王建国 张文兴 等 著

支持向量机建模及其智能优化

清华大学出版社

王建国 张文兴 等 著

支持向量机建模及其智能优化

清华大学出版社
北京

内 容 简 介

本书以实际工业过程为背景，研究了冶金生产过程、化工生产过程，并获取了大量的实际生产数据，利用数据挖掘方法，挖掘出数据中隐含的生产规律，提出了一些改进的建模方法和优化方法，这些方法不仅适用于冶金和化工，还可广泛用于装备制造、材料、航天航空等领域。

本书的建模方法主要以支持向量机为基础，针对数据分布不平衡和海量数据的建模问题、模型的在线学习和优化问题进行了相关的算法研究，如粒度支持向量机、主动学习的增量支持向量机、误差校正的混合核函数在线支持向量机、粒子群智能优化方法和蚁群智能优化方法等，这些方法均配有仿真实验和实际生产数据实验，用于验证方法的有效性。

本书作为专业学术类参考书，可供高年级本科生、研究生、工程师、高校教师等人员阅读参考。

图书在版编目(CIP)数据

支持向量机建模及其智能优化／王建国，张文兴等著. --北京：清华大学出版社，2015
ISBN 978-7-302-40088-2

Ⅰ. ①支… Ⅱ. ①王… ②张… Ⅲ. ①向量计算机－系统建模－研究 Ⅳ. ①TP338

中国版本图书馆 CIP 数据核字(2015)第 089642 号

责任编辑：庄红权　洪　英
封面设计：常雪影
责任校对：王淑云
责任印制：沈　露

出版发行：清华大学出版社
网　址：http://www.tup.com.cn，http://www.wqbook.com
地　址：北京清华大学学研大厦 A 座　　邮　编：100084
社总机：010-62770175　　邮　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：三河市君旺印务有限公司
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　印　张：11.5　　字　数：213 千字
版　　次：2015 年 9 月第 1 版　　印　次：2015 年 9 月第1 次印刷
印　　数：1～1000
定　　价：39.00 元

产品编号：062725-01

前　言

FOREWORD

20世纪90年代，Vapnik等人在VC维理论和结构风险最小化原则的基础上，提出了针对小样本、非线性、高维问题的机器学习方法——支持向量机(support vector machine，SVM)。历经十多年的发展，支持向量机已成为国内外专家、学者的研究热点，并在模式识别、光谱分析、数据挖掘、人工智能、医学、经济、社会等各个领域得到广泛应用。

多年来，作者一直从事基于数据驱动的质量建模和群智能优化方面的研究，特别是在支持向量机建模及粒子群和蚁群智能优化方法方面取得了一些有价值的研究成果。本书主要以支持向量机理论为基础，针对数据分布不平衡和海量数据的SVM建模问题、模型的在线学习和优化问题以及相关的算法研究，同时利用公开标准数据集和实际生产过程产生的实际数据进行了方法验证。

本书可供机械、信息、冶金、化工等领域的高年级本科生、研究生、工程技术人员和科技工作者阅读参考。本书具有如下特点。

(1) 着重从应用的角度出发，强调理论与工程实际的紧密结合，突出实用性。

(2) 内容由浅入深且易于理解。

(3) 每章均有典型的应用实例，并给出了详实的步骤及其结果分析。

本书各章的主要内容包括：第1章介绍统计学习理论、支持向量机的分类模型和回归模型、支持向量机的研究现状以及支持向量机的参数优化方法；第2章分析支持向量机二次规划问题中的优化目标和约束条件问题，介绍一种新的支持向量机快速建模方法——连续过松弛和严格凸二次规划的支持向量机，并将该方法应用于冷轧带钢热镀锌生产过程建模；第3章讨论数据分布不平衡情况下的建模方法，介绍粒度计算和模糊C均值聚类方法，在此基础上详细分析了模糊C均值和共享最近邻(shared nearest neighbor，SNN)相似度的粒度支持向

量机方法，最后将该方法用于甲醇合成过程的质量建模；第 4 章讨论大数据的增量建模方法，分析凸壳理论的几何意义及其求解方法，结合 KKT 条件，介绍基于凸壳和 KKT 条件的增量支持向量机建模方法；第 5 章讨论支持向量机在线建模方法，分析混合核函数及误差校正方法对在线建模的影响；第 6 章讨论粒子群智能优化算法，从粒子群的更新策略、惯性权重、多种群 3 个方面分析了粒子群方法的特点，并用粒子群算法优化支持向量机的结构参数；第 7 章讨论蚁群智能优化方法，分析蚁群优化方法的原理和特点，介绍分段蚁群优化方法和权重蚁群优化方法。

本书涉及的研究内容、成果得到了国家自然科学基金“基于支持向量机和群智能的煤制甲醇合成过程建模及优化方法研究”(21366017)和内蒙古自然科学基金重大项目“基于数据驱动的带钢热镀锌生产过程工艺参数优化的研究”(2011ZD08)资助。本书由王建国和张文兴完成主要撰写工作，此外，秦波参与了第 2 章和第 3 章、杨斌参与了第 4 章和第 7 章、刘文婧参与了第 6 章的撰写工作，云海滨、丛宽、范凯、陈良武、武丽明、赵元元、张志杰、陈肖洁等硕士研究生参与了部分研究工作，在此向他们表示感谢。

当今世界正处于知识大爆炸的时代，知识的更新日新月异，由于作者理论水平有限，以及所做研究工作的局限性，书中难免存在不妥之处，恳请广大读者批评指正。

作　者

2015 年 6 月

目 录

CONTENTS

第1章

绪　论

1.1　引言

人类智慧中一个重要的方面表现为从过去的数据和以往的知识中学习的能力。通过归纳学习，得到对客观世界的认识和规律。这些规律可以帮助人类认识现有世界，同时帮助人类对未知现象做出正确的预测和判断。人们把从过去的数据和以往的知识中学习并获取规律的能力称为学习能力。用获得的规律不仅可以解释已知实例，而且能够对未知现象做出正确的预测和判断，这种能力称为推广能力。

在人工智能的研究中，人们用计算机来模拟这种学习能力的问题常常被称为基于数据的机器学习问题。基于数据的机器学习的任务就是设计某种方法和模型，通过对已知数据的学习，找到数据内在的相互依赖关系，从而对已知数据进行预测和判断，其最终目的是使机器具有良好的推广能力。

统计学在机器学习中起着重要的基础作用，但是传统的统计学研究的主要是渐进理论，即当样本趋于无穷多时的统计性质。所以基于传统统计理论的各种学习方法，都是以样本无穷多为前提来推导算法。然而现实中，我们面对的样本数量往往是十分有限的，通常的方法是仍以样本无穷多为假设进行算法推导和建模。当样本数较少时，用这种方法得到的结果有时是难以令人满意的。例如，在普通的神经网络学习中，当样本数有限时，本来很好的学习机器却表现出很差的推广能力，这种现象被称为过学习现象。

为了解决这类问题，人们进行了坚持不懈的研究工作，直到 20 世纪 70 年代，Vapnik 等人开始建立一种研究有限样本情形下统计规律及学习方法性质的理论即统计学习理论(statistical learning theory，SLT)[1,2]，它为有限样本的机

器学习问题建立了一个良好的理论框架，较好地解决了小样本、非线性、高维数和局部极小点等实际问题，其核心思想就是学习机器要与有限的训练样本相适应。直到1995年，Vapnik和他的合作者们明确提出一种新的通用学习方法——支持向量机(support vector machine，SVM)[1,2]后，该理论才受到广泛重视并应用到不同领域。支持向量机方法目前仍处于研究阶段，它仍存在许多有待进一步研究和探讨的问题，如为保证学习机器具有最佳推广能力而需进行模型参数的优化选择问题；在处理大规模数据和数据分布不平衡时，为提高模型的精度和速度问题；为了提高模型动态实时性的在线建模问题等。

1.2 模型预测方法

模型预测方法是机器学习的一个重要研究领域，绝大部分机器学习算法的应用都涉及对数据的预测性学习，目的是从已知数据中估计相关性以达到准确预测未来数据变化。经典的模式识别方法、各种统计方法以及近些年来出现的各种神经网络算法都被应用于估计这种数据中固有的相关性，试图建立一个能预测未来数据的模型。虽然，这些应用在一些特殊领域取得了一定进展，并且提出了一些新的概念和方法，但是，这些方法有其固有的算法缺陷：由于不知道数据的分布密度以及有限的训练数据，常常导致病态的训练结果，预测效果往往不尽如人意。人们迫切需要关于预测学习算法共同的概念和理论框架以及更高效的预测学习方法。到目前为止，几乎所有的预测学习算法都可以归结为以下三类。

(1) 传统的统计预测方法

这类方法是在参数结构形式已知的前提下，通过训练数据，预测各参数的值，如最小二乘法。应用这些预测方法除了需要很强的先验知识外，还需预先知道模型的结构形式。但是，在处理大量的实际预测问题时，常常不知道背景知识，面对一大堆采样数据，也不知道模型的结构形式。由于传统统计学研究的前提是样本数目趋于无穷大时的渐进理论，而参数预测方法几乎都是建立在这一前提基础之上的。因此只有当采样数据趋于无穷时，参数方法的训练结果才趋于真实模型。由于实际样本数目是有限的，很难满足这一前提。

(2) 经验非线性预测方法

20世纪80年代发展起来的人工神经网络和柔性统计方法就属于这一类。虽然，这些新方法克服了参数预测方法的部分弱点，能够依照需要，假设数据内在相关性而构造非线性模型。然而，这些非线性方法缺乏统一的数学理论基础，通常是从生物学的理论和一些学术流派中得到灵感，对于诸如神经网络中的结

构选择和权重初值的设定，仍需要借助于经验，得到的模型通常是局部最优解，而非全局最优。

(3) 统计学习理论

早期的统计学习理论是从 20 世纪 60 年代发展起来的，随着 VC 维理论和结构风险最小化原则的提出，进一步丰富和发展了统计学习理论，使它不仅是一种理论分析工具，还是一种能构造具有多维预测功能的预测学习算法的工具，使抽象的学习理论能够转化为通用的实际预测算法。

1.3 统计学习理论基础

统计学习理论被认为是目前针对小样本统计估计和预测学习的最佳理论。它从理论上系统地研究了经验风险最小化原则成立的条件、有限样本下经验风险与期望风险的关系，以及如何利用这些理论找到新的学习原则和方法等问题。

1.3.1 VC 维

统计学习理论是关于小样本进行归纳学习的理论，其中一个重要的概念就是 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里的函数按照所有可能的 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。有界实函数集的 VC 维可以通过用一定的阈值将它转化成指示函数集来定义。

VC 维反映了函数集的学习能力，一般而言，VC 维越大则学习机器越复杂，学习容量越大。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。一般地，对于 n 维空间$\mathbb{R}^n$ 中，最多只能有 n 个点是线性独立的，因此$\mathbb{R}^n$ 空间超平面的 VC 维是 $n+1$。

但是对于非线性学习机器而言，VC 维与独立参数的个数之间并没有明确的对应关系，非但如此，在非线性情况下学习机器的 VC 维通常是无法计算的。但是实际中在应用统计学习理论时，可以通过变通的办法巧妙地避开直接求 VC 维的问题。

1.3.2 经验风险最小化原则

学习的目的是根据给定的训练样本求系统输入输出之间的依赖关系。学习问题可以一般地表示为变量 y 与 $\boldsymbol{x}$ 之间存在的未知依赖关系，即遵循某一未知

的联合概率 $F(\boldsymbol{x},y)$。机器学习问题就是根据 n 个独立同分布观测样本

$$(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n) \tag{1.1}$$

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

$$R(\boldsymbol{w})=\int L(y,f(\boldsymbol{x},\boldsymbol{w}))\mathrm{d}F(\boldsymbol{x},y) \tag{1.2}$$

最小。其中，$\{f(\boldsymbol{x},\boldsymbol{w})\}$称为预测函数集；$\boldsymbol{w}$ 为函数的广义参数。$\{f(\boldsymbol{x},\boldsymbol{w})\}$可以表示任何函数集，$L(\boldsymbol{y},f(\boldsymbol{x},\boldsymbol{w}))$为由于用 $f(\boldsymbol{x},\boldsymbol{w})$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。

在传统的学习方法中，采用了所谓的经验风险最小化(empirical risk minimization，ERM)原则，即用样本定义经验风险：

$$R_{\mathrm{emp}}(\boldsymbol{w})=\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(\boldsymbol{x}_i,\boldsymbol{w})) \tag{1.3}$$

机器学习就是要设计学习算法使 $R_{\mathrm{emp}}(\boldsymbol{w})$最小化。

在早期的神经网络研究中，人们总是把注意力集中在如何使 $R_{\mathrm{emp}}(\boldsymbol{w})$更小，但很快发现一味追求训练误差小并不是总能达到好的预测效果。在某些情况下，当训练误差过小反而会导致推广能力下降。此即几乎所有神经网络研究者都曾遇到过的过学习问题。出现过学习问题的原因有：一是学习样本不充分，二是学习机器设计不合理。这两个问题是相互关联的，在神经网络学习中，对于有限样本，如果神经网络的学习能力过强，足以记住每一个训练样本，此时经验风险很快就可以收敛到很小甚至为零，但是无法保证它对新训练样本能够得到好的预测。因此，在有限样本情况下，经验风险最小化并不一定意味着期望风险最小，而且学习机器的复杂性不仅要与所研究的系统有关，还要与有限的学习样本相适应。即有限样本下学习机器的复杂性和推广能力之间存在着矛盾，采用复杂的学习机器容易使学习误差更小，但却往往丧失了推广能力。

1.3.3 结构风险最小化原则

传统机器学习方法中普遍采用的经验风险最小化原则在样本数据有限时是不合理的，因为经验风险和置信范围需同时最小。实际上，在传统方法中，选择学习模型和算法的过程就是优化置信范围的过程，若选择的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结果。譬如在神经网络中，需要根据问题和样本的具体情况来选择不同的网络结构(对应不同的 VC 维)，然后进行经验风险最小化。

为解决经验风险和置信范围这两项最小化风险泛函问题，引出结构风险最

小化(structural risk minimization,SRM)原则。

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 $R_{\text{emp}}(\boldsymbol{w})$ 和实际风险 $R(\boldsymbol{w})$之间以至少 $1-\eta$ 的概率满足如下的关系:

$$R(\boldsymbol{w}) \leqslant R_{\text{emp}}(\boldsymbol{w}) + \sqrt{\frac{h[\ln(2n/h)+1]-\ln(\eta/4)}{n}} \tag{1.4}$$

其中,h 是函数集的 VC 维;n 是样本数;η 是满足 $0 \leqslant \eta \leqslant 1$ 的参数。

由此可见,统计学习的实际风险 $R(\boldsymbol{w})$由两部分组成:一是经验风险(训练误差),另一部分称为置信范围(VC confidence)。置信范围反映了真实风险和经验风险差值的上界,反映了结构复杂所带来的风险,它和学习机器的 VC 维 h 及训练样本数 n 有关,式(1.4)可以简单地表示为

$$R(\boldsymbol{w}) \leqslant R_{\text{emp}}(\boldsymbol{w}) + \Phi(h/n) \tag{1.5}$$

由式(1.4)右边第二项可知,$\Phi(h/n)$随 h 加大而增长,因此,在有限训练样本下,学习机器的复杂性越高,VC 维越高,则置信范围越大,就会导致真实风险与经验风险之间可能的差别越大。

这里的泛化界限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。当 $h/n>0.37$ 时,这个界限肯定是松弛的。当 VC 维无穷大时,这个界限就不再成立。

为了构造适合于小样本学习的归纳学习原理,可以通过控制学习机器的泛化能力来达到此目的。对于完全有界非负函数 $0 \leqslant f(\boldsymbol{x},\boldsymbol{w}) \leqslant B$,以至少 $1-\eta$ 的概率满足以下不等式:

$$R(\boldsymbol{w}) \leqslant R_{\text{emp}}(\boldsymbol{w}) + \frac{B\varepsilon}{2}\left(1+\sqrt{1+\frac{4R_{\text{emp}}(\boldsymbol{w}_1)}{B\varepsilon}}\right) \tag{1.6}$$

运用以上公式可以控制基于固定数目的经验样本之上的风险函数最小化过程。控制最小化过程有几种方法:一是最小化经验风险,根据上面的公式可以看出,风险的上界随着经验风险值的减少而减少;二是最小化式(1.6)中右边的第二项,将样本数目与 VC 维作为控制变量,后者适合于样本数较小的情况。

为了寻找使得结构风险达到最小的函数 $f(\boldsymbol{x},\boldsymbol{w})$,统计学习理论提出了一种新的策略,它将函数集 $\boldsymbol{S}=\{f(\boldsymbol{x},\boldsymbol{w}),w\in\Lambda\}$看成具有一定的结构,并由一系列嵌套的函数子集 $\boldsymbol{S}_k=\{f(\boldsymbol{x},\boldsymbol{w}),\boldsymbol{w}\in\Lambda_k\}$组成,它们满足

$$\boldsymbol{S}_1 \subset \boldsymbol{S}_2 \subset \cdots \subset \boldsymbol{S}_k \subset \cdots \subset \boldsymbol{S}_n$$

函数嵌套子集构成的函数集结构如图 1.1 所示。

为了选择合适的 $\boldsymbol{S}_k$ 作为学习函数,可以将式(1.4)右边划分为两部分,即第一项经验风险和第二项置信范围。如果给定样本数目 n,那么,随着 VC 维数目 h 的增加,经验风险逐渐变小,而置信范围逐渐递增。如图 1.2 所示,真实风险

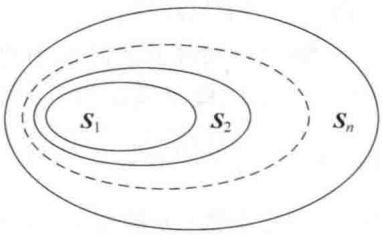

图 1.1 函数嵌套子集构成的函数集结构

的界是经验风险和置信范围之和，随着结构元素序号的增加，经验风险将减小，而置信范围将增加。最小的真实风险上界是在结构的某个适当的元素上取得的。综合考虑经验风险与置信区的变化，可以求得最小的风险边界，它所对应的函数集的中间子集 $\boldsymbol{S}^*$ 可以作为具有最佳泛化能力的函数集合。

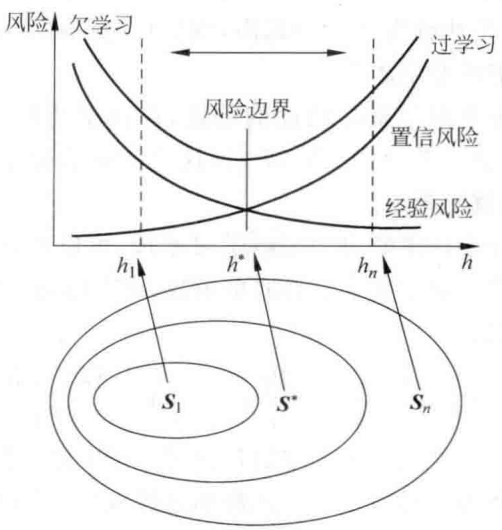

图 1.2 结构风险最小化原理图

根据图 1.2，可以得到以下两种运用结构风险最小化原理构造的学习机器的思想。

(1) 给定一个函数集合，按照上面的方法来组织一个嵌套的函数结构，在每个子集中求取最小经验风险，然后选择经验风险与置信风险之和最小的子集。当子集数目较大时，此方法较费时，甚至不可行。

(2) 构造函数集合的某种结构，使得在其中的各函数子集均可以取得最小的经验风险(如使得训练误差为 0)。然后，在这些子集中选择适当的子集使置信风险最小，则相应的函数子集中使得经验风险最小的函数就是所求解的最优

函数。

可见，利用结构风险最小化原理的思想，就可以完美解决传统机器学习中的过学习问题。支持向量机通过最大化分类边界以最小化 VC 维，也即在保证经验风险最小的基础上最小化置信风险，从而达到最小化结构风险的目的，因此支持向量机方法实际上就是结构风险最小化思想的具体实现。

1.4 支持向量机的提出

统计学习理论的 VC 维理论和结构风险最小化原则的提出为支持向量机算法打下了坚实的理论基石。

20 世纪六七十年代，非参数统计学以及不适定问题的解决方法的发现，特别是算法复杂度的概念及其与归纳推理的关系的提出，推动了统计学发展。算法复杂度的思想是统计学和信息论中最伟大的思想之一。到 20 世纪 70 年代，在算法复杂度思想的基础上，形成了对于学习问题的最小描述长度(minimum description length，MDL)归纳推理原则。如何在有限数量的经验数据基础上进行依赖关系估计的问题，这些思想改变了人们对算法复杂度的认识。1963 年，Vapnik、Lerner 以及 Chervonenkis 开始研究描述学习理论的非线性普遍算法，他们开创了实现统计机器学习算法的先河。从此，以 Vapnik 为代表的许多学者对统计机器学习的具体内容进行了不断地丰富和发展[3]。

1968—1971 年，Vapnik 和 Chervonenkis 在《The Necessary and Sufficient Conditions for the Uniforms Convergence of Averages to Expected Values》一文中提出了一个重要的理论基础——VC 维理论[4]。利用 VC 熵和 VC 维的概念，发现对指示函数空间的大数定律及其与模式识别问题的关系，并创造出对于经验风险最小化原则下的模式识别的一个一般的非渐近理论。

1974 年，Vapnik 形成结构风险最小化原则[5]，1982 年，Vapnik 在《Estimation of Dependences Based on Empirical Data》[6]一书中首次提出这一具有划时代意义的结构风险最小化原理，堪称为支持向量机算法的基石。

1974—1979 年，Vapnik 创造基于 ERM 和 SRM 原则上的一般的非渐近学习理论。

1981 年，Vapnik 把大数定律推广到实值函数空间[4]。

1987 年，Vidyasaga 出版了专著《ATheory of Learningand Generalization》，该书从数学角度详细讨论了统计学习理论。

1991 年，Vapnik 发现 ERM 原则与最大似然方法一致的充分必要条件[7]，完成了对经验风险最小化归纳推理的分析。