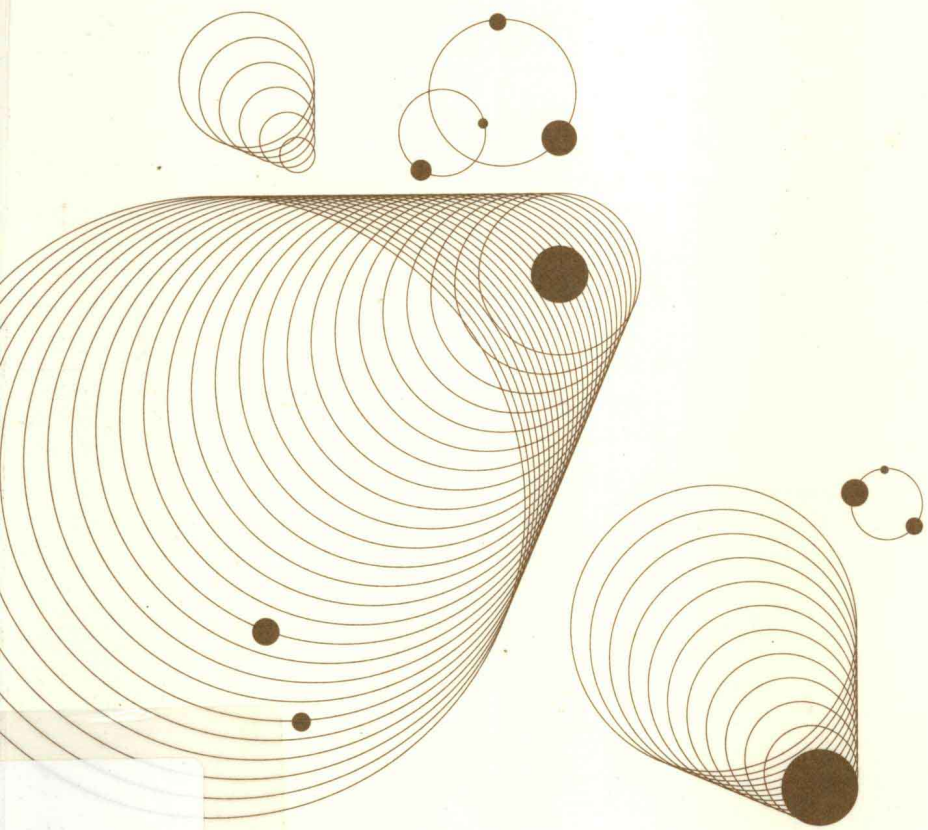
鲁棒最小二乘支持向量机研究与应用

Research and Application of Robust Least Squares Support Vector Machines

刘京礼 著



经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

鲁棒最小二乘支持向量机研究与应用

Research and Application of Robust Least Squares Support Vector Machines

刘京礼 著

经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

图书在版编目（CIP）数据

鲁棒最小二乘支持向量机研究与应用/刘京礼著. —北京：经济管理出版社，2012.5
ISBN 978-7-5096-1849-3

Ⅰ.①鲁… Ⅱ.①刘… Ⅲ.①鲁棒控制—研究 Ⅳ.①TP273

中国版本图书馆 CIP 数据核字（2012）第 088365 号

出版发行：经济管理出版社
北京市海淀区北蜂窝 8 号中雅大厦 11 层
电话：(010)51915602　　邮编：100038

印刷：北京银祥印刷厂　　经销：新华书店

组稿编辑：申桂萍　　责任编辑：孙　宇
责任印制：黄　铄　　责任校对：超　凡

720mm×1000mm/16　　10 印张　　180 千字
2012 年 6 月第 1 版　　2012 年 6 月第 1 次印刷

定价：35.00 元

书号：ISBN 978-7-5096-1849-3

前 言

二分类问题是模式识别、机器学习、统计学习理论以及人工智能中研究的一个重要问题。在二分类问题的模型中，支持向量机模型采用结构风险极小化原则和核函数方法来构造分类模型，模型比较简单，解具有唯一性。但由于支持向量机模型（SVM）需要求解二次规划，因此当数据量较大时，其计算需要大量的时间。作为对支持向量机模型的改进模型，最小二乘支持向量机模型（LS-SVM）通过使用误差均方和作为目标函数，把不等约束改为线性约束，使得支持向量机模型中二次规划的求解转化成求解线性方程组，克服了支持向量机模型求解二次规划计算量大的问题。但是，最小二乘支持向量机模型中的等式约束以及目标函数中的误差均方和使得模型的解丢失了稀疏性，降低了解的鲁棒性。

由于随机过程或者非随机过程的存在，现实生活中的数据经常带有噪声和不确定性。数据的噪声以及不确定性会影响统计学习分类算法模型的性能，降低分类的准确率以及分类模型的推广能力。支持向量机和最小二乘支持向量机模型都是采用了固定范数的目标函数，这种建立模型的方法不能够很好地适应各种各样的数据结构，从而使得模型的适应能力较弱。为了加强最小二乘支持向量机模型的鲁棒性和稀疏性，增强其推广能力，使模型能够根据数据结构自动进行调整，本书主要包括以下几个方面的工作：

（1）系统整理了文献中对支持向量机模型和最小二乘支持向量机模型中改进鲁棒性的方法，并指出这些改进模型存在的问题和缺

陷，从而得到了本书将要研究的主要问题，即以加强最小二乘支持向量机模型的稀疏性、鲁棒性和可解释性为目的，对原有模型进行了较大的改进，给出了基于最小二乘支持向量机模型的有效二分类算法模型。

（2）针对最小二乘支持向量机模型丢失稀疏性和鲁棒性的原因，提出了使用核主成分法对样本数据中存在的噪声特征进行剔除，并借鉴先前的增强最小二乘支持向量机模型稀疏性的方法，对特征进行压缩，给出了一个双层 L_1 范数最小二乘支持向量机模型——KPCA–L_1–LS–SVM。通过使用 KPCA 方法，可以有效地进行特征抽取和提取。同时以 L_1 范数作为目标函数，可以有效地消除噪声点对模型推广能力的影响，并使模型的解更稀疏，从而可以降低计算的复杂度。在仿真数据集和基准数据库上对该模型的测试表明该方法是有效的。

（3）在实际的二分类问题中，由于噪声点或者噪声特征的存在，使得样本的标签会出现不确定的情况。分类模型应该能够自动判别哪些是相对重要的样本、哪些是受噪声点影响较大的样本，从而在分类函数的构造中剔除这种样本。模糊隶属度的概念则可以用来描述样本标签的不确定性。本书采用 L_1 范数作为目标函数以及模糊隶属度的概念，可以构造出一个具有稀疏性和鲁棒性的基于最小二乘支持向量机的分类模型——模糊 L_1–范数 LS–SVM（FL_1–LS–SVM）。在测试数据集上的测试表明这个模型同样可以消除噪声点的影响，并具有较好的可解释性。

（4）在分类问题中，不同的样本在分类函数的构造中所起的作用是不同的。在分类函数的构造中，样本所包含的判别信息越是重要，相应的样本对分类模型的构造所起的作用就越大。因此，为了区别不同样本对于决策函数构造的不同作用，应该对包含重要信息的样本赋予较大的权重，而包含次要信息的样本所对应的权重就会较小。通过这种赋权的方法也可以消除噪声点对分类模型的影响，

使得模型具有鲁棒性特征。无论是支持向量机还是最小二乘支持向量机模型，在目标函数中都使用固定的 L_p 范数，这是一种基于先验知识的建模方法，不能适应各种各样复杂的数据结构。从模型更好的适应数据的角度出发，本书提出了一个鲁棒赋权最小二乘支持向量机模型——RW-L_p-LS-SVM。在仿真数据集以及 UCI 基准数据库上的测试表明该模型具有鲁棒性特征，稀疏性好，具有较好的解释能力。

（5）信用评估数据库所包含的数据类型比较特殊，其类别比例极不均衡。为了检验本书所提出的三个模型的分类性能，我们使用这三个模型在三个信用数据库上进行测试，所得到的结果说明模型能够较好地适应信用数据库类别不均衡的特点，因而可以作为信用风险评价的备选模型。

本书的读者对象设定为对模式识别、机器学习、优化算法有一定了解的人士。对于支持向量机、最小二乘支持向量机的理论，本书做了系统严谨的论述，适合于不同层次的读者。对理论部分感兴趣的读者，可以参考本书的理论证明部分；仅关心应用的读者，可以略去这些证明，阅读本书中的实例部分。希望本书的出版能够进一步推广支持向量机在实际领域中的应用，促进其深入研究。

本书的出版要感谢国家自然科学基金（71171123）、山东省博士基金（BS2011SF011）以及山东工商学院出版基金的资助，也要感谢山东工商学院各级领导的支持和技术创新重点学科的资助。同时，还要感谢经济管理出版社编辑同志的大力协助，使得本书能够顺利出版。

由于作者水平有限，错误和疏漏之处在所难免，欢迎读者批评指正。

目 录

附 图

附 表

1 绪论

本章主要介绍本书的研究背景和意义、本书的主体鲁棒最小二乘支持向量机二分类问题的文献综述以及研究要点、本书的研究内容和结构安排以及本书所使用的技术路线。

1.1 研究背景和意义

信息是现代社会人类各种活动的基石。随着信息、通信和网络技术的飞速发展，各种各样的数据库例如银行客户数据库、超市交易信息数据库、发明数据库、医疗和统计数据库的数据规模也在不断快速的增长。比如，一个有很多分支机构的商业银行，其网络终端会记录银行客户每一次交易的时间、 地点、存取款的数量、使用金融卡消费的金额等，这种记录每天都会产生几个 GB 的数据量。这些数据往往呈现出一种非线性、高维的特征。而且，这些数据库中包含了大量隐含和丰富的信息，可以用来帮助我们更好地进行决策。如何对这些数据进行高效的分析和管理，从隐含在数据库的信息中提取出有用的知识是人工智能、模式识别以及机器学习中一个非常关键的问题。

一个学习问题可以被描述为 (Vapnik，1995)：给定空间 Z 上的一个概率测度 F(z)，F(z) 是未知的，给定的样本 z_1，…，z_l 是

独立分布的。考虑函数集合 $Q(z, a)$，$a \in A$，学习问题的目标是极小化风险函数 $R(a) = \int Q(z, a)dF(z)$，$a \in A$，其中 $Q(z, a)$ 是特定的损失函数，例如 0~1 损失函数、二次损失函数等。从这个定义来看，学习问题的任务就是从给定的数据中推断隐含在数据中的函数依赖关系，并根据这种依赖关系，对未来的数据做出预测和推断（Vapnik，1998）。

分类是学习问题的一种，是统计学中最古老和研究最多的一个问题，也是人工智能、模式识别和机器学习理论研究的重点问题。现实中的许多问题比如手写汉字识别、疾病检测、电子邮件过滤、语音识别、图像分类、违约贷款预测、网页检索、信用数据分类、分子特性的确定等都属于分类问题。分类在统计学中被称为判别分析，而在工程领域中则被称为模式识别。在 20 世纪 60 年代之前，统计学家所研究的大多数问题的样本量都比较小，而且样本的维数较低。对于这种数据，可以假定样本数据服从带有未知参数的某一种分布，从而可以采用各种估计方法来求解参数，这属于传统的参数统计方法。但在处理非线性高维数据时使用传统的参数统计方法会出现维数灾难的问题，因此需要寻找传统的统计分析方法的替代方法。

对于一个特定的分类问题来说，很重要的一点是确认分类中需要用到哪些最重要的属性（特征），这个问题被称为特征选择。分类问题中另一个重要的部分是构造决策函数（分类函数），也就是如何利用有限的样本数据来选择合适的分类模型，这是分类问题的核心。在分类问题中，决策函数的形式也就是决策函数的结构是模式分类中一个研究的热点问题。带标签的观测值的数量、每一个观测值的维数、观测值所选择的特征、观测值的独立性、噪声点的多少等问题都会对分类函数的选择及分类的准确率有很大的影响。

在现实生活中，由于随机过程或者非随机过程的存在，所收集

到的样本数据具有不确定性，数据可能有噪声，甚至是相互矛盾的。那么，对于使用这些样本数据来进行学习所构造的分类模型来说，需要模型对这些数据具有鲁棒性，即模型所导出的分类函数受数据中不确定性因素的影响较小。鲁棒这个词最早出现在 1953 年（Huber，1981），主要是在统计理论中分析数据的较小的摄动对统计模型的影响，后来被广泛应用到控制理论中，研究带有摄动的数据对模型稳定性的作用。传统的规划模型中处理包含不确定性因素的方法是采用随机规划（Kall 和 Wallace，1994），以及在管理学中常采用的敏感性分析（Dantzig，1955；Dantzig 和 Infanger，1993；Ben-Tal 和 Nemirovski，1998）。采用这些分析可以发现数据的不确定性对于模型的影响，从而在构造模型时能够尽量减小这种影响。

1.2 鲁棒支持向量机研究综述

本书主要研究的是对最小二乘支持向量机模型的鲁棒性和稀疏性加以改进的方法。这一部分我们将对支持向量机和最小二乘支持向量机的鲁棒性改进方法加以详细论述，包括支持向量机模型最小二乘支持向量机模型的统计学习理论基础、模型存在的问题以及支持向量机模型的鲁棒改进方法和最小二乘支持向量模型的鲁棒改进方法。二分类学习算法和支持向量机模型以及最小二乘支持向量机模型，我们将在下一章做具体介绍。

1.2.1 统计学习理论的发展

统计学习理论的发展可以分为四个阶段（Vapnik，1995），分别是学习机的构造、基础学习理论的构建、神经网络的出现和神经网络替代方法（支持向量机模型的出现）的发展。

1.2.1.1 感知器的出现（20 世纪 60 年代）

1962 年，心理学家 F.Rosenblatt 在其撰写的《神经动力学原理：感知器和人脑机制原理》一书中提出了感知器的最初模型。这是一个模仿人类学习过程的线性分类模型，能够使用机器对人类的学习方式进行模拟并使用函数关系来对这个过程进行刻画。Novikoff（1962）对感知器可以分类训练数据的定理的证明是学习理论研究的一个里程碑。

1.2.1.2 基础学习理论的发展（20 世纪 60~70 年代）

这一时期出现了许多学习理论，包括经验风险极小化理论（VC 熵，VC 维，结构风险极小化归纳原则（V.Vapnik，1982）等）、求解病态问题的理论（规范化理论）（A.N.Tikhonov，1963）、无参数概率密度估计理论（V.Vapnik，1978）、算法复杂性理论（R.J.Solomonoff，1960）等。

1.2.1.3 神经网络的出现（20 世纪 80 年代）

1986 年，几位研究人员通过使用 Sigmoid 函数来代替感知器输出中的符号函数，解决了同时计算所有神经元系数的问题。这就是后来被称为神经网络的后向传播感知器方法。神经网络以及决策树都可以用来处理样本数据中存在的非线性关系。但是，对于神经网络模型容易出现过拟合以及局部最优的问题。

1.2.1.4 支持向量机模型的出现（20 世纪 90 年代至今）

20 世纪 90 年代初，Vapnik（1995）提出了支持向量机模型。该模型是基于统计学习理论，以结构风险极小化原则（V.N.Vapnik 和 A.Ja.Chervonenkis，1974）和极小描述长度原则（Rissanen，1978）从小样本（V.Vapnik，1998）数据中进行学习的一种分类模型。SVM 模型的形式简单，有唯一解，通过使用核函数可以处理非线性的高维数据。SVM 在核学习方法和模式识别之间建立了一座桥梁（Shawe-Taylor 和 Cristianini，2004），它提供了一个可推理的统一框架并可以处理各种类型的数据。

1.2.2 支持向量机模型的鲁棒性

SVM 模型采用极大间隔方法以及结构风险极小化原则来对分类函数进行学习，其模型是一个二次规划模型，具有较好的稀疏性。但该模型对于噪声点也比较敏感，因此出现了许多改进 SVM 模型的鲁棒性的模型。对该模型的鲁棒性进行改进的方法主要是要消除噪声点对分类超平面构造的影响，使得样本数据中包含的信息量与样本点对于构造分类超平面的作用是成比例的。

Chen（1994）为了能够从信号表示的过完备词典中找到信号表示的最优叠加，而采用了基追踪的方法。该方法中的最优准则是采用矩阵等式中系数的 L_1 范数形式，也就是在满足约束 $\Phi\alpha = s$ 的条件下极小化 $\min \| \alpha \|_1$。其实质也是对等式约束中的系数 α 进行压缩，从而得到信号的一个稀疏表示。该规划模型的求解是采用了原始对偶对数障碍法。Bradley 等（1998）提出了一个能够进行特征选择的凹函数形式，该方法在目标函数中使用了 L_1 范数。Herbrich 等（1999）提出了一个自适应间隔的 SVM 模型，这个模型中，当输入点 x_i 的映射 $\varphi(x_i)$ 要成为一个奇异值的时候，根据算法就会给它对应的拉格朗日乘子 α_i 一个较大的值，从而使得奇异值对分类平面的影响较小。Zhu 等（2004）提出了一个 L_1 范数-支持向量机模型，并给出了求解局部极小的一个多项式迭代算法。该模型的损失函数采用了铰链损失函数 $(1 - y_i(\omega^T x + b))_+$，并把标准 L_2 范数的正则化项 $C \| \omega \|_2$ 改为岭惩罚函数 $C \| \omega \|_1$。把 L_2 范数改为 L_1 范数的主要目的是压缩 ω 的非零元素个数，使得 ω 的多数元素都为 0，从而简化了模型，增强模型的可解释性，同时达到稀疏性的目的。该模型主要用来处理样本数据存在多余的噪声特征的情况。Liu（2007）则针对 SVM 模型提出了一个自适应的 L_q 范数惩罚项，范数 q 可以根据数据自动选择数值，改进的模型具有较好的鲁棒性。Jung 等（2008）给出了一个带有自适应约束的原始对偶