\ j=1,\cdots,N$$

$$\sum_{i=1}^{N} w_i K(x_j,x_i)+b-y_j \leqslant \varepsilon+\zeta_j^*,\ j=1,\cdots,N$$

$$\zeta,\zeta_j^* \geqslant 0,\ j=1,\cdots,N$$

上述问题可以用Lagrangian乘子法转化为无约束优化问题，然后再用常用的优化方法如最小二乘法、最速下降法、牛顿法、共轭梯度法、拟牛顿法等等常用的数值优化方法来求解这个问题。这个优化问题凸优化问题，并不存在一个局部极小点，而且也考虑到噪音的影响，使之更加具有鲁棒性。从计算复杂度上说这个优化问题的复杂性只取决于训练样本集的大小，而与样本的维数无关，因此当样本集很大时上述二次型规划问题的计算复杂性很大。为了解决这个问题，1995年Vapnik提出了组块(chunking)方法，1997年Osuna提出了另一个分解方法，同年Platt提出了序列最小化方法，这种方法被认为是SVM学习的“误差后置法”。上述的优化过程实质上是支持向量的选择过程，支持向量对应于二次型规划中的解向量中的非零项。也有很多学者利用线性规划(LP)来选择支持向量，这样也可以克服大样本时所带来的计算复杂性问题。Smola在1998年，Bennet在1999年，Weston et在1999年，Graepel et在1999年都用线性规划(LP)来选择支持向量，用LP来选择支持向量是支持向量机技术上的一个很重要的发展。下面以回归问题为例来说明这种方法的原理。假设已给定样本集(x_i,y_i)和核函数$K(x,x_i)$(i=1,…,N)，支持向量机的输出为：

$$f(x,W)=\sum_{i=1}^{N} w_i K(x,x_i)+b$$

求极小值：

$$J=\|W\|_1$$

约束条件是：

$$y_j-\sum_{i=1}^{N} w_i K(x_j,x_i)-b \leqslant \varepsilon,\ j=1,\cdots,N$$

$$\sum_{i=1}^{N} w_i K(x_j,x_i)+b-y_j \leqslant \varepsilon,\ j=1,\cdots,N$$

为了简化推导这里没有引入松弛变量，ε是SVM的不灵敏区，它定义了最大允许误差。这里问题的提法和前面的不一样，目标函数在L_1空间取范数，L_1空间的范数值为矢量内元素的绝对值和。上述的约束优化问题可以转化为一个标准的线性规划(LP)问题。

首先目标函数$\|W\|_1=\sum_{i=1}^{N}|w_i|$并不是一个标准的线性规划形式$C^TW$，可以利用如下变换：

$$w_i=w_i^+-w_i^-$$

$$|w_i|=w_i^++w_i^-$$

其中w_i^+, w_i^-是两个非负数。这样目标函数就可以变成线性规划的标准形式了。

其次约束条件也不是线性规划的标准形式，也必须进行转换。上述的约束条件可以重写为下式：

$$y_j-b-\varepsilon \leqslant \sum_{i=1}^{N}(w_i^+-w_i^-)K(x_j,x_i) \leqslant y_j-b+\varepsilon,\ j=1,\cdots,N$$

$$w_j^+, w_j^- \geqslant 0, j=1,\cdots,N$$

这样上述的优化问题就可以转化为标准的线性规划问题，写成矩阵的形式如下：

目标函数：

$$J=C^TW=[11111\cdots1]\times\begin{bmatrix}w_1^+\\ \cdot\\ \cdot\\ w_N^+\\ w_1^-\\ \cdot\\ \cdot\\ w_N^-\end{bmatrix}$$

约束条件：

$$\begin{bmatrix}K & -K\\ -K & K\end{bmatrix}\begin{bmatrix}W^+\\ W^-\end{bmatrix}\leqslant\begin{bmatrix}Y-B+\varepsilon1\\ -Y+B+\varepsilon1\end{bmatrix}$$

$$W^+,W^-\geqslant0$$

其中：

$$K=\begin{bmatrix}\sum_{i=1}^{N}K(x_1,x_1) & \cdot & \cdot & \sum_{i=1}^{N}K(x_N,x_1)\\ \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot\\ \sum_{i=1}^{N}K(x_1,x_N) & \cdot & \cdot & \sum_{i=1}^{N}K(x_N,x_N)\end{bmatrix}$$

$Y=[y_1,\cdots,y_N]^T$，$B=[b,\cdots,b]^T$，$\varepsilon1=[\varepsilon,\cdots,\varepsilon]^T$，$W^+=[w_1^+,\cdots,w_N^+]^T$，$W^-=[w_1^-,\cdots,w_N^-]^T$，N 是样本集数。

尽管现在还没有从理论上分析用二次型规划(QP)和线性规划(LP)训练 SVM 的优劣，但有很多的仿真研究已经表明当训练样本集很大时，LP 比 QP 有更好的结果[4]，且 LP 比 QP 速度更快，鲁棒性更强。

4 支持向量机与神经网络的异同

相同点：神经网络和支持向量机都是从经验数据集中学习的，是数据驱动(data driven)的学习机。支持向量机和单隐层神经网络在结构上并没有什么区别，都可以表示为网络结构，数学表达也相似。在功能上它们都可以作为通用逼近器以任意精度逼近任何的函数，它们的区别是在学习方面。

不同点：神经网络与支持向量机最大的不同在于学习上，神经网络学习的目标函数是经验风险最小化，而支持向量机的学习的目标函数是结构风险最小化。神经网络学习所得到的是一个小的训练误差，而支持向量机学习所得到的是一个小的泛化误差。观察如下的三个学习目标：

$$E=\sum_{i=1}^{N}(y_i-f(x_i,W))^2=R_{emp}(W)\qquad(1)$$

$$E=\sum_{i=1}^{N}(y_i-f(x_i,W))^2+\lambda\|Pf\|=R_{emp}+\lambda\|Pf\|\qquad(2)$$

$$E=\sum_{i=1}^{N}(y_i-f(x_i,W))^2+\Phi(h/n)=R_{emp}+\Phi(h/n)\qquad(3)$$

其中 h 是学习机的 VC 维，n 是训练样本数。(1)式的 E 是标准神经网络的学习目标，也就是经验风险 $R_{emp}(w)$，学习时只追求经验风险 $R_{emp}(w)$极小，但是训练误差小并不总能导致好的预测、推广和泛化效果，某些情况下，训练误差过小反而导致泛化能力的下降，这就是过拟合或过学习问题。(2)式的 E 是正规化神经网络的学习目标，是经验风险与调整因子的和，第二项的作用是调整拟合函数的平滑度，学习时不但追求经验风险小，而且也要求拟合函数有一定的平滑度，这在一定程度上经验式地提高学习系统的泛化能力。(3)式的 E 是支持向量机的学习目标，它是经验风险和置信范围之和，也叫结构风险。它在学习时追求的是结构风险的最小化，已经有学者证明[2]结构风险是实际风险的上界，即：

$$R(W)\leqslant R_{emp}(W)+\Phi(h/n)$$

其中 $R(W)$是学习机的实际风险。这说明支持向量机的泛化误差是能控制的，这样就能有效防止过拟合现象，因此比神经网络有更好的泛化能力。从设计方面讲，必须先验地设计神经网络的结构，而支持向量机的结构设计完全自动化了，支持向量的选取过程就是上述的优化过程。另外，如果是从稀疏数据集中、从含有噪音的数据集中、从高维数据集中学习，SVM 比神经网络具有更好的运算速度和结果精度[5]。最后神经网络从某种意义上说是一种启发式的学习机，本身有很大经验的成分，而支持向量机却具有严格的理论基础和数学基础。

5 支持向量机的应用

目前支持向量机技术主要用来解决模式识别问题和函数回归问题，在实际中有很多的具体应用，例如基本粒子辨识[6]，脸型检测[7]，字符分类[8]，非线性系统建模[9]，时间序列预测[10]，疾病诊断[11]。其中最有代表性的一个应用实例是美国的邮政部门用不同的方法设计分类器来进行手写数字字符识别[2]，结果发现：人工方法的错误率是2.5%，决策树的错误率是16.2%，两层神经网络的错误率是5.9%，五层神经网络的错误率是5.1%，多项式 SVM 的错误率是4.0%，RBFSVM 的错误率是4.1%，感知器 SVM 的错误率是4.2%，可以明显地看出 SVM 的性能要好于其他的分类器技术。比较起神经网络在工程应用中的深度和广度，支持向量机的应用才刚刚起步，很多的研究还处于实验室阶段，但它的应用前景是很诱人的。

结论 支持向量机是基于统计学习理论的新一代学习机器，具有很多吸引人的特点，它的理论基础是扎实、严谨、清晰、明确的；它的设计技术是系统可行而又简单实用的；它在函数表达能力、推广能力和学习效率上都要优于传统的人工神经网络[2]；它不但可以处理数值信息，还可以处理符号信息；在实际应用中也解决了许多实际问题，且效果良好。但从另一方面说，SVM 研究刚刚开始，它还处于发展阶段，有些方面还不成熟，还存在一些需要解决的问题，比如：许多理论目前还只有理论上的意义，尚不能在实际算法中实现；而有关 SVM 算法某些理论解释也并非完美，J. C. Burges. 在文[12]就曾提到结构风险最小化原理并不能严格证明 SVM 为什么有好的推广能力；此外，对于一个实际的学习机器的 VC 维的分析尚没有通用的方法；SVM 方法中如何根据具体问题选择适当的内积函数(核函数)也没有理论依据。但支持向量机是一个快速发展的领域，所取得的成果是激动人心的，很多学者指出它是机器学习研究的一个非常具有前途的一个发展方向。目前，国际上对这一理论的讨论和进一步研究逐渐广泛，而我国国内尚未在此领域开展系统的研究，从事支持向量机研究的学者也不多，发表的文献也很少，因此我们需要及时学习掌握有关理论，开展有效的研究工作，使我们在这一有着重要意义的领域中能够尽快赶上国际先进水平。

参 考 文 献

1 何振亚. 神经智能—认知科学中若干重大问题的研究. 长沙：湖南科学技术出版社，1997

2 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)

(下转第142页)

对事务数据进行数据品质处理。

假设有一个事务数据库D由数据库进行数据挖掘我们可得以下关联规则见表1。

表1

关联规则	置信度
肝胆湿热⇒脉弦数	80%
肝胆湿热⇒小便短赤	60%
肝胆湿热⇒胁肋胀痛	60%
肝胆湿热⇒大便溏结	50%
⋮	⋮

现在我们要对再进入系统的基础数据进行数据品质评价。

定义一个概念为得分用S表示$S\in R_t^+$，每个事务的得分S由前面数据挖掘得到的关联规则来计算。设R为关联规则的集合。对一个关联规则r=X⇒Y，其中X为前提记为body(r)=X，Y结论记为head(r)=Y。

定义事务对规则的违反度，对一个事务T有四种情况：

1)body(r)⊂T∧head(r)⊂T；

2)body(r)⊂T∧head(r)⊄T；

3)body(r)⊄T∧head(r)⊂T；

4)body(r)⊄T∧head(r)⊄T；

对第3和第4种情况前提都不满足，规则也就没有意义不存在违反或不违反的情况。第1种情况是满足规则的。第2种情况就只满足前提是违反规则的情况。

这样我们定义事务违反规则的公式为：

$$V(T,r)=\begin{cases}1 & \text{如 body}(r)\subset T\ \wedge \text{head}(r)\\ & \quad\subset T,\text{即第2种情况；}\\ 0 & \text{其它三种情况；}\end{cases}$$

事务的得分是本事务违反规则集合R的和，记为：

$$S(T,R)=\sum_{r\in R}C(r)^t\times V(T,r)$$

其中：C(r)是规则r的置信度，t是一个指数参数。

当t=7时，事务号1的得分：$S(1,R)=0.8^7+0.6^7+0.6^7+0.5^7$，其它同理如表2。

从表2可以看出t的不同值，得分的分布不同。t值越小，得分分布越疏。同时，t的不同值可以反映不同的情况。如事务3不违反的是置信度比较大为0.9的脉弦数的项，违反其它三项置信度比较小的分别为0.6、0.6和0.5。而事务2则反之。从表2我们可以看出在不同的t值下它们的得分是不同的，并且变化趋势也不同。这样在实际应用中进行调整就能从得分上体现使用者想要知道和关心的方面。再通过设置得分S的阈值就可以衡量事务数据的品质。

表2

事务号	事务内容	t=7时，得分	t=5时，得分
1	肝胆湿热⇒脉沉数　小便不利　腰痛　大便燥结	0.33	0.7
2	肝胆湿热⇒脉沉数　小便短赤　胁肋胀痛　大便溏结	0.21	0.33
3	肝胆湿热⇒脉弦数　小便不利　腰痛　大便燥结	0.13	0.37
4	肝胆湿热⇒脉弦数　小便短赤　胁肋胀痛　大便溏结	0	0
5	膀胱湿热⇒脉数　小便短赤　腰痛　舌红苔黄	0	0

我们举的例子只是清楚地表示应用数据挖掘技术进行数据品质的挖掘的过程。它的应用领域将十分广泛。

总结　现在，由于基础数据的品质问题和其它一些因素，数据挖掘的结果数据十分庞大，而实际可使用的信息的量不大，并且其使用效率也不高。在数据挖掘的整个过程都发展相关的数据品质技术将可以有效地解决这些相关的问题，数据挖掘在数据品质在数据挖掘的应用工具将是一个决定性因素。数据品质及其技术将使数据挖掘得到进一步的发展。

参考文献

1 Han Jiawei, Kambr M. 数据挖掘——概念与技术(影印版). 高等教育出版社, 2001

2 Srikant R, Agrawal R. Mining generalized association rules. In: Proc. of the 21st Conf. on Very Large Databases (VLDB'95), Switzerland, 1995

3 Jeusfeld M, Quix C, Jarke M. Design and Analysis of Quality Information for Data Warehouses. In: Proc. of Conf. on Information Quality Cambridge, Mass., 1997. 98～112

4 Theodoratos D, Bouzeghoub M. Data Currency Quality Factors in Data Warehouse Design(DMDW'99), 1999

5 张华珠主编. 中医学. 中国科技出版社, 1992

(上接第137页)

3 张学工译. 统计学习理论. 北京：清华大学出版社, 2000. 9

4 Kecman V. Learning and soft computing, The MIT Press, Cambridge, MA, 2001

5 de Freitas N, Milo M, Clarkson P. Sequential support vector machines, Neural Networks for Signal Processing IX. In: Proc. of the 1999 IEEE Signal Processing Society Workshop, 1999. 31～40

6 Barabino N, Pallavicini M. Support vector machines vs multi-layer perceptrons in particle identification. In: Proc. of the European Symposium on Artificial Neural Networks'99, 1999. 257～262

7 Osuna E, Freund R. Training Support Vector Machine : an Application to Face Dection. In: Proc. of CVPR'97, Puerto Rico, 1997

8 Drucker H, Wu D, Vapnik V. Support vector machine for spam categorization. IEEE Trans. on Neural Networks, 1999, 10: 1048～1054

9 Suykens J A K. Nonlinear modeling and support vector machines. IEEE instrumentation and measurement technology conference, Budapest, Hungary, 2001. 21～23

10 Mukherjee S, Osuna E, Girosi F. Nonlinear prediction of chaotic time seris using support vector machines. In: Proceedings of IEEE NNSP'97, 1997. 24～26

11 Mangasarian O L, Street W N. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 1995, 43(4): 570～577

12 Burges J C. A Tutorial on Support Vector Machines for Pattern Recognition. Bell Laboratories, Lucent Technologies. 1997