



ZHICHI XIANGLIANGJI ZAI

QILUNJI ZHUANGTAI JIANCE ZHONG DE YINGYONG

支持向量机在 汽轮机状态监测中的应用

王 雷◎著



北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP

北京师范大学出版社

图书在版编目(CIP)数据

支持向量机在汽轮机状态监测中的应用 / 王雷著. —北京:
北京师范大学出版社, 2012.6
ISBN 978-7-303-14026-8

I. ①支… II. ①王… III. ①向量计算机—算法理论—
应用—汽轮机运行—研究 IV. ①TK267

中国版本图书馆 CIP 数据核字(2012)第 017847 号

营销中心电话 010-58802181 58805532
北师大出版社高等教育分社网 <http://gaojiao.bnup.com.cn>
电子邮箱 beishida168@126.com

出版发行: 北京师范大学出版社 www.bnup.com.cn

北京新街口外大街 19 号

邮政编码: 100875

印刷: 北京京师印务有限公司

经销: 全国新华书店

开本: 170 mm × 230 mm

印张: 10.25

字数: 184 千字

版次: 2012 年 6 月第 1 版

印次: 2012 年 6 月第 1 次印刷

定价: 20.00 元

策划编辑: 胡 宇 责任编辑: 岳昌庆 胡 宇

美术编辑: 毛 佳 装帧设计: 李尘工作室

责任校对: 李 茵 责任印制: 李 啸

版权所有 侵权必究

反盗版、侵权举报电话: 010-58800697

北京读者服务部电话: 010-58808104

外埠邮购电话: 010-58808083

本书如有印装质量问题 请与印制管理部联系调换。

印制管理部电话: 010-58800825

前 言

随着汽轮机发电机组向高科技、大容量发展，极大地提高了电力生产能力，带来了瞩目的社会效益。与此同时，人工智能技术和电子信息技术的广泛应用，对电力生产过程中性能监测与故障诊断提出了更高的要求，也为其提供了更丰富的信息来源和更强大的信息处理手段，极大地促进了性能监测和故障诊断理论与技术在该领域的发展。机器学习是人工智能在监测与诊断技术中应用的重要研究领域，传统的基于经验风险最小化原则的学习机器(如人工神经网络)面临着如何从有限样本中得到具有较大推广能力的决策函数的问题，因其渐近性的前提条件得不到满足，而难以取得令人满意的结果。在这种情况下，基于经验风险最小化原则的学习机器就很难得到有价值的、具有推广意义的应用。因此，在工程实际中对解决小样本学习问题的方案有着迫切的需求。

20 世纪 90 年代中期，在统计学习理论的基础上，发展出了一种新的方法——支持向量机。在解决小样本、非线性及高维模式识别问题中，它表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。虽然统计学习理论和支持向量机方法尚有许多问题需要进一步研究，但它们正在成为继模式识别和神经网络研究之后机器学习领域新的研究热点。

支持向量机算法一经提出，就得到国内外学者的高度关注，学术界普遍认为它是继神经网络之后一个新的研究方向。支持向量机算法的优良性能使得其在短短十几年时间内得到广泛的应用，取得了一系列令人瞩目的研究成果。从最初的线性分类，到各种模式识别问题、函数回归分析以及时间序列预测分析

等方面都有出色的表现，在众多领域都能找到支持向量机的影子。

近几年来，随着对电厂运行设备性能监测与故障诊断需求的提高，使得支持向量机在这个领域也开始得到应用，并取得了较好的效果。但是，目前在电站热力设备性能监测与故障诊断方面上应用支持向量机算法的相关著作较少，汽轮机组的性能监测与故障诊断方面还没有。在此背景下，作者完成了此书，目的在于促进支持向量机在汽轮机组性能监测和故障诊断领域的推广与应用。本书主要以作者在东南大学徐治皋教授指导下完成的博士学位论文为基础，引用了国内外研究者特别是作者的一些最新研究成果，全面介绍支持向量机算法在汽轮机组性能监测和故障诊断领域中的应用研究。

全书共分七章。第1章是绪论，介绍了汽轮机组状态监测和故障诊断的内容、意义和当前的发展状况、存在的问题；第2章为支持向量机算法的介绍；第3章主要介绍基于支持向量回归在汽轮机组性能监测中的应用，研究了模型建立的基本步骤及其参数的选择方法；第4章和第5章分别研究了基于支持向量机算法的时间序列参数预测模型和运行参数异常值检测方法；第6章把支持向量机多值分类算法引入到汽轮机辅助设备故障诊断中，研究了支持向量机用于故障诊断的关键问题；第7章总结了全书的主要内容并对支持向量机算法在电厂中的应用前景进行了展望。

本书可供热能与动力工程领域设计和应用的科研和工程技术人员参考使用，对于从事支持向量机算法在工程应用的其他领域的工程技术人员与研究者也具有一定的参考作用，也可作为上述相关专业的高等学校教师、学生的选修教材或教学参考书。

在本书成文的过程中，得到了东南大学徐治皋教授、东北电力大学李勇教授的悉心指导与帮助，李勇教授就书中问题提出了许多有益的建议，进行了大量的讨论，并慷慨分享了大量的文献资料。衷心感谢沈阳工程学院能源与动力工程系的同事，他们对本书的编写工作给予了热情的关怀、鼓励和支持。本书的部分研究内容还得到了沈阳工程学院博士启动基金的资助。

此外，本书参考了大量的国内外参考文献，特别是支持向量机原理部分，引用了较多前人整理的材料，引文出处已尽可能列在每章后，但难免有疏漏，特别是反复引用但无法查到出处的。这里对所有参考文献的作者表示诚挚的感谢。

限于作者的水平与实践经验，书中难免出现错误和不当之处，敬请读者批评指正。

王雷

2011年10月于沈阳

目 录

第 1 章 绪 论	(1)
1.1 引言	(1)
1.2 汽轮机状态监测简述	(5)
1.3 基于机器学习的状态监测方法	(8)
1.4 本书的安排	(16)
参考文献	(17)
第 2 章 统计学习理论与支持向量机	(25)
2.1 机器学习的表示	(26)
2.2 统计学习理论主要内容	(30)
2.3 分类支持向量机	(36)
2.4 回归支持向量机	(43)
2.5 支持向量机与神经网络	(46)
2.6 小结	(48)
参考文献	(49)
第 3 章 汽轮机性能参数监测	(51)
3.1 概述	(51)
3.2 回归方法简介	(53)
3.3 凝汽器真空应达值估计	(57)

3.4	汽轮机热耗率计算	(69)
3.5	主蒸汽流量软测量	(75)
3.6	小结	(77)
	参考文献	(78)
第4章	时间序列参数预测	(80)
4.1	概述	(80)
4.2	时间序列预测简述	(82)
4.3	凝汽器清洁系数时间序列预测	(92)
4.4	汽轮机转子振动状态的预测	(99)
4.5	小结	(109)
	参考文献	(110)
第5章	汽轮机运行异常参数检测	(112)
5.1	概述	(112)
5.2	基于支持向量机的数据异常值检测	(114)
5.3	小结	(120)
	参考文献	(121)
第6章	汽轮机组及其附属设备故障诊断	(123)
6.1	概述	(123)
6.2	支持向量机多分类方法	(127)
6.3	故障征兆模糊处理	(130)
6.4	凝汽设备故障诊断	(135)
6.5	高压加热器故障诊断	(148)
6.6	小结	(153)
	参考文献	(154)
第7章	结论与展望	(156)

第1章 绪论

1.1 引言

经济发展,电力先行,能源是国民经济的基础产业,对经济持续快速发展和人民生活水平的改善发挥着十分重要的保障与促进作用。电力是唯一能够大规模利用煤炭、水力、核能的二次能源,具有能高度集中和无限划分、清洁干净和易于控制、可大规模生产和远距离输送等诸多优良特性。因此,电力的使用已深入到社会经济、生活的各个领域,电力的发展和应用程度成了衡量社会现代化水平高低,以及物质文明和精神文明高低的重要标志之一。

我国是生产大国,到2010年我国发电装机超过9.5亿kW,如果按此速率,我国发电装机将在未来两年内超过美国达到世界首位。目前我国电力工业结构中,火电设备容量占总装机的2/3左右。“十二五”期间,全国规划煤电开工规模3亿kW,其中煤电基地开工1.97亿kW(占66%),到2015年我国煤电装机预计达到9.33亿kW。因此,煤炭仍然是我国的主要能源。到2015年,煤炭占能源需求总量的比例达到58.5%,2020年仍将占到51%,到2050年我国的一次能源消费总量可能达到200多亿吨标准煤。如此巨大的不可再生能源消耗,对我国煤炭工业造成空前的压力。从国际上来看,我国常规能源资源占世界总量的10.7%,但人均能源资源占有量远低于世界平均水平,能源资源形势十分严峻,而且能源利用率处于较低水平,火电机组年平均供电煤耗比国外先进水平高约20克标准煤/千瓦时^[1]。因此,深入研究火电机组的优化

运行,实现火电厂的高效经济运行,不仅能降低能耗,提高电厂自身的经济效益,有效地缓解我国能源紧张的问题,还对我国国民经济和社会可持续发展具有重要的战略意义。

随着电力工业的不断发展和电力体制改革的深入开展,电力系统的市场经济体系逐渐建立,传统的电力生产和供应垄断体制将被打破,厂网分开、竞价上网已成为必然。发电企业将面临巨大的市场竞争,企业成本控制也会越来越受到重视。大型火力发电机组是电力生产的主要设备,其运行的安全性与经济性问题在这种竞争条件下更显突出。由于大型发电设备结构的复杂性和运行环境的特殊性,机组出现性能退化甚至故障不可避免。由于运行管理水平的制约,即使机组设备状态良好,也难以维持最佳经济性。大型机组单机容量大、热力参数高,其性能退化与设备故障以及运行水平对安全性、经济性的影响愈发严重。及时诊断、预防机组的故障,以及决策运行操作和故障维修,提高设备的可用率,保障机组安全、经济地运行,仍然是发电企业提高经济效益、增强竞争力的有效途径之一。因此,在当前形势下,进一步开展机组性能监测、故障诊断的研究工作具有显著的现实意义。

2
汽轮机及其附属设备系统是火力发电机组进行能量转换的主要场所,其设备及总体运行状态的优劣将直接影响整个机组的发电效率。良好的汽轮机系统运行状态是整个发电机组安全、经济运行的保障,研究针对汽轮机及其附属设备系统的状态监测和故障诊断,开发具有实际应用价值的系统是目前亟待深入的工作之一。

现代信息技术的发展,在整个社会的各个领域发挥着巨大作用。信息技术在电力行业的广泛应用,极大地提高了电力生产能力,带来了瞩目的社会效益。人工智能技术在现代电厂的广泛应用,一方面对性能监测与故障诊断提出了更高的要求,另一方面也为其提供了更丰富的信息来源和更强大的信息处理手段,并极大地促进了性能监测与故障诊断理论和技术的发展^[2,3]。

同时,各种数学理论与数学方法也不断渗透到电力生产中,模糊数学、神经网络、灰色系统理论、粗糙集理论、分形几何、混沌理论等在性能监测与故障诊断领域的应用研究异常活跃,并取得可喜的成果。这为汽轮机组的性能监测与故障诊断提供了强有力的数学工具。以人工智能技术为核心的监测与诊断技术在电站中的综合与应用,不但推动了汽轮机组性能监测与故障诊断的发展,而且代表了其今后的研究方向。

以人工智能技术为核心的性能监测与故障诊断技术是以常规监测和诊断技

术为基础,以人工智能为核心,研究内容和实现方法与传统的监测和诊断技术对比,已经发生了重大的变化。监测和诊断过程的知识化使得人们致力于研究知识的获取、知识的表示与组织、推理方法的结构模型的建立以及监测与诊断策略的形成等问题。

基于人工智能的性能监测与故障诊断技术代表了监测与诊断技术的发展方向,同时,其发展与人工智能技术的发展密切相关,特别是专家系统在监测与诊断技术中的应用为性能监测与故障诊断的智能化提供了可能性,使监测与诊断技术进入了新的发展阶段。继专家系统之后,机器学习成为人工智能在监测与诊断技术中应用的又一重要研究领域。机器学习是提高性能监测与故障诊断智能水平的主要途径,一旦监测与诊断系统具有自学习能力,它就能从环境的变化中学习新知识并不断实现自我完善,它不同于只是依赖于经验知识库却不能创新和发展的专家系统。作为机器学习的一个发展阶段和代表算法,在过去的十年里,人工神经网络理论的研究及其应用获得了长足的进步,成为计算机与人工智能等相关专业的研究热点。由于人工神经网络具有很强的自学习能力,因而吸引了众多的研究学者^[4~18]。

人工神经网络的算法基础是传统统计学,所研究的主要是渐近理论,即当样本量趋向于无穷多时的统计性质。但在现实问题中,如故障诊断,特征知识的获取通常有一定的约束,样本的数目通常是有限的,甚至是十分有限的。在这种情况下,我们希望在有限的样本数据下尽可能地发现其中蕴涵的知识,强调学习机器具有较强的推广能力,即对符合某规律但没有学习过的样本也能给出合理的结论,这是学习机器体现其智能性的最为重要的一个方面。但对于神经网络算法,当样本数有限时,训练效果良好的一个算法结构却可能表现出很差的推广能力,即所谓的神经网络过学习问题^[19,20]。

进一步分析传统的学习机器,如人工神经网络是基于经验风险最小化原则(empirical risk minimization, ERM),即仅仅试图使经验风险最小化,并没有使实际风险最小化,因而造成了推广性方面较为严重的缺陷。近年来随着人工神经网络研究的不断深入,人们认识到神经网络的学习算法缺乏定量分析和机理完备的理论结果,人工神经网络的研究遇到了较大的困难,从而使新的学习算法的研究成为机器学习的研究热点和关键问题。

统计学习理论是建立在坚实的理论基础之上的,为解决小样本学习问题提供了统一的框架,致力于寻求小样本情况下学习问题的最优解,而不需要利用样本量趋于无穷大的渐近性条件。这使得统计学习理论在小样本情况下同样具

有推广的价值，因此，适合于监测和诊断这种小样本情况的工程实际问题^[20]。

机器学习是人工智能中最具智能特性和最前沿的研究领域之一，机器学习的研究取得重大进展往往意味着人工智能甚至整个计算机科学向前迈出坚实的一步。应当看到，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，毕竟只有几年时间，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起步。

本书正是在此背景下，研究和探讨将由统计学习理论发展而来的支持向量机(support vector machine, SVM)算法应用于汽轮机组性能监测和故障诊断领域需要考虑的问题和具体的应用实现。

1.2 汽轮机状态监测简述

1.2.1 汽轮机状态监测的目的

随着汽轮机组向高参数、大容量、高自动化方向发展,其系统越来越复杂,设备出现故障的可能性越来越大,故障的危害性也越来越大。近几十年来,国内外已发生多起汽轮机组整机毁坏事故,因设备故障而导致重大经济损失和人员伤亡的事件时有发生。因此,及时掌握机组的运行状况,保证汽轮机组的安全经济运行是十分必要的。

汽轮机组运行的安全性和可靠性取决于两个方面:一是设备设计与制造的各项技术指标的实现;二是设备安装、运行、管理、维修和诊断措施的实施。其中,机组性能监测和故障诊断是确保汽轮机组安全运行的重要手段,其目的是:

1. 保证机组发挥最大的设计能力,制定合理的维修制度,以便在允许的条件下充分挖掘设备潜力,延长服役期限和使用寿命,降低设备寿命的周期费用。
2. 及时、正确地对各种异常状态或故障状态做出诊断,并及时预防或消除故障,对设备的运行进行必要的指导,提高设备的可靠性、安全性和有效性,从而将故障损失降低到最低水平。
3. 预报机组潜在的故障,避免经济损失和人身伤亡事故的发生。

1.2.2 汽轮机状态监测的特点

汽轮机组属于大型设备,由于工作环境的特殊性,决定了其故障的复杂性和多态性,从而决定了监测与诊断任务的特殊性。具体表现在以下几个方面^[21]:

1. 监测多元化。具体含意是:采集信号种类多样化,不仅要采集模拟量信号,而且还对开关量信号进行采集;同一类信号测点分布化,即对重要参数在机组的不同部位布置测点,并在数据采集板上留出足够的通道数,以保证全面地了解该类信号所反映的机组运行状况;不仅对机组的稳态工况进行信号采集,而且还对非稳态工况(瞬态)进行监测,比如对启停过程、过临界点过程以及异常工况的信号进行采集。

2. 监测连续化. 汽轮机组的工作过程是一个动态过程, 为了确保机组安全、可靠地运行, 提高其生产效率, 必须进行连续监测, 以便及早地发现并处理机组存在的隐患.

3. 数据采集与处理的实时化. 为保证实时地跟踪、监测机组的运行变化, 数据采集与处理要满足一定的时间限制, 以便能快速反映机组运行性能状况.

4. 数据记录历史化. 在进行故障诊断时不仅需要当前数据, 同时也需要历史数据, 从机组的状态变化趋势中, 确认机组所处的状态, 并对当前数据和历史数据进行记录.

5. 监测和诊断的并行与协作. 为了提高系统的实时性和模块化, 必须使监测功能和诊断功能并行执行. 两者既要相互联系, 又要相对独立, 不能因为诊断而中断监测.

1.2.3 汽轮机状态监测的内容

机组性能监测与故障诊断是通过电站设备及系统的参数进行在线测量和分析, 以一定的安全性、经济性指标对机组当前的运行状况作出评价, 从而对机组的运行状况与调整方案进行决策. 其主要包括以下几项内容:

1. 性能指标计算. 性能指标是指电站全厂及主要系统的性能参数, 如热耗、效率、煤耗等, 这些指标由现场采集的压力、温度、流量等原始参数通过一定计算获得. 其算法遵守一些试验规程, 如美国机械工程师学会(American Society of Mechanical Engineers, ASME)的电站性能试验规程以及国家汽轮机热力试验规程等. 由于试验规程对测量元件与现场测点等要求比较严格, 最初的性能计算颇受局限. 但随着测量技术、传感技术以及计算机技术的不断进步, 目前已经能够在线完成全厂及主要系统的性能分析^[22].

2. 运行目标值优化. 这是指在机组正常带负荷运行阶段, 调整机组参数, 使机组达到或者接近运行优化目标值, 提高机组经济性. 运行优化目标值包括机组当前最优化运行状况的各个运行参数和性能指标的优化目标值, 它为运行人员提供了机组在不同外部条件(负荷、环境、燃料等)下的最佳运行方式和参数控制. 它建立在现有设备基础上(包括热力系统结构、设备的运行状态等), 主要通过运行调整实现, 其目的就是使机组一直处于最优状态运行^[23].

3. 偏差分析. 这是指机组主要运行参数偏离其运行目标值而对机组热耗率(煤耗率)的影响进行定量分析, 从而反映影响机组运行性能的主要因素. 通过参数的偏差分析, 可以监督考核运行人员的运行水平, 为运行人员提供优化

运行指导。运行参数的偏差对机组的经济性影响与机组负荷、运行环境和热力系统结构等因素有较大的关系，在偏差分析时必须根据机组运行的具体情况进行分析^[23~26]。

4. 故障诊断。机组故障诊断的任务是通过获取当前运行工况下足够的设备数据信息，通过过程异常变化的征兆与过程系统故障之间的内在联系，对整个机组或某单一设备的运行状态进行分析和判断，预测或查明故障发生的时间、位置、幅度、故障模式以及故障发展趋势^[27]。

1.3 基于机器学习的状态监测方法

1.3.1 传统的监测方法

1. 基于直接测量信号的方法

系统在正常工作情况下，其输出信号绝对值或者变化趋势必然在一个合理范围内变化。如果超出了这个合理范围，那么系统必然处于非正常工作状态。因此，可以根据直接测量的信号幅度及其变化趋势诊断过程故障，这是直接诊断的方法^[28,29]。系统的输出信号在幅值、相位、频率及相关性上与故障源存在着某种关系，因此利用这种关系，即可从测量的输出信号中确定系统的故障，这是信号处理的方法。信号处理方法在旋转机械振动故障诊断中应用较多，如谱分析、快速傅里叶变换、小波变换等。

基于直接测量信号的故障诊断方法简单、实用、可靠，在工程中应用广泛。但是，该方法只有在故障发展到相当程度并影响到外部特征时才有效，而且只能对故障的范围作出粗略的估计，大多数情况下不能直接定位故障^[30]。

2. 基于机器学习方法

由于电厂运行环境的复杂性与多变性，机组运行偏离最优工况的原因与多种因素有关，而对其机理研究往往不透彻，并且有些情况下非线性程度高、存在不确定性问题，精确的数学模型难以建立，因此基于数学模型监测和诊断技术的应用受到一定限制。基于机器学习的监测和诊断方法，可利用历史经验知识，通过多种渠道获得故障信息，不受数学模型的限制，因此基于机器学习的监测和诊断方法成为当前应用研究的重点。目前，较常见的方法有专家系统方法、神经网络方法、模糊数学方法以及各种综合方法。

(1) 专家系统方法^[31]。专家系统是指具有大量专门知识，并能运用这些知识解决特定领域中实际问题的计算机程序。完整的专家系统包括四部分：知识库、推理机、知识获取模块和解释界面。知识存储和管理各种知识，这些知识以事实和规则的形式存储起来。事实是短期信息，可以发生快速的变化，而规则是长期信息，相对稳定。推理机是协调控制整个系统工作的机构或规则解释机构，它根据知识库中的事实、规则，按一定的推理策略求解当前的问题。知识获取模块解决如何将各种信息转化为计算机可用的知识。解释界面负责向用户说明系统如何进行推理和如何得到结论。

(2)模糊数学方法. 不确定的、模糊的状态和特征等在实际系统中普遍存在, 模糊集合理论具有对不确定性的表达能力, 因此可以用模糊集合理论处理故障诊断中不确定的、模糊的问题. 常用方式是模糊推理、模糊聚类、模糊综合评判等. 模糊推理一般用于建立模糊专家系统, 是模糊集理论与专家系统的综合应用, 其知识用模糊隶属度表示、推理机采用模糊推理方法. 模糊聚类可将模糊集分为不同水平的子集, 由此判断故障最有可能属于的子集. 模糊综合评判就是根据模糊关系矩阵 R 建立被诊断对象的特征集上的模糊矢量 a 与故障状态集上的模糊矢量 f 之间的映射关系, 即 $f=R \cdot a$, 再根据判断准则大致判断有无故障.

(3)神经网络方法. 人工神经网络是由大量简单处理单元, 以某种特殊拓扑方式相互连接而成的、复杂的非线性动力学系统, 它从微观上对人脑的智能行为进行描述. 信息处理通过各单元之间的大规模连接权值与作用函数的并行运算实现. 信息存储表现为神经元之间分布式权值联系, 学习表现为神经元连接权值的动态调整. 神经网络故障诊断的主要方式有基于征兆的诊断方式^[32]、基于测量信号的诊断方式^[33]、基于残差生成的诊断方式^[34]和基于自适应误差补偿诊断的方式^[35]等. 与传统信息处理方法相比, 神经网络方法在很多方面具有明显的优势, 主要表现在具有很强的非线性表达能力、能够以任意精度逼近任意非线性函数、分布式信息存储与并行计算结构具有独特的容错性等. 强大的自适应学习能力增强了对环境变化的适应能力, 在故障诊断领域的应用研究一直是一个热点.

1.3.2 传统机器学习的不足之处

在实际问题中, 例如大机组故障诊断, 可用的样本量往往是有限的. 首先, 大机组系统非常昂贵, 其运行状态关系到企业的整个生产, 因此不可能让其长时间带故障运行; 其次, 故障的发生有一定的突然性, 而且往往是可遇不可求的, 并且多数情况下信号与故障之间的对应关系是非常模糊的, 而且有可能是一种信号对应多种故障. 因此, 诸多原因造成大机组的典型故障的样本量是非常有限的^[36].

传统统计学主要研究渐近理论, 即当训练样本的样本量趋向无穷大时的解决方案, 应用经验风险最小化来优化学习机器的参数, 但是理论表明, 经验风险与实际风险之间具有一定的差异, 在小样本情况下这种差异尤其明显. 以神经网络为代表的学习机器的重要理论基础之一就是传统统计学, 传统统计学是

样本量趋向于无穷大时的渐近理论，现有的学习方法也多基于此假设。但在实际问题中，可用于训练的样本数目往往是有限的，因此一些基于传统统计学的、理论上非常优秀的学习方法在实际应用中的表现却不尽如人意^[37,38]。

由传统统计学理论得到的机器学习原则是基于经验风险最小化原则，即保证学习机器对数目有限的训练样本学习能力最优，但实践结果证明，在有限样本上学习能力最优并不能保证对未知样本的推广能力最好。在某些情况下，太小的经验风险（对有限样本太好的学习能力）反而可能导致学习机器推广能力的下降，例如神经网络中的过学习现象。这是由于经验风险是利用有限数目的样本，即小样本计算得到的，而不是理论基础所要求的样本量趋于无穷时得到的结果。因此，实际问题中的小样本情况与理论基础所要求的无穷渐近性之间的矛盾，成为传统学习机器表现不好的一个主要原因。

传统学习机器的另外一个难点是结构选择问题。学习机器的结构越复杂，学习能力越强。例如神经网络，隐层节点数越多，结构越复杂，对有限样本的学习能力越强，反映在训练结果上是可以得到非常小的经验风险，但推广能力很差；相反，如果隐层节点数太少，则其学习能力不足，难以有效地从有限样本中学习到规律，从而产生欠学习现象。可见，学习机器的复杂度应该与实际问题的复杂度相对应。但就如何控制网络复杂程度，神经网络中却没有任何理论依据和实际可行方法，其结构的确定往往取决于使用者的经验和不断地调整，不但需要耗费大量的时间和精力，而且在某些问题上成功的经验很难直接应用到其他问题中去。

传统学习机器的第三个难点是局部极小值问题。神经网络主要的优化方法是梯度修正算法，而梯度修正算法只能保证收敛到局部极小值点，难以保证最后得到的网络有良好的推广能力。另外，梯度修正算法中参数的选择也非常重要，例如 BP 网络中有两个参数：步长和惯性项。这两个参数对网络能否收敛和网络的收敛速度有着非常重要的影响，但这两个参数的选择目前尚没有有效的方法，需要用多组值进行多次尝试。

传统学习机器面临的第四个难点是高维问题。而对高维问题，传统学习机器往往没法正常工作。这一方面是由于高维空间存在更多的局部极小值点，而不同的极小值点之间存在较大的差异，传统学习机器很容易收敛于这些局部极小值而表现出很大的随机性；另一方面，当数据维数很高时，神经网络训练和测试过程的计算量急剧增加，从而使训练和测试过程慢得难以忍受。因此，面

对高维问题,传统学习机器往往需要将数据通过一定的技术手段压缩至低维方可处理。

这种情况下,基于渐近性分析的学习机器就很难得到有价值的、具有推广意义的应用,因此,在工程实际中对解决小样本学习问题的有效方案有着迫切的需求。

1.3.3 基于统计学习理论的机器学习

早在20世纪70年代诞生的统计学习理论(statistical learning theory, SLT)系统地研究了机器学习问题,它为有限样本情况下的统计学习问题提供了一个有效的解决途径,弥补了传统统计学的不足。

统计学习理论的核心概念是VC维(Vapnik-Chervonenkis Dimension)^[39],它用VC维来描述学习机器的复杂度,并以此为出发点导出了学习机器推广能力的界的理论。该理论致力于寻找小样本情况下学习问题的最优解,而不需要利用样本量趋于无穷大的渐近性条件,从而使得统计学习理论在小样本下同样能够具有推广的价值。正因为统计学习理论是针对小样本情况进行研究的,所以从推广性的角度来看,统计学习理论更适用于汽轮机组性能监测和故障诊断这种实际的工程问题。

统计学习理论是一种专门研究小样本情况下机器学习规律的基本理论和数学构架,也是小样本统计估计和预测学习的最佳理论。Vapnik等人从20世纪六七十年代开始致力于此方面研究,到90年代中期,随着该理论的不断发展和成熟,产生了基于统计学习理论体系的新的通用机器学习方法,即SVM。它与神经网络完全不同:神经网络学习算法的构造是受模拟生物启发;而SVM的思想来源于最小化错误率的理论界限,这些界限是通过对学习过程的形式化分析得到的。基于这一思想产生的SVM算法,不但具有良好的数学性质(如解的唯一性、不依赖输入空间的维数等),而且在应用中也表现出了良好的性能,它所得到的最优解超过了传统学习方法(如神经网络)。由于统计学习理论为解决小样本情况下的机器学习问题提供了有力的理论基础,以及基于这一理论体系下的SVM所表现出的优良特性,人们已经开始重视起这一新的学术方向,尤其是Vapnik在1995年发表的著作《统计学习理论的本质》(《The Nature of Statistic Learning Theory》),极大地促进了统计学习理论的研究。