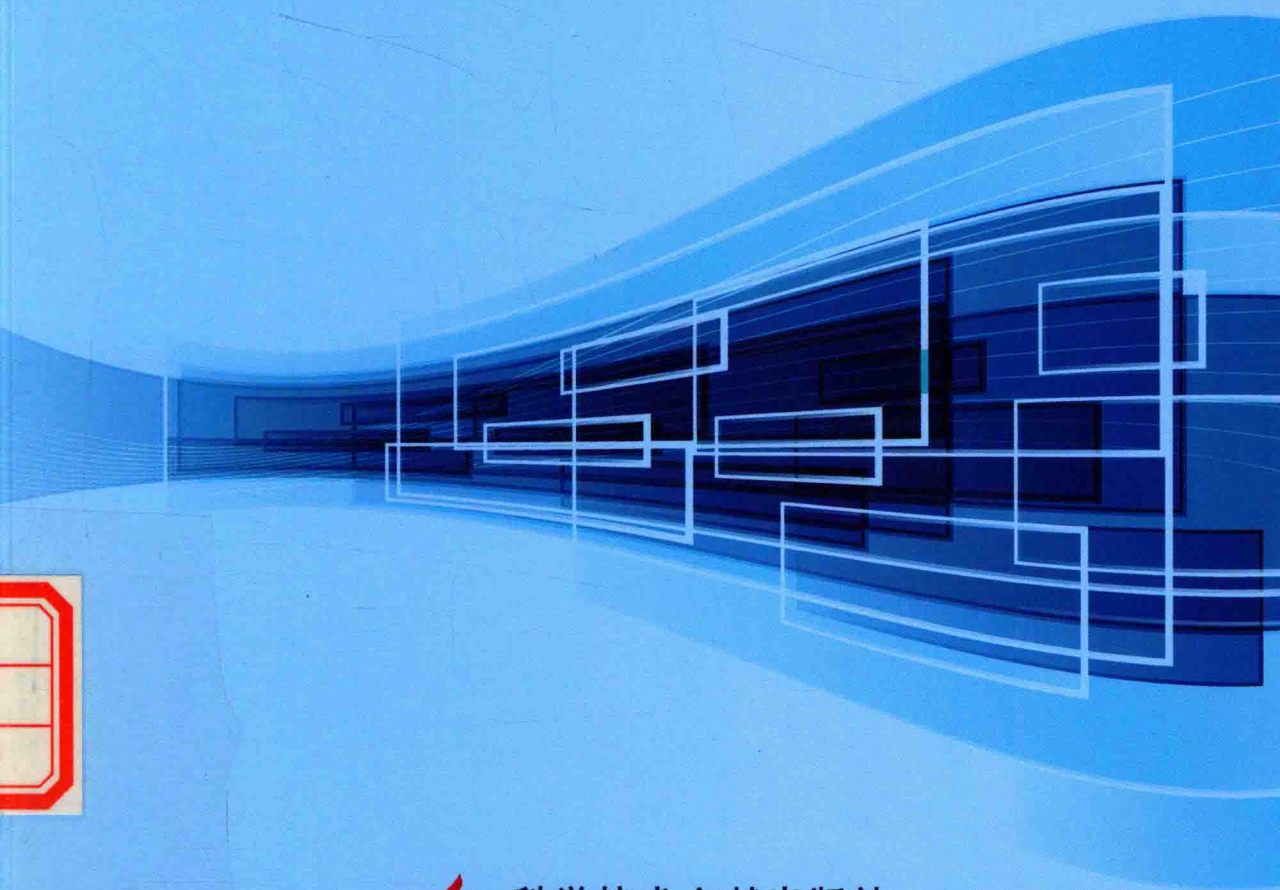
基于模糊信息的应用技术研究

Research on Application Technology Based on Fuzzy Information

王爱民　葛彦强　周宏宇◎著



科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

基于模糊信息的应用技术研究

王爱民　葛彦强　周宏宇◎著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS
·北京·

图书在版编目（CIP）数据

基于模糊信息的应用技术研究 / 王爱民，葛彦强，周宏宇著. —北京：科学技术文献出版社，2016.10
ISBN 978-7-5189-2044-0

Ⅰ.①基… Ⅱ.①王… ②葛… ③周… Ⅲ.①数据处理—研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2016）第 249073 号

基于模糊信息的应用技术研究

策划编辑：崔灵菲　　责任编辑：王瑞瑞　　责任校对：赵　瑗　　责任出版：张志平

出 版 者　科学技术文献出版社
地　　址　北京市复兴路15号　　邮编　100038
编 务 部　（010）58882938，58882087（传真）
发 行 部　（010）58882868，58882874（传真）
邮 购 部　（010）58882873
官方网址　www.stdp.com.cn
发 行 者　科学技术文献出版社发行　全国各地新华书店经销
印 刷 者　北京教图印刷有限公司
版　　次　2016 年 10 月第 1 版　2016 年 10 月第 1 次印刷
开　　本　787×1092　1/16
字　　数　685千
印　　张　28.25
书　　号　ISBN 978-7-5189-2044-0
定　　价　138.00元

前　言

随着信息时代的到来，人类面临着越来越多的数据存储、组织和检索等信息处理问题。这些问题的特点表现在层次结构上越来越复杂，空间活动的规模上越来越大，时间尺度上越来越快，后果和影响上越来越广泛和深远。解决这类问题的关键是基于数据的智能（机器学习）技术。研究从已知数据集中挖掘出各种数据模型、寻找规律，利用这些学来的“知识”对未来数据或无法观测的数据进行预处理。

统计学习理论（SLT）是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。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（Support Vector Machine），它能够较好地处理分类和回归问题。该技术已成为智能化数据处理领域的研究热点，并在很多领域得到了成功的应用。但是，作为一种尚未成熟的新技术，支持向量机目前还存在局限。例如，客观世界存在大量的模糊信息，如果支持向量机的训练集中含有模糊信息，那么支持向量机将无能为力。

模糊数学是处理模糊信息的有力工具。模糊聚类分析、模糊模式识别是模糊数学中处理模糊分类问题的有效方法，但模糊模式识别方法与含有模糊信息的支持向量机方法二者的已知条件和所解决的问题都是不一样的。因此不能简单地用模糊模式识别方法解决含有模糊信息的支持向量机所要解决的问题。特别是，由于现实问题的复杂性，人们还需要在实际应用的求解过程中对现有支持向量机进行不断完善，更多的智能化数据分析（预测）还需要多类算法的组合应用，企望达到更好的应用效果。本书的研究主要从训练集中含有模糊信息的支持向量机分类算法和具有复杂信息的数据挖掘算法两个方面展开。研究成果在实际问题中得到了应用。

本书在理论上的研究主要是：阐述了数据挖掘的研究现状，阐述了支持向量机分类和回归算法，并且分析了支持向量机的理论基础——统计学习理论。引入模糊系数规划的模型和解法，给出了分类问题中的模糊信息表示方法。将模糊分类问题转化为求解模糊系数规划问题。分别建立了 Fuzzy 线性可分问题、Fuzzy 广义线性可分问题和 Fuzzy 非线性问题的支持向量分类机（算法）。研究了构建模糊支持向量分类机时，模糊系数规划中最佳阈值的确定方法。对 LS-SVM 遗传算法中的参数优化问题进行了研究，对非线性支持向量机回归算法进行

了改进。以 Fuzzy 聚类、Fuzzy 规划、支持向量机回归算法为关键技术，设计了烧结矿化学成分预测辅助系统，并取得了理想的预测效果。对“最短距离聚类算法”进行了改进，提出了邻近聚类算法，使其适用于任意形状的聚类。提出并证明了 2 个可以快速搜索到最近邻点的搜索定理。根据搜索定理提出了相应的搜索算法：“基于距离的聚类”算法、“基于阈值的邻近聚类”算法、“邻近聚类”算法、“多重聚类”算法。对以上算法用实验数据进行了验证，实验结果表明该算法是有效的，可以快速产生不同层次的高质量聚类。基于仿生计算技术，分析了基因表达式编程（Gene Expression Programming，GEP）中采用的随机初始化策略，虽然实现简单，占用计算资源少，但产生的初始种群基因多样性有限。在 GEP 算法中，随机产生初始种群策略有时还会产生“最高适应度为负”的种群，从而导致进化很难开始。提出了精英个体产生方案和基因空间均匀分布初始种群产生方案。前者可以产生适应度较高的个体，后者可以提高初始种群基因多样性，并且产生的初始种群的最高适应度一般为正。证明了基因表达式编程译码空间定理。通过 GEP 模拟 3 个标准函数的挖掘过程，结果表明：优化后的基因表达式编程方案可以大大提高进化成功率。

本书在技术应用上的研究主要是：开发了“中医症状鉴别智能诊断系统”。在理论与实例的结合上，具体设计了图像分割与影像快速融合、安全设计与信息检索优化方法、优化设计与决策支持、智能控制技术、第四方物流多属性指派决策机制等多个研究领域的问题求解方案。

目　录

第1章 引 论

1.1 研究背景

随着信息时代的到来，人类面临着越来越多的数据存储、组织和检索等信息处理问题。这些问题的特点表现在层次结构上越来越复杂，空间活动的规模上越来越大，时间尺度上越来越快，后果和影响上越来越广泛和深远。

解决这类问题的关键是基于数据处理的智能（机器学习）技术。它旨在从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识。研究从已知数据集中发现各种模型、导出值的过程、寻找规律，利用这些学来的“知识”对未来数据或无法观测的数据进行预处理。它是一个多领域的交叉学科，涉及数据库技术、统计学、神经网络、知识工程、高性能计算等诸多领域，并在工、农、商、经、医等众多行业得到了广泛的应用。

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，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[1-3]（Support Vector Machine，SVM），它能够较好地处理分类和回归问题。该技术已成为智能化数据处理领域的研究热点，并在很多领域得到了成功的应用。但是，作为一种尚未成熟的新技术，支持向量机目前还存在局限。例如，客观世界存在大量的模糊信息，如果支持向量机的训练集中含有模糊信息，那么支持向量机将无能为力。

文献[58]、文献[59] 提出的 FSVM 方法，只是对支持向量机做了一些改进，没有从算法的本质上建立模糊支持向量机。而且 FSVM 所处理的模糊信息仅是一般模糊信息的特例，缺乏一般情况下含有模糊信息的支持向量机的研究。

模糊数学[61]是处理模糊信息的有力工具。模糊模式识别是模糊数学中处理模糊分类问题的有效方法，但模糊模式识别方法与含有模糊信息的支持向量机方法提法不同，即已知条件和解决的问题不一样。因此不能简单地用模糊模式识别方法解决含有模糊信息支持向量机所要解决的问题。

由于现实问题的复杂性，人们还需要在实际应用的求解过程中对已有的基于数据处理的

智能技术进行不断完善，企望达到更好的应用效果。

1.2 基于模糊信息的智能技术研究现状

随着数据库技术的迅猛发展及数据库管理系统的广泛应用，人们积累的数据越来越多，这些海量数据中包含有大量模糊的、不完全的、有噪声的、随机信息。由于缺乏有效的数据处理方法导致了“数据爆炸但知识贫乏”的状况。为了解决这一问题，数据挖掘应运而生，使数据库技术进入一个更高级的阶段，可以快速、有效地处理（合成）各类复杂信息，实现对海量数据的智能化管理，从而促进信息传递。数据挖掘是数据库技术与智能计算相结合的一个交叉学科，融合了数据库、数据仓库、人工智能、机器学习、模糊集合与模糊逻辑、模糊信息处理、神经网络信息处理、模糊神经网络信息处理、统计学、模式识别、信息检索、遗传算法等多学科内容。目前，数据挖掘已是计算机领域内最为活跃的一个分支，研究成果的应用也已相当普及。

数据挖掘是一个新兴的研究方向，从数据库中发现知识（Knowledge Discovery in Database，KDD）这一术语首次出现在1989年举行的第十一届国际联合人工智能学术会议上，面向模糊信息的数据挖掘——模糊系统与知识发现（Fuzzy Systems and Knowledge Discovery，FSKD）是会议的一个重要研究专题。本次会议之后人们将FSKD和KDD统称为KDD。IEEE的《Knowledge and Data Engineering》会刊率先在1993年出版了KDD技术专刊。随着KDD在学术界和工业界的影响越来越大，由美国人工智能协会主办的KDD国际研讨会1995年已由原来的专题讨论会发展成为国际学术大会（International Conference on Data Mining & Knowledge Discovery in Databases），1995年在加拿大蒙特利尔市召开了第一届KDD国际学术会议，以后每年举办一届。1998年筹备了ACM SIGKDD兴趣组，并于1999年将KDD国际会议纳入ACM兴趣组系列的重要会议。数据库和人工智能方面的重要国际会议，如AAAI、IJCAI、ACM SIGMOD、VLDB、PODS和PAKDD等都将面向模糊信息的智能技术列为重要的研究专题。此外，并行计算、计算机网络和信息工程等其他领域的国际学会、学刊也把数据挖掘和面向模糊信息的知识发现列为专题和专刊讨论。

数据挖掘在我国起步较晚，但发展速度非常快。

1997年亚太地区在新加坡组织了第一次规模较大的PAKDD学术研讨会，面向模糊信息的知识发现被列为重要的研究专题。2000年6月，中国计算机协会数据库专委会在上海召开了第一届WAIM（Web-Age Information Management）国际会议，KDD、FSKD是其中的2个重要的研究议题。该会议每年举办1次，FSKD方面的文章每次都占相当大的比例。

2002年11月，在新加坡召开了首届模糊系统和知识发现国际会议（International Conference on Fuzzy Systems and Knowledge Discovery，ICFSKD）。自2005年至今，FSKD国际会议与自然计算国际会议（International Conference on Natural Computation，ICNC）一直联合召开，每年召开1次。

2005年8月，第二届模糊系统与知识发现暨首届自然计算联合国际会议在长沙召开。大会得到国际神经网络协会的支持，会议论文集由Spring出版。

2006年起，该联合会议得到IEEE计算智能协会或IEEE计算机协会的支持。2006年9月，第三届模糊系统与知识发现国际会议暨第二届自然计算国际会议在西安召开。2007年8

月，第四届模糊系统与知识发现国际会议暨第三届自然计算国际会议在海南召开。2008 年 10 月，第五届模糊系统与知识发现国际会议暨第四届自然计算国际会议在山东召开。第六届模糊系统与知识发现国际会议暨第五届自然计算国际会议将于 2009 年 8 月在天津召开。

国内的许多科研单位和高等院校竞相开展知识发现基础理论及其应用研究，并取得了卓有成效的研究成果。

目前，它的研究重点也逐渐从最初的发现方法转向系统应用，注重多种发现策略和技术的集成，以及多学科之间的相互渗透。

支持向量机（Support Vector Machine，SVM）是数据挖掘的一种新方法，是 Vapnik 等人[1-3]提出的。SVM 基本上不涉及概率测度及大数定律等，因此不同于以往的统计方法。从本质上看，它避开了从归纳到演绎的传统过程，实现了从训练样本到预测样本的高效“转导推理”，大大简化了通常的分类和回归等问题。

支持向量机的核心内容基本是在 1992—1995 年形成的。从那以后，关于支持向量机方面的论文及专著如雨后春笋，逐渐成为国际上机器学习领域新的研究热点。对支持向量机理论的研究，主要集中在对统计学习理论本身、核技术及对支持向量机算法实现的研究上。

支持向量机面世以来，学者们从理论上解释了支持向量机的优势所在。例如，Yi Lin[7]指出，对于模式识别问题，支持向量机不是去试图估计样本的密度分布 $p(x)$，而是通过核技术实现对决策函数 sgn $(p(x)-0.5)$ 的逼近，后者可以看作是 Bayes 最佳估计的近似。

在理论研究上支持向量机具有突出优势，但其应用研究相对比较滞后。随着理论的不断完善，到目前为止，SVM 在模式分类、回归分析、函数估计等领域都有广泛应用[36]。支持向量机应用最广泛的当属模式识别领域，已成功地用于许多模式识别问题。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了实验，取得了较大的成功。实验结果显示：人工识别平均错误率为 2.5%，专门针对该特定问题设计的 5 层神经网络错误率为 5.1%（其中利用了大量的先验知识），而用 3 种 SVM 方法（采用 3 种核函数）得到的错误率分别为 4.0%、4.1% 和 4.2%，且是直接采用 16×16 的字符点阵作为输入，表明了 SVM 的优越性能。

原始数据的多样性和复杂性，要求人们在 SVM 的计算方法上，必须注重多种信息处理策略和技术的集成。

关于支持向量机的最新动态，可访问 http：//www. svms. org/。

作为统计学习理论的具体实现，支持向量机具有坚实的数学理论基础和严格的理论分析，是机器学习中的一种新方法和研究热点。但是目前支持向量机还处在不断的完善发展中。主要表现在以下几个方面。

（1）支持向量机算法某些理论解释并非完美无缺。例如，Burges[15]就曾经提到结构风险最小化原则并不能严格证明支持向量机为什么具有好的推广能力；对 VC 维的分析尚无通用的方法等。

（2）支持向量机的训练过程，虽然有了很多快速训练算法，但当样本规模较大时，算法的收敛速度还是仍然较慢，特别是当支持向量的原始信息中含有模糊信息时，直接应用目前的模糊数学知识难以保证较高的实时性要求。

（3）核函数的选择及核参数的确定，尚无理论上的依据。Keerthi 证明了径向基函数核在适当选择参数时可以代替多项式核，一般情况下选用径向基函数核的效果不会太差，但对于具体问题仍需相应的专业知识及对象特性合理选择核函数。目前核及参数的选择，还是一个

公开的难题。

（4）对模式分类来说，支持向量机本质上属于两类分类算法，在多类分类问题上，支持向量机还存在构造学习机器及分类效率低的缺点。

需要指出的是，支持向量机算法中的这些不足，不只是SVM算法具有。许多机器学习领域的别的算法也面临训练和多类分类效率低的问题，而目前基于核技巧成为从非线性空间寻求线性判别的通用方法，也具有难以选择核及其参数的问题。对这些问题的深入研究，同时也将促进机器学习领域的发展。

1.3 智能技术在中医症状鉴别领域的研究现状

中医药是人类和疾病长期斗争的知识结晶，是中华民族留给世人的宝贵财富。然而在药理上，中医药还有很多不完善的地方，传统研究方法在中医药的研究中遇到了很多困难，严重影响了对传统中医药的继承和发展[4]。

把智能技术应用于传统的中医药方剂学理论和中医症状鉴别的研究，从古今大量验方中挖掘出对中医药研究有价值的信息，不仅可以为智能化数据处理技术开辟新的应用领域，还可以为中医研究提供新途径、新思路，为推动传统中医的发展和繁荣做出贡献。

经过历代中医专家的不断完善，人们已经基本了解了每一单味中药的性、味、归经、功效等药理特性。然而中医治病一般都是通过多味中药进行配伍组成方剂，每首方剂中的药相须相生，有君臣佐使之分。按照现有中医理论，各味药之间关系错综复杂，融合了自然科学知识和阴阳五行学说，以及传统儒家思想等人文知识。由于中医理论的不完善，中医对方剂药理特征的认识还停留在师带徒式的心传口授阶段。

经过数千年的临床实践，历代流传下来数百万首中药方剂。这些历代验方是对中医临床实践的真实记载，其中蕴含了前人临床用药的宝贵经验，体现了中医药的理、法、方、药的方方面面。

目前，我国已建有多个有关中医药的信息网和一批中医药文献、方剂、偏方、诊疗等数据库系统，初步满足了中医药界文献检索的需要。

智能技术在中医症状鉴别和药理分析等领域的应用还处于起步阶段，其在中医药领域的应用尚处于实验探讨阶段。目前，已有不少学者在进行这方面的研究，出现了一些可喜的研究成果。

西南交通大学的李力等把关联规则和粗糙集应用于中药方剂的分析，可实现对一些中药复方的初步分析，得到中药复方中的药组和药对[45]。但该系统操作复杂，功能简单，所用方法也较少。这是目前一套比较完整的中药复方数据挖掘系统。

赵蔡斌等采用相似系数来描述方剂中各味单药之间的差异程度[46]，相似系数的大小反映了各单药之间的相似程度。这种方法对小柴胡汤的分析具有较好的结果，揭示了其中一些中药配伍规律。

王咏梅等运用模糊聚类和模糊欧几里得距离分析了药物之间的配伍[47]，得出诸药间的相互作用，实验结果符合传统中医药理论的认识。

苏薇薇通过对中药的性、味、归经进行量化处理，然后用聚类方法研究了药物的配伍规律[48]，将方中的诸药分成了君、臣、佐、使药群，所得结果与传统中医药理论相吻合。

高媛等采用相对剂量计算出方剂的综合性、味、归经、功效及主治证[49]，客观地展现了方剂的性能和组方特点。

在河南省重大科技攻关课题（编号：60402）的支持下，安阳师范学院智能技术研究所与河南省中医学院、安阳市中医院的老中医专家合作，把基于数据处理的智能技术应用到中药方剂、中医症状鉴别的研究，试图从大量中医验方和类方中挖掘出未知的行医用药知识，从而指导中医药理论研究、指导中医临床用药、辅助医生进行中医症状鉴别。

1.4 主要研究内容

本书基于智能化数据处理的实际需要：研究解决一般情况下支持向量机中含有模糊信息问题的算法，LS-SVM 的参数优化及其应用，以支持向量机为支撑建立新的数据聚类算法，并且将这些研究直接应用到了“烧结矿化学成分的预测”和“中医症状鉴别智能诊断”系统中。

其内容包括如下几方面。

（1）基于 Fuzzy 集合理论（聚类、规划），构建模糊线性可分问题、模糊广义线性可分问题和模糊非线性问题的支持向量分类机（算法）。

（2）对最小二乘支持向量机参数的遗传算法进行改进。

（3）将支持向量机应用于烧结矿化学成分的预测。

（4）基于 Fuzzy 模糊聚类、模糊分类理论和“同类相近”的思想，对“最短距离聚类算法”[73]进行改进。使其适应于任意形状的聚类，在聚类时只考虑数据的最近邻点，而不考虑数据某一邻域范围内的其他点。

（5）提出了精英个体产生和初始种群产生方案，给出了有 SVM 支撑的仿生计算与基因表达式编程算法。

（6）设计了“中医症状鉴别智能诊断”系统。

第2章　支持向量机的基本理论

支持向量机是数据挖掘中的一种新方法，能非常成功地处理回归问题（时间序列分析）和模式识别（分类问题、判别分析）等诸多问题，并可推广于预测和综合评价等领域，可应用于理科、工科和管理等多种学科。目前国际上支持向量机在理论研究和实际应用两方面都正处于飞速发展阶段。它广泛应用于统计分类及回归分析中，支持向量机属于一般化线性分类器，这族分类器的特点是它们能够同时最小化经验误差与最大化几何边缘区，因此支持向量机也被称为最大边缘区分类器。

我们通常希望分类的过程是一个机器学习的过程，这些数据点是 n 维实空间中的点，我们希望能够把这些点通过一个超平面分开，通常这个被称为线性分类器。有很多分类器都符合这个要求，但是我们还希望找到分类最佳的平面，即使得属于2个不同类的数据点间隔最大的那个面，该面亦称为最大间隔超平面。如果我们能够找到这个面，那么这个分类器就称为最大间隔分类器。

支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的超平面的两边建有2个互相平行的超平面。建立方向合适的分隔超平面使2个与之平行的超平面间的距离最大化。其假定为：平行超平面间的距离或差距越大，分类器的总误差越小。

所谓支持向量是指那些在间隔区边缘的训练样本点。这里的“机（machine，机器）”实际上是一个算法。在机器学习领域，常把一些算法看作是一个机器。

本章简要介绍支持向量机分类机和支持向量回归机[3,5,39,96-98]，并分析支持向量机的理论基础——统计学习理论[63]。

2.1　最大间隔超平面

定义2.1　称集合 $H=\{x \mid p^{\mathrm{T}}x=\alpha\}$ 为 R^n 中的一个超平面，其中 p 是 R^n 中非零向量，α 是一个实数。称向量 p 为该超平面的法向量。

定义2.2　设训练集为

$$S=\{(x_1, y_1), \cdots, (x_l, y_l)\}, \tag{2-1}$$

其中，$x_j \in R^n$，$y_j \in \{-1, 1\}$，$(j=1, \cdots, l)$。所谓分类问题就是寻找 R^n 上的一个实值函数 $g(x)$，以便用决策函数 $f(x)=\mathrm{sgn}(g(x))$ 推断任一模式 x 相对应的 y 值。