

**Support Vector Machine and its Application
in Mixed Gas Infrared Spectrum Analysis**

支持向量机理论及 工程应用实例

❖ 白 鹏 张喜斌 张 斌 编著
李 彦 谢文俊 刘君华

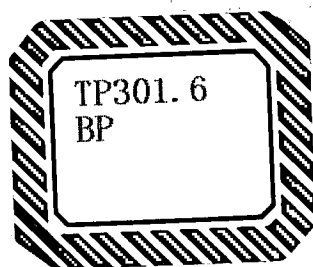


西安电子科技大学出版社

<http://www.xduph.com>

支持向量机理论及 工程应用实例

白 鹏 张喜斌 张 斌 编著
李 彦 谢文俊 刘君华



西安电子科技大学出版社

2008

内 容 简 介

本书从机器学习的基本问题开始,循序渐进地介绍了相关的内容,包括线性分类器、核函数特征空间、推广性理论和优化理论,从而引出了支持向量机的算法,进而将支持向量机应用到实际的工程实例中。

本书共分为8章,第1章统计学习理论基础,第2章支持向量机基础,第3章支持向量机的分类、回归问题及应用,第4章应用背景及混合气体红外光谱分析基础,第5章基于SVM和红外光谱的含烃类混合气体分析方法,第6章含烃类混合气体分析方法的实际应用研究,第7章层次式SVM子集含烃类混合气体光谱分析框架研究,第8章石油天然气红外光谱分析系统的集成应用。

本书适合高等院校高年级本科生、研究生、教师和相关科研人员及相关领域的工作者使用。本书既可作为研究生教材,也可作为神经网络、机器学习、数据挖掘等课程的参考教材。

图书在版编目(CIP)数据

支持向量机理论及工程应用实例/白鹏等编著. —西安:西安电子科技大学出版社, 2008.8

ISBN 978-7-5606-2051-0

I. 支… II. 白… III. 向量计算机—算法理论 IV. TP301.6

中国版本图书馆CIP数据核字(2008)第072493号

策 划 云立实

责任编辑 许青青 云立实

出版发行 西安电子科技大学出版社(西安市太白南路2号)

电 话 (029)88242885 88201467 邮 编 710071

http://www.xduph.com E-mail: xdupfxb001@163.com

经 销 新华书店

印刷单位 陕西光大印务有限责任公司

版 次 2008年8月第1版 2008年8月第1次印刷

开 本 787毫米×960毫米 1/16 印张 10.375

字 数 239千字

印 数 1~4000册

定 价 16.00元

ISBN 978-7-5606-2051-0/TP·1060

XDUP 2343001-1

*** 如有印装问题可调换 ***

本社图书封面为激光防伪覆膜,谨防盗版。

前 言

支持向量机是在 20 世纪 90 年代由 Vapnik 等人研究并迅速发展起来的一种基于统计学习理论的机器学习算法。支持向量机以统计学习理论为理论体系,通过寻求结构风险最小化来实现实际风险的最小化,追求在有限信息的条件下得到最优结果。

以往困扰机器学习方法的很多问题,如模型选择与学习问题、非线性和维数灾难问题、局部极小问题等,通过支持向量机可得到一定程度上的解决。随着支持向量机理论的不断发展和成熟,加之神经网络等学习方法在理论上缺乏实质性进展,支持向量机开始受到越来越广泛的重视。

本书共分为 8 章,第 1 章统计学习理论基础,第 2 章支持向量机基础,第 3 章支持向量机的分类、回归问题及应用,第 4 章应用背景及混合气体红外光谱分析基础,第 5 章基于 SVM 和红外光谱的含烃类混合气体分析方法,第 6 章含烃类混合气体分析方法的实际应用研究,第 7 章层次式 SVM 子集含烃类混合气体光谱分析框架研究,第 8 章石油天然气红外光谱分析系统的集成应用。

通过本书内容的学习,读者可掌握统计学习的基本理论,学会用支持向量机理论处理信息的基本方法,了解支持向量机理论及应用的最新研究与进展,为开展科学研究打好基础。

诚然,SLT 理论和 SVM 方法处在发展阶段,很多方面尚不完善。例如,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。有关 SVM 算法的某些理论解释也并非完美。SVM 方法中如何根据具体问题选择适当的内积函数也没有理论依据。因此,在这方面我们可做的事情是很多的。

希望本书的出版能促进支持向量机在我国各个应用领域的普及,以期能给相关领域的理论研究者 and 应用工作者提供一些思路和帮助。

空军工程大学电路教研室的各位老师对本书的完成给予了大力支持和帮助,在此表示衷心感谢。

本书得以顺利出版还要感谢西安电子科技大学出版社,尤其要感谢云立实副编审的支持和帮助。

由于编著者水平有限,书中不当之处在所难免,敬请读者批评指正,不吝赐教。

编著者

2007 年 12 月于空军工程大学

欢迎选购西安电子科技大学出版社教材类图书

~~~~~国家级、部级重点教材~~~~~

计算机系统结构 (第四版) (李学干)	25.00
雷达对抗原理 (赵国庆)	15.00
雷达原理 (第三版) (丁鹭飞)	23.00
通信网的安全——理论与技术 (王育民)	42.00
模拟电子线路基础 (傅丰林)	16.00
移动通信 (第四版) (李建东)	30.00

~~~~~计算机提高普及类~~~~~

计算机应用基础 (第三版) (丁爱萍)	22.00
计算机文化基础 (高职) (游鑫)	27.00
计算机文化基础上机实训及案例 (高职)	15.00
计算机科学与技术导论 (吕辉)	22.00
计算机应用基础 (高职) (赵钢)	29.00
计算机应用基础——信息处理技术教程	31.00
计算机应用基础 (高职) (冉崇善)	
(Windows XP & Office 2003 版)	23.00

《计算机应用基础》实践技能训练	
与案例分析 (高职) (冉崇善)	18.00
办公自动化设备的使用和维护	
(第二版) (陈国先) (高职)	18.00

~~~~~计算机网络类~~~~~

计算机网络管理 (雷震甲)	20.00
网络安全技术 (高职) (廖兴)	19.00
网络安全技术 (高职) (李卓玲)	17.00
网络信息安全技术 (周明全)	17.00
动态网页设计实用教程 (蒋理)	30.00
Windows网络程序设计 (夏靖波)	26.00
局域网组建实例教程 (高职) (尹建璋)	20.00
WindowsServer 2003组网实训教程 (高职)	27.00
WindowsServer2003组网技术	
(含光盘) (高职) (陈伟达)	30.00
综合布线技术 (高职) (王趾成)	18.00
计算机网络应用基础 (武新华)	28.00
计算机网络 (张璟)	26.00
计算机网络 (第二版) (袁家政) (高职)	26.00

计算机网络 (第三版) (蔡皖东)	27.00
计算机组网实验教程 (王宣政)	23.00
计算机网络基础及应用 (高职) (向隅)	22.00

~~~~~计算机技术类~~~~~

计算机应用基础教程 (第四版) (陈建铎)	
(for Windows XP/Office XP)	30.00
计算机系统结构 (陈智勇)	26.00
计算机系统结构与组成 (吕辉)	26.00
计算机系统安全 (第二版) (马建峰)	30.00
计算机信息检索实用教程 (高职) (张晓芳)	15.00
电子商务概论 (宋沛军) (高职)	20.00
电子商务基础与应用 (第五版) (含盘)	39.00
《电子商务基础与应用 (第五版)》	
学习指导与练习 (杨坚争)	20.00
电子商务基础与实务 (第二版) (高职)	16.00
数据结构——使用 C++语言 (第二版) (朱战立)	23.00
数据结构 (高职) (周岳山)	15.00
数据结构教程——Java 语言描述 (朱振元)	29.00
计算方法与实习 (高职) (田祥宏)	11.00
算法设计与分析 (霍红卫)	15.00
编译原理教程 (第二版) (胡元义)	18.00
离散数学 (马光思)	22.00
离散数学 (武波)	24.00
软件工程 (第二版) (邓良松)	22.00
软件技术基础 (高职) (鲍有文)	23.00
信息系统分析与设计 (高职) (卫红春)	20.00
信息系统分析与设计 (第二版) (陈圣国)	15.00
人工智能技术导论 (第三版) (廉师友)	24.00
~~~~~计算机辅助技术类~~~~~	
电子工程制图 (含习题集) (高职) (童幸生)	25.00
机械制图与计算机绘图 (含习题集) (高职)	40.00
计算机绘图 (第二版) (许社教)	25.00
DSP应用技术 (高职) (赵明忠)	25.00
电子线路 CAD 实用教程 (潘永雄) (第三版)	27.00
AutoCAD 基础教程 (高职) (石高峰)	15.00

电子工艺与电子 CAD (高职) (朱旭平)	14.00	~~~~~数据库及计算机语言类~~~~~	
EDA 技术及应用 (第二版) (谭会生)	27.00	数据库原理 (第二版) (郭盈发)	16.00
数字电路 EDA 设计 (高职) (顾斌)	19.00	C 程序设计与实例教程 (曾令明)	21.00
~~~~~操作系统类~~~~~		程序设计与 C 语言 (第二版) (马鸣远)	32.00
计算机操作系统 (第二版) (颜彬) (高职)	19.00	C 语言程序设计教程 (高职) (朱接文)	20.00
计算机操作系统 (修订版) (汤子瀛)	24.00	C 语言程序设计课程与考试辅导 (王晓丹)	25.00
计算机操作系统 (第三版) (汤小丹)	27.00	Visual Basic.NET 程序设计 (高职) (马宏锋)	24.00
计算机操作系统 (方敏)	28.00	Visual FoxPro 数据库程序设计教程 (康贤)	24.00
Linux 网络操作系统应用教程 (高职) (王和平)	25.00	Visual FoxPro 6.0 数据库原理与应用 (高职)	21.00
Linux 操作系统实用教程 (高职) (梁广民)	20.00	SQL Server 2000 应用基础与实训教程 (高职)	22.00
~~~~~图形处理类~~~~~		网络数据库技术及应用 (高职) (范剑波)	24.00
多媒体软件开发 (高职) (含盘) (牟奇春)	35.00	C++ 程序设计语言 (揣锦华)	20.00
多媒体技术及应用 (第二版) (高职) (王坤)	25.00	Visual C++ 基础教程 (郭文平)	29.00
多媒体软件设计技术 (第二版) (陈启安)	20.00	汇编语言程序设计 (第二版) (韩海)	18.00
多媒体技术基础与应用 (曾广雄)	20.00	面向对象程序设计与 VC++ 实践 (揣锦华)	22.00
计算机图形学 (丁爱玲)	14.00	面向对象程序设计与 C++ 语言 (第二版)	18.00
计算机图形学 (研究生系列) (璩柏青)	26.00	面向对象程序设计——JAVA (第二版)	32.00
数字图像处理 (何东建)	23.00	跨平台程序设计语言——JAVA (赵毅)	24.00
三维动画案例教程 (含光盘) (高职)	25.00	Java 语言程序设计教程 (张席)	18.00
~~~~~微机与控制类~~~~~		Java 程序设计 (高职) (陈圣国)	18.00
微型计算机原理 (第二版) (王忠民)	27.00	Java 程序设计教程 (曾令明)	23.00
微型计算机原理及接口技术 (第二版)	36.00	JavaWeb 程序设计基础教程 (高职) (李绪成)	25.00
微机接口技术及其应用 (李育贤)	19.00	Access 数据库应用技术 (高职) (王趾成)	21.00
微型计算机组成与接口技术		Web 应用开发技术: JSP (含光盘)	33.00
(第二版) (高职) (赵佩华)	28.00	~~~~~电子及通信类~~~~~	
单片机应用技术 (第二版) (高职) (刘守义)	30.00	测试与计量技术基础 (周渭)	19.00
单片机原理与应用技术 (黄惟公)	22.00	光的电磁理论 (石顺祥)	28.00
单片机原理与程序设计实验教程 (于殿泓)	18.00	传感器原理及工程应用 (第二版)	20.00
单片机实验与实训指导 (高职) (王曙霞)	19.00	传感器应用技术 (高职) (王煜东)	27.00
单片机原理及接口技术 (第二版) (余锡存)	19.00	传感器原理及应用 (郭爱芳)	24.00
单片机原理与应用技术 (喻宗泉)	19.00	电路 (高赞)	26.00
新编单片机原理与应用 (第二版) (潘永雄)	22.00	电路分析基础 (第三版) (张永瑞)	28.00
MCS-51 单片机原理及嵌入式系统应用	26.00	电路基础 (高职) (孔凡东)	13.00
计算机控制技术 (高职) (温希东)	12.00	模拟电子技术 (高职) (钱聪)	13.00
微机外围设备的使用与维护 (高职) (王伟)	19.00	模拟电子技术及应用 (高职) (王朱芳)	21.00
微机装配调试与维护教程 (王忠民)	25.00	模拟电子技术 (第二版) (教育部高职) (周雪)	17.00
《微机装配调试与维护教程》实训指导	22.00	模拟电子技术 (第二版) (江晓安)	21.00
		数字电子技术 (高职) (蒋卓勤)	15.00

机械工程科技英语(程安宁)	15.00	模具制造技术(高职)(刘航)	22.00
机械设计基础(张京辉)(高职)	24.00	模具设计(高职)(曾霞文)	18.00
机械基础(安美玲)(高职)	20.00	冷冲压模具设计(高职)(刘庚武)	21.00
机械 CAD/CAM(欧长劲)	21.00	塑料成型模具设计(高职)(单小根)	37.00
机械 CAD/CAM 上机指导及练习教程	20.00	液压与气动技术(第二版)(朱梅)	23.00
机械制图(含习题集)(高职)(孙建东)	29.00	液压传动技术(高职)(简引霞)	23.00
机械设备制造技术(高职)(柳青松)	33.00	发动机构造与维修(高职)(王正键)	29.00
机械制造工艺装备(高职)(朱派龙)	22.00	汽车典型电控系统结构与维修(李美娟)	31.00
机械制造技术(高职)(邵堃)	24.00	汽车机械基础(高职)(姜万军)	29.00
机械制造基础(高职)(郑广花)	21.00	汽车底盘结构与维修(高职)(张红伟)	28.00
机械加工技术(高职)(魏康民)	24.00	汽车车身电气设备系统及附属电气设备	
数控编程与操作--SINUMERIK 数控系统	15.00	(高职)(颜培钦)	23.00
数控加工与编程(第二版)(高职)(詹华西)	22.00	汽车单片机与车载网络技术(于万海)	20.00
数控加工工艺(高职)(赵长旭)	24.00	汽车故障诊断技术(高职)(王秀贞)	19.00
数控加工工艺课程设计指导书(赵长旭)	12.00	汽车及配件营销(边伟)(高职)	25.00
数控加工工艺(高职)(刘长伟)	22.00	汽车营销技术(高职)(孙华宪)	15.00
数控加工实训教程(高职)(胡相斌)	24.00	汽车使用性能与检测技术(高职)(郭彬)	22.00
数控加工编程与操作(高职)(刘虹)	15.00	汽车电工电子技术(高职)(黄建华)	20.00
数控编程与操作(高职)(周虹)	19.00	汽车电气设备与维修(高职)(李春明)	25.00
数控机床故障分析与维修(高职)(潘海丽)	19.00	汽车车身结构与维修(高职)(吴兴敏)	26.00
数控机床电气控制(高职)(姚勇刚)	21.00	汽车使用与技术管理(高职)(边伟)	25.00
现代数控机床(高职)(刘瑞己)	25.00	汽车空调(高职)(李祥峰)	16.00
数控应用专业英语(高职)(黄海)	17.00	汽车概论(高职)(邓书涛)	20.00
机床电器与 PLC(高职)(李伟)	14.00	现代汽车典型电控系统结构原理	
电机及拖动基础(高职)(孟宪芳)	17.00	与故障诊断(高职)(徐生明)	25.00
电机拖动与控制(高职)(刘保录)	25.00	有色金属材料焊割技术(高职)(袁国义)	24.00
电机与电气控制(高职)(冉文)	23.00	互换性与技术测量(高职)(屈波)	19.00
电机原理与维修(高职)(解建军)	20.00	互换性与几何量测量技术(张帆)	30.00
供配电技术(高职)(杨洋)	25.00	电子CAD(Protel 99 SE)实训指导书(高职)	12.00
电切削加工技术(高职)(詹华西)	13.00	Pro/ENGINEER应用教程(高职)(朱玉红)	23.00
金属切削与机床(高职)(聂建武)	22.00	安全评价技术(张乃禄)	23.00
模具 CAD/CAM 实用教程——		安全检测技术(张乃禄)	28.00
Pro/ENGINEER Wildfire2.0(高职)	18.00		

欢迎来函索取本社最新书目和教材介绍, 欢迎投稿!

通信地址: 西安市太白南路2号 西安电子科技大学出版社发行部 邮 编: 710071
 邮购业务电话: (029) 88201467 传 真: (029) 88213675

目 录

第 1 章 统计学习理论基础	1
1.1 机器学习	1
1.1.1 基于数据的机器学习问题模型	2
1.1.2 经验风险最小化	5
1.2 统计学习理论的产生	5
1.3 学习过程的一致性条件	7
1.4 VC 维理论	8
1.5 推广性的界	9
1.6 结构风险最小化	10
1.6.1 结构风险最小化的含义	10
1.6.2 SRM 原则的实现	11
第 2 章 支持向量机基础	13
2.1 SVM 的基本思想	13
2.1.1 最优分类面	14
2.1.2 广义的最优分类面	15
2.2 核函数	17
2.3 支持向量机的经典算法剖析	20
2.3.1 算法概述	21
2.3.2 SMO 算法	23
2.3.3 C-SVM 算法及其变形算法	28
2.3.4 ν -SVM 算法	30
2.4 SMO 算法的自适应学习改进方法	32
2.4.1 SMO 算法与二次规划	33
2.4.2 自适应学习方法	33
2.4.3 实验结果	34
2.5 LibSVM 软件	36
2.5.1 LibSVM 软件简介	36
2.5.2 LibSVM 软件的使用方法	37
第 3 章 支持向量机的分类、回归问题及应用	41
3.1 分类问题的提出及 SVM 分类原理	41
3.1.1 线性分类	43
3.1.2 近似线性分类	44
3.1.3 非线性分类	45

3.2 多类分类问题	46
3.2.1 多类分类原理	47
3.2.2 经典多类分类算法简介	48
3.3 SVM 回归原理	53
3.3.1 SVM 回归问题的描述	53
3.3.2 线性支持向量机回归	55
3.3.3 非线性支持向量机回归	56
3.4 基于关联规则的 SVM 在线分类方法	57
3.4.1 基本思路	58
3.4.2 混合气体分布模式中心集的生成	59
3.4.3 分布模式 SVM 的在线学习	60
3.5 压力传感器支持向量机的校正方法	62
3.5.1 支持向量机与压力传感器的特性校正原理	62
3.5.2 实验及分析	64
3.6 支持向量机回归用于分类	66
3.6.1 思路及推导证明	66
3.6.2 应用实例	67
3.7 基于支持向量机的设备备件供应研究	69
3.7.1 设备备件供应保障概述	69
3.7.2 基于 LS-SVM 的设备备件多元分类	70
3.7.3 基于 SVM 的设备备件需求预测	73
第 4 章 应用背景及混合气体红外光谱分析基础	77
4.1 研究背景与意义	77
4.2 气测录井中混合气体分析的发展、现状及存在的问题	79
4.2.1 气相色谱分析	79
4.2.2 红外光谱分析	82
4.3 支持向量机与红外光谱分析结合	87
4.3.1 问题的提出	87
4.3.2 研究方法	88
4.3.3 技术路线	88
4.4 红外光谱分析理论基础	88
4.4.1 Lambert-Beer 吸收定律	89
4.4.2 红外光谱理论	90
4.4.3 傅立叶变换红外光谱仪	92
4.5 红外光谱数据的描述及分析方法	95
4.5.1 红外光谱数据的描述	95
4.5.2 红外光谱分析方法	96
4.6 红外光谱分析	97
4.6.1 红外光谱的定性分析	97

4.6.2 红外光谱的定量分析	100
第5章 基于SVM和红外光谱的含烃类混合气体分析方法	105
5.1 分析方法的基本工作模式及处理流程	105
5.1.1 分析思路	106
5.1.2 含烃类混合气体组分浓度的SVM分析原理	107
5.1.3 实验装置	108
5.1.4 处理流程	108
5.2 含烃类混合气体的基本工作模式分析	110
5.2.1 混合气体样本的建立	110
5.2.2 光谱数据预处理	112
5.2.3 SVM校正模型的训练与检验	112
5.2.4 SVM校正模型的实现	113
第6章 含烃类混合气体分析方法的实际应用研究	115
6.1 含烃类混合气体分析方法的参数选择	115
6.1.1 SVM校正模型参数对分析结果的影响	115
6.1.2 光谱数据预处理——归一扩展法	117
6.1.3 光谱分析波段对结果的影响	119
6.1.4 红外光谱仪扫描间隔对结果的影响	120
6.1.5 参数选择结果	121
6.2 二氧化碳气体浓度分析	122
6.2.1 混合气体样本的建立	122
6.2.2 光谱数据预处理	124
6.2.3 SVM校正模型的训练与检验	124
6.2.4 SVM校正模型的实现	124
6.3 含烃类混合气体组分浓度分析	125
6.3.1 单组分气体实验	125
6.3.2 两组分气体实验	127
6.3.3 多组分气体实验	128
6.4 含烃类混合气体组分种类分析	129
6.4.1 SVM校正模型用于混合气体组分种类识别的原理	129
6.4.2 实际应用	130
第7章 层次式SVM子集含烃类混合气体光谱分析框架研究	131
7.1 层次式SVM子集混合气体分析框架	131
7.1.1 基本概念	131
7.1.2 层次式SVM子集原理	132
7.1.3 实验结果	134
7.2 层次式SVM子集含烃类混合气体的光谱分析集成框架	135
7.2.1 分析集成框架的结构	135
7.2.2 混合气体分布子模式的建立	136

7.2.3 SVM模型的建立及优化	138
7.2.4 分析结果	140
第8章 石油天然气红外光谱分析系统的集成应用	141
8.1 实验系统	141
8.1.1 配气系统	141
8.1.2 红外光谱数据样本制作系统	144
8.1.3 SVM校正模型	148
8.2 系统集成框架	148
8.2.1 系统总体框架	149
8.2.2 系统硬件	149
8.2.3 系统软件	150
8.2.4 系统功能	153
8.3 现场应用	153
8.3.1 性能指标评估	153
8.3.2 运行状况	155
参考文献	156

第1章 统计学习理论基础

机器学习(Machine Learning, ML)是人工智能(Artificial Intelligence, AI)最具智能特征、最前沿的研究领域之一。基于数据的统计学习问题是现代智能计算技术的一个重要分支。基于数据的统计学习不同于传统的以渐进理论为基础的统计学,它模拟人类从实例中学习归纳的能力,主要研究如何从一些观测数据中挖掘出目前尚不能通过原理分析得到的规律,并利用这些规律去分析客观对象,对未知数据或无法观测的新数据进行预测和判断。

本章对机器学习、统计学习理论的产生、学习过程的一致性条件、VC维理论、推广性的界、结构风险最小化等统计学习理论的基本内容进行了概述,为后续章节的学习提供理论基础。

1.1 机器学习

学习能力是学习行为的一个非常重要的特征,但至今人类对学习的机理尚不清楚。研究人员曾对学习给出过各种定义。

H. A. Simon 认为:学习是系统所作的适应性变化,以使得系统在下次完成同样或类似的任务时更为有效。

R. S. Michalski 认为,学习是构造或修改对于所经历事物的表示。

从事具体智能系统研制的学者则认为学习是知识的获取。

上述观点各有侧重,第一种观点强调学习的外部行为效果,第二种则强调学习的内部过程。

机器学习主要研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构,使之不断改善自身的性能。机器学习就是根据生理学、认知科学等对人类学习机理的了解,建立人类学习过程的计算模型或认识模型,发展各种学习理论和学习方法,研究通用的学习算法并进行理论上的分析,建立面向任务的具有特定应用的学习系统。机器学习是人工智能的核心问题,是使计算机具有人工智能的根本途径。机器学习的应用遍及人工智能的各个领域。机器学习主要使用归纳、综合,而不是演绎。

机器学习在人工智能的研究中占有十分重要的地位。一个不具有学习能力的智能系统难以称得上是一个真正的智能系统,而以往的智能系统普遍都缺少学习的能力。例如,遇到错误时不能自我校正,不会通过经验改善自身的性能,不会自动获取和发现所需要的知识。它们的推理仅限于演绎而缺少归纳,因此至多只能证明已存在的事实、定理,而不能发现新的定理、定律和规则等。

随着人工智能的深入发展,这些局限性表现得更加突出。正是在这种情形下,机器学习逐渐成为人工智能研究的核心之一。机器学习的应用已遍及人工智能的各个分支,如专家系统、自动推理、自然语言理解、模式识别、计算机视觉、智能机器人等领域。其中,尤其典型的是专家系统中知识获取瓶颈的问题,研究人员一直在努力试图采用机器学习的方法加以克服。

机器学习的目标就是尽可能地让机器能具有类似于人类学习知识的能力。由于机器本身具有一些特性,因而机器学习也具有其自身的一些特点。

(1) 机器学习结果的正确性往往不能保证。机器学习一般采用归纳推理进行概括,而归纳推理总能得出多个结论。虽然机器学习中的归纳推理也使用了各种启发信息、假设和约束,经过提出假设,再经实践验证等多个过程,但是绝对严格地验证各种假设是非常困难的,故学习结果的正确性不能得到保证。

(2) 机器一旦具有学习能力,则其学习的效率是惊人的。人类学习的速度是非常缓慢的,往往需要十几年的时间,而机器学习的速度则很快。机器学习到有关知识后就能马上应用。同人类知识的传播极其困难不同,机器中知识的传播是非常方便的。

(3) 机器学习也具有一定的风险。因为学习的特点就是获取前所未有的新知识,机器学习在获得巨大成效的同时,也存在着超出人类期望和控制的潜在危险,所以轻易使用机器学习解决重大问题有可能导致意外的困难和危险。

1.1.1 基于数据的机器学习问题模型

基于数据的机器学习问题作为人工智能研究领域的一个重要方面,其研究的主要问题是从小组观测数据集的数据出发,得到一些不能通过原理分析而得到的规律,进而利用这些规律对未来的数据或者无法观测到的数据进行预测和分析。

将这种研究方法应用于抽象的分类问题时,这类机器学习的问题就是模式识别问题。从广义的角度看,传统的基于统计学的模式识别问题可以视为基于数据的机器学习问题的一个特例。

基于数据的机器学习的目标是利用给定有限数量的训练数据对某系统输入/输出之间的依赖关系进行估计,根据输入/输出之间的依赖关系对未知输出做出尽可能准确的预测。基于数据的机器学习问题的基本模型可以用图 1-1 表示。

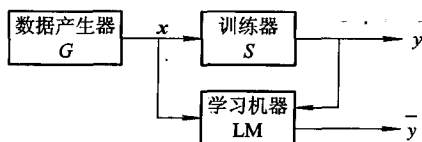


图 1-1 基于数据的机器学习问题的基本模型

基于数据的机器学习模型一般由数据产生器 G 、训练器 S 、学习机器 LM 三部分组成,即 $M = \langle G, S, LM \rangle$ 。

G 是数据产生器,能从固定但未知概率的分布函数 $F(x)$ 中独立产生随机向量 $x, x \in \mathbf{R}^n$ 。

S 是训练器,对每个输入向量 x 返回一个确定的输出 y ,输出根据的是同样固定但未知条件的分布函数 $F(y|x)$,根据联合分布的条件有: $F(x, y) = F(x)F(y|x)$ 。从其中抽取出 l 个独立分布的观察数据 $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$,便构成了训练数据集。

LM 是学习机器,能在一组函数集 $f(x, \alpha) (\alpha \in \Lambda, \Lambda$ 为参数集合)中选择出使输出 y 最好地逼近训练响应 y 的函数 $f(x, \alpha)$ 。训练之后的学习机器必须对任意输入 x 给出输出 $\bar{y}$ 。最小化风险泛函 $R(\alpha)$ 为

$$R(\alpha) = \int L(y, f(x, \alpha)) dF(x, y) \quad (1.1-1)$$

其中: $f(x, \alpha)$ 称为学习函数集或预测函数集; α 为函数的广义参数, $\alpha \in \Lambda$; Λ 是参数集合。 $L(y, f(x, \alpha))$ 为在给定输入 x 下,训练器 S 输出 y 与学习机器 LM 给出的 $f(x, \alpha)$ 之间损失的期望。通过选择不同形式的损失函数可以构成不同类型的学习机器。

图 1-1 给出了基于数据的机器学习问题的一个基本模型,这个模型的结构和表述是非常广义的,涵盖了很多特殊的问题。通常来说,这个模型主要包括了以下三类典型问题:模式识别、回归估计和概率密度估计。

1. 模式识别

对于简单的一类分类问题,图 1-1 模型的输出 y 就转化成模式分类器的输出标号。对于两类分类问题,系统输出向量 y 可以简单表示成形式为 $y = \{0, 1\}$ 的二值函数。预测函数集在这里称为指示函数集,即模式识别问题中的判别函数。在模式识别问题中的损失函数可以定义成如下形式:

$$L(y, f(x, \alpha)) = \begin{cases} 0 & y = f(x, \alpha) \\ 1 & y \neq f(x, \alpha) \end{cases} \quad (1.1-2)$$

对应这个损失函数, 式(1.1-1)的泛函确定了训练器 S 和指示函数 $f(x, \alpha)$ 所给出的答案不同的概率, 这在模式识别中通常称为平均错误率。使平均错误率最小的模式识别方法就是贝叶斯决策, 模式识别中选用不同的损失函数代表了不同的决策方法。

2. 回归估计

令图 1-1 中的输出 y 为实数值, 并令 $f(x, \alpha)$ 为实函数集合, $\alpha \in \Lambda$ 。其中, 包含的回归函数为

$$f(x, \alpha_0) = \int y dF(y | x) \quad (1.1-3)$$

这里的回归函数就是在损失函数为

$$L(y, f(x, \alpha)) = (y - f(x, \alpha))^2 \quad (1.1-4)$$

的情况下, 使式(1.1-1)最小化风险泛函的函数。

因此, 回归估计问题可以表示成在概率测度 $F(x, y)$ 未知, 但数据集 $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$ 已知的情况下, 使采用式(1.1-4)的损失函数的风险泛函式(1.1-1)最小化。

实际上, 如果能够将函数的输出 y 通过一个值域转化为一个二值函数, 则回归估计的问题也就变成了一个模式识别问题。

3. 概率密度估计

概率密度估计问题就是根据数据集, 考虑从密度函数集 $p(x, \alpha)$, $\alpha \in \Lambda$ 中确定 x 的概率分布。对于这个问题, 考虑下面的损失函数:

$$L(p(x, \alpha)) = -\ln p(x, \alpha) \quad (1.1-5)$$

待求的密度函数在损失函数式(1.1-5)下使风险泛函式(1.1-1)最小化。因此, 根据数据集估计概率密度函数的问题就是, 在相应的概率测度 $F(x)$ 未知, 但给出了独立同分布数据集 $x_1, x_2, \dots, x_n$ 的情况下, 使风险泛函式(1.1-1)最小化。

更一般地, 基于数据的机器学习问题可以定义为如下问题:

设定义在空间 Z 上的概率为 $F(z)$, 考虑函数的集合 $Q(z, \alpha)$, $\alpha \in \Lambda$, 基于数据的机器学习的目标是最小化风险泛函:

$$R(\alpha) = \int Q(z, \alpha) dF(z), \alpha \in \Lambda \quad (1.1-6)$$

其中, $Q(z, \alpha)$ 就是待定的损失函数, 概率 $F(z)$ 未知, 但给定了独立同分布数据集:

$$z_1, z_2, \dots, z_l \quad (1.1-7)$$

z 代表数据 (x, y) 。

1.1.2 经验风险最小化

一般基于数据的机器学习问题的目标在于使期望风险 $R(\alpha)$ 达到最小化。但是, 由于基于数据的机器学习问题已知的全部信息只有数据集的期望风险, 且无法计算, 因此, 传统的方法采用了经验风险最小化(Empirical Risk Minimization, ERM)原则, 即用式(1.1-7)的数据集定义经验风险泛函:

$$R_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \quad (1.1-8)$$

在回归估计中, 广泛采用的最小二乘法 and 密度估计的最大似然法是 ERM 原则的一个特例。对于模式识别损失函数, 经验风险最小化就是训练数据集的错误率; 对于函数逼近的损失函数, 经验风险就是最小二乘法; 采用概率密度估计的损失函数的 ERM 原则就等价于最大似然方法。

仔细研究经验风险最小化原则和基于数据的机器学习问题中期望风险最小化要求可以发现, 从期望风险最小化到经验风险最小化并没有可靠的理论依据, 只是直观上合理的、想当然的做法。

首先, $R_{\text{emp}}(\alpha)$ 和 $R(\alpha)$ 都是 α 的函数, 概率论中大数定理只说明了在一定条件下, 当数据集趋于无穷多时 $R_{\text{emp}}(\alpha)$ 将在概率意义上趋近于 $R(\alpha)$, 并没有保证使 $R_{\text{emp}}(\alpha)$ 最小的 α^* 与使 $R(\alpha)$ 最小的 α' 是同一个点, 更不能保证 $R_{\text{emp}}(\alpha^*)$ 能够趋近于 $R(\alpha')$ 。

其次, 即使有办法使这些条件在数据集数目无穷大时得到保证, 也无法认定在这些前提下得到的经验风险最小化方法在数据集数目有限(即小样本数据)的情况下仍能得到好的结果。

尽管有上述的未知问题, 但经验风险最小化原则作为解决模式识别等基于数据的机器学习问题的基本思想统治了这一领域几乎所有的研究。多年来, 研究人员一直将大部分注意力集中到如何更好地求取最小经验风险。

1.2 统计学习理论的产生

统计学方法是面对大量的观测数据而又缺乏合适的理论模型时唯一的分析手段。统计学方法研究的是在数据集的数目趋于无穷大时的极限特性。统计学中关于估计一致性、无偏性、估计方法的界以及分类错误率的诸多结论均以这种极限特性为基础。

传统的统计模式识别问题是在数据集的数目足够大的前提下进行研究的, 在此前提下提出的各种有效方法也只有当数据集数目趋于无穷大时, 在

理论上才有保证。然而,在现实应用中,数据集的数目通常是有限的,有时数据集样本的获取甚至是非常困难的。数据集的数目趋于无穷大或者数据集的数据样本量大这个最基本的前提,在实际应用中往往得不到满足,特别是当所研究的问题处于一个高维的特征空间中时,这一矛盾显得尤为突出。

例如,有一混合气体由 7 种气体组成,每种气体浓度分别为 $100\% \sim 0\%$ 、 $80\% \sim 0\%$ 、 $60\% \sim 0\%$ 、 $40\% \sim 0\%$ 、 $30\% \sim 0\%$ 、 $30\% \sim 0\%$ 、 $30\% \sim 0\%$ 。如果每种气体浓度按 1% 的间隔标定,则 7 组分混合气体需 5.184×10^{11} 个数据样本。若要实现在较大分布范围内的混合气体组分浓度分析,则必须构造数量足够多且分布合理的混合气体数据样本,以供训练使用。但在实际工作中,构造数量足够多且分布合理、组分浓度范围大的混合气体数据集是不容易实现的,甚至无法获得如此海量的混合气体数据集。

在实际应用中,包括人工神经网络方法(Artificial Neural Networks, ANN)、贝叶斯理论等众多基于数据的机器学习理论和方法都面临着这一困难。因此,研究在小样本数据下的统计学习规律是一个非常具有实用价值的问题。

Vapnik 等人提出了统计学习理论(Statistical Learning Theory, SLT),他们早在 20 世纪 60 年代中期就开始研究在有限数据集的情况下基于数据的机器学习问题。由于当时的研究尚不完善,在模式识别问题中取得的结果往往趋于保守,加之该理论在数学上比较艰涩,因此直到 20 世纪 90 年代初期都没有提出能将该理论付诸实践的应用方法。

20 世纪 90 年代中期,统计学习理论才逐渐发展和成熟,并形成了一个比较完善的统计学习理论体系。

在同一时期,机器学习领域其他比较流行的方法(如人工神经网络等),则遇到了另外一些问题的困扰,例如,网络模型结构的确定问题、“过学习”和“欠学习”的问题以及局部最小点的问题等。这些问题的出现在一定程度上给统计学习理论在基于数据的机器学习领域中重新受到关注提供了机会。

统计学习理论从理论上系统地研究了经验风险最小化原则成立的条件,有限数据集下经验风险与期望风险的关系及如何利用这些理论找到新的学习原则和方法等问题。其主要内容包括以下 4 个方面:

- (1) 经验风险最小化原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数据样本归纳推理原则;
- (4) 实现这些新原则的实际方法或算法。

由于统计学习理论是一种专门研究在有限数据集的情况下基于数据的机器学习规律的理论,因此它为研究在有限数据集情况下的统计模式识别和更广泛的机器学习问题构建了一个较好的理论基础。同时,也发展了一种新的