

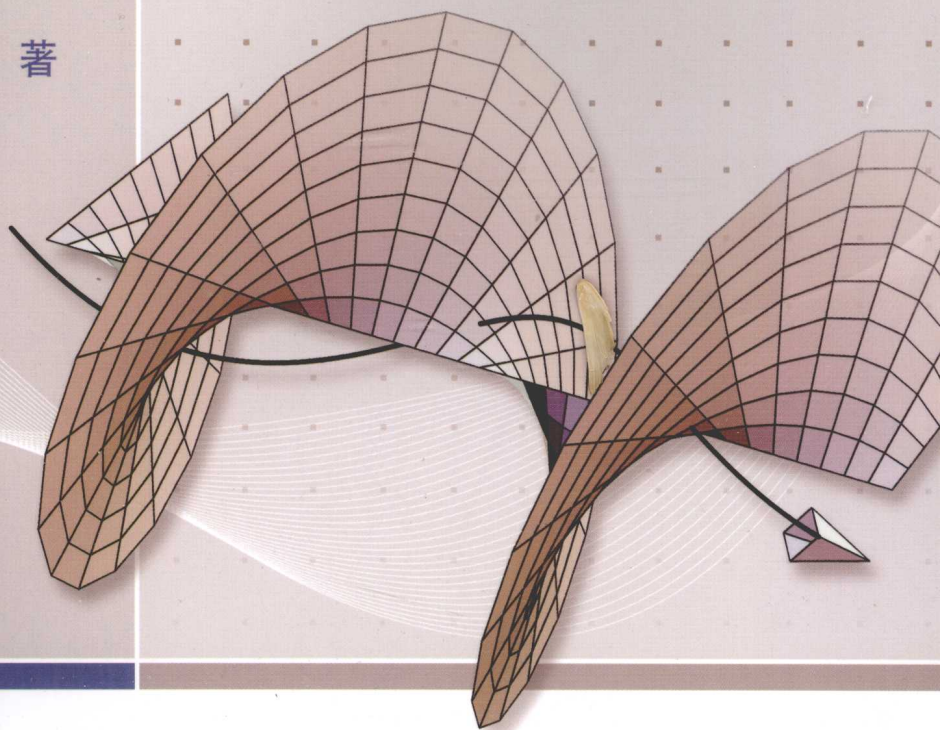
核方法与累积量 随机学习方法研究

*The Study of Kernel Methods
and Stochastic Learning of the Cumulative*

赵 峰
刘淑娥
安志勇
李大社

著

<http://www.phei.com.cn>



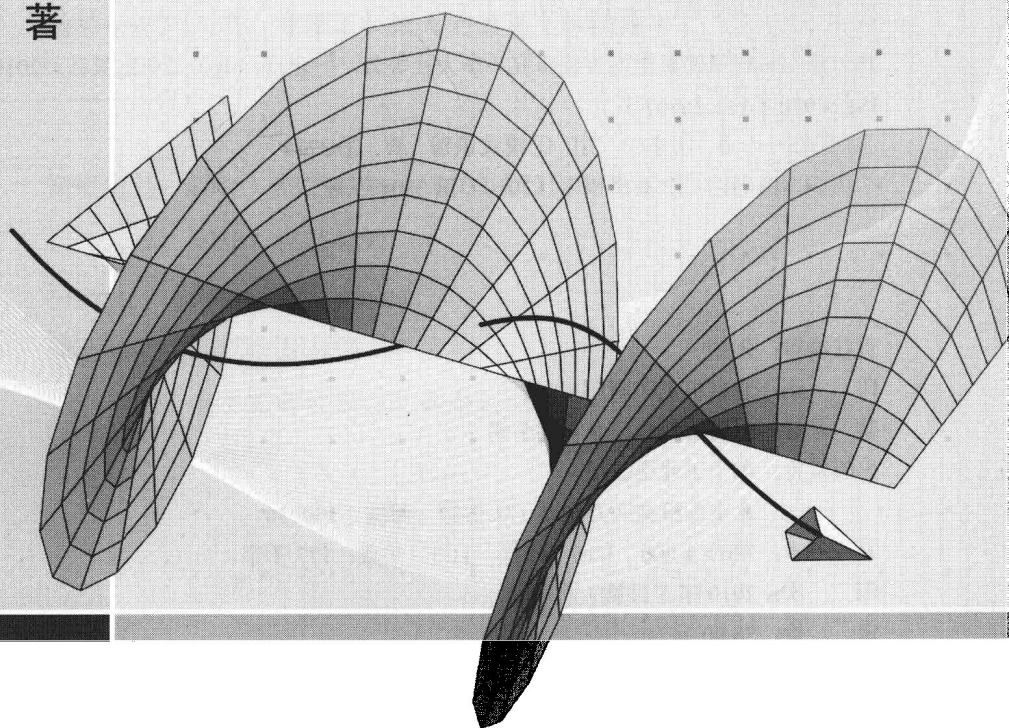
电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

核方法与累积量 随机学习方法研究

*The Study of Kernel Methods
and Stochastic Learning of the Cumulative*

赵 峰
刘淑娥
安志勇
李大社

著



電子工業出版社

Publishing House of Electronics Industry

北京 · BEIJING

内 容 简 介

模式分类是模式识别的一项重要内容,在许多领域已经得到成功的应用。本书主要针对近年来模式分类领域广泛关注的核方法和累积量随机学习方法进行了探讨和研究,内容包括绪论、特征空间中数据的几何结构表示、KPCA 的关键问题研究、KFDA 的关键问题研究、SVDD 的关键问题研究、新的核方法的结构和基于累积量随机学习方法的雷达目标识别。

本书适合从事机器学习研究和应用的科研人员阅读,也可作为高等学校研究生学习模式识别的教学用书。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有,侵权必究。

图书在版编目(CIP)数据

核方法与累积量随机学习方法研究 / 李大社等著. —北京:电子工业出版社, 2010.5

ISBN 978-7-121-10807-5

I. ①核… II. ①李… III. ①模式识别 IV. ①O235

中国版本图书馆 CIP 数据核字(2010)第 080473 号

责任编辑:张剑

印 刷:北京机工印刷厂

装 订:三河市鹏成印业有限公司

出版发行:电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本:720×1 000 1/16 印张:11.5 字数:177 千字

印 次:2010 年 5 月第 1 次印刷

定 价:58.00 元

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010) 88254888。

质量投诉请发邮件至 zltts@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线:(010) 88258888。

前 言



模式分类是模式识别的一项重要内容,在许多领域已经得到成功的应用。本书主要从以下两个方面对模式分类进行了探讨和研究。

(1) 针对近年来模式分类领域广泛关注的核方法,分别就参数调节问题、计算复杂度问题及新的核方法构造问题等进行了一些探讨。

(2) 针对模式分类的一个重要应用领域——基于高分辨距离像的雷达目标识别,研究了累积量随机学习方法在该领域的应用。

本书共分7章。第1章为绪论,第2章介绍特征空间中数据的几何结构表示,第3章论述了KPCA的关键问题及解决方法,第4章对KFDA的关键问题处理方法进行了阐述,第5章讨论了SVDD的关键问题研究,第6章针对核方法构造问题给出了一种新的核方法,第7章主要讨论了累积量随机学习方法在雷达目标识别领域的应用。全书由赵峰、刘淑娥、安志勇、李大社写作。

本书的研究工作得到国家自然科学基金(60673153、60970105、60773053)、山东省信息产业发展专项资金项目(2008X00039)、山东省优秀中青年科学家奖励基金(2008BS01026、BS2009DX038)、山东省自然科学基金(2009ZRB01620、ZR2009GL011)、山东省高校科研发展计划(J09LG22)、山东省教育厅科技计划(J08LJ06)和烟台市科技发展计划(2008158)等项目的资助;参加研究工作的还有李晋江博士、刘敬博士、梁军利博士、李维勤博士等,在此一并致谢!

另外，要特别鸣谢华臻教授、范辉教授、原达教授、谢青松副教授和刘培强副教授，感谢他们为本书的写作提供的帮助和支持。

本书可作为从事机器学习研究和应用的科技人员的参考用书，也可作为高等学校研究生学习模式识别的教学用书。

由于撰写时间和作者水平有限，书中难免存在缺点和错误，敬请读者批评指正。

作者

2010.01.10

目 录



第一章 绪论	1
1.1 模式分类概述	1
1.1.1 基本概念	1
1.1.2 模式分类的基本方法	3
1.1.3 模式分类系统	5
1.1.4 模式分类系统性能评价	6
1.2 模式分类中的核方法概要	7
1.2.1 核方法的基本思想	7
1.2.2 核函数的构建	9
1.2.3 核方法的研究现状及趋势	14
1.2.4 核方法的应用	16
1.3 累积量随机学习方法简述	18
1.3.1 概率密度估计方法介绍	18
1.3.2 累积量随机学习方法简述	19
1.4 本书内容概要	21
1.4.1 本书研究工作概述	21
1.4.2 本书的章节安排	22
第二章 特征空间中数据的几何结构表示	24
2.1 基本思想	24



2.2	子空间标准正交基的确定	25
2.2.1	子空间基的确定	25
2.2.2	子空间标准正交基的确定	28
2.2.3	投影坐标的显式表示	28
2.3	本章小结	29
第三章	KPCA 的关键问题研究	30
3.1	KPCA 算法描述	31
3.1.1	PCA 的数学描述	31
3.1.2	KPCA 简介	34
3.2	KPCA 快速算法描述	36
3.2.1	KPCA 复杂度分析	36
3.2.2	核 Hebbian 算法 (Kernel Hebbian Algorithm, KHA)	38
3.2.3	分块核主成分	40
3.2.4	基于张成子空间的简化 KPCA	43
3.2.5	实验分析	46
3.3	KPCA 的核参数优化	49
3.3.1	核参数优化思路	50
3.3.2	高斯性测度的核参数优化算法	51
3.3.3	实验	53
3.4	小结	56
第四章	KFDA 的关键问题研究	58
4.1	KFDA 算法描述	59
4.1.1	Fisher 判别分析简述	59
4.1.2	核 Fisher 判别分析简述	61
4.2	KFDA 的奇异性分析	63
4.2.1	KFDA 的奇异性问题	63



4.2.2	KDDA 算法	64
4.2.3	OKFDA 算法	67
4.2.4	NKFDA 算法	70
4.2.5	OKDV 算法	71
4.3	KFDA 的快速算法	73
4.3.1	KFDA 计算复杂度分析	73
4.3.2	KFDA 的快速算法描述	74
4.3.3	实验分析	76
4.4	KFDA 的核参数优化	78
4.4.1	KFDA 的二次规划形式	79
4.4.2	优化核参数方法的推导	79
4.5	小结	82
第五章	SVDD 的关键问题研究	83
5.1	支撑向量域描述 (SVDD) 算法介绍	83
5.1.1	研究背景	83
5.1.2	SVDD 算法描述	84
5.2	SVDD 的简化算法	89
5.2.1	简化算法介绍	89
5.2.2	乘性规则算法 (SVDCM)	90
5.2.3	增量学习算法	92
5.3	SVDD 的核参数优化	95
5.3.1	SVDD 的核参数优化问题	95
5.3.2	核参数优化算法的思路	95
5.3.3	核参数优化测度的构造	96
5.3.4	最优化核参数的算法步骤	98
5.3.5	实验及结果分析	98
5.4	本章小结	102



第六章 新的核方法的构造	106
6.1 基本思想	107
6.2 最优变换和聚类中心算法描述	109
6.2.1 最优变换	109
6.2.2 最优聚类中心的选择	110
6.3 核最优变换和聚类中心算法	112
6.3.1 KOT-CC 算法推导	113
6.3.2 性能比较分析	115
6.4 核最优变换和聚类中心的简化算法	115
6.5 实验分析	117
6.5.1 鉴别特征提取能力分析	117
6.5.2 计算性能分析	119
6.5.3 与核 Fisher 鉴别分析的比较	121
6.6 本章小结	123
第七章 基于累积量随机学习方法的 雷达目标识别	125
7.1 HRRP 的特性分析	126
7.1.1 HRRP 的获取	126
7.1.2 HRRP 的敏感性及其预处理	127
7.1.3 帧内 HRRP 的统计特性分析	131
7.2 基于累积量随机学习方法的雷达目标识别	133
7.2.1 方法提出的背景	133
7.2.2 累积量随机学习方法 (SLC)	135
7.2.3 SLC 的主要步骤	137
7.2.4 运用 SLC 进行识别的基本步骤	137
7.2.5 基于外场实测数据的实验分析	138
7.2.6 结论	142



7.3 基于“Gamma—SLC”混合密度估计的雷达目标识别.....	143
7.3.1 方法提出的背景	143
7.3.2 密度估计方法的描述	145
7.3.3 密度估计方法的选择	146
7.3.4 采用“Gamma-TCL”进行识别的基本步骤	147
7.3.5 基于外场实测数据的实验分析	148
7.3.6 结论	150
7.4 本章小结	151
参考文献	153

第一章



绪 论

模式识别是计算机领域的一个重要研究分支,在其他各个领域都有着广泛的应用,具有很大的研究价值。模式分类是模式识别的一项关键技术,从 20 世纪中叶起,许多研究人员就已经开始这方面的研究,提出了多种理论和方法,积累了大量的研究成果。今天,模式分类已广泛应用于工业产品分拣、医学图像诊断、生物特征识别中,展现了广阔的发展前景。但是,由于模式分类涉及很多更复杂的问题,现有的理论和方法尚不能完全解决这些问题。

本章重点对模式分类的两种方法,即核方法与累积量随机学习方法,进行概念性的介绍,并对后续章节的内容安排做出说明。

1.1 模式分类概述

1.1.1 基本概念

模式识别作为一个研究领域,迅速发展于 20 世纪 60 年代。模式识别是一个多领域交叉学科,涉及统计学、工程学、人工智能、计算机科学、认知科学、心理学和生理学等学科。模式识别所研究的理论和方法在很多科学和技术领域中得到了广泛的重视,推动了人工智能系统的发



展, 扩大了计算机应用的可能性。几十年来, 模式识别研究取得了大量的成果, 在很多领域得到了成功的应用。

模式识别 (Pattern Recognition, PR) 中的模式 (Pattern) 一词, 最初是用于表示一个值得模仿或复制的理想人或物, 后来, 模式一般是指某种事物的标准形式或使人可以照着做的标准样式。

关于模式识别, 第六届国际模式识别会议主席, 德国爱尔朗根大学 H.Niemann 教授在《模式分类》专著中指出: “模式识别是研究模式的自动处理和判读的数学技术问题, 它既包含简单模式的分类, 也包含复杂模式的分析。” 针对本书要讨论的内容, 我们对模式、模式识别作如下狭义的定义: 模式是对某些感兴趣的客体的定量的或结构的描述。有人也习惯上把模式称为样本。模式类则是指模式所属的类别或同一类模式的总体, 简称类别或类。模式识别是研究一种自动技术, 依靠这种技术, 机器将自动 (或人尽量少地干涉) 把待识别的模式分配到各自的模式类中去。

通常, 根据学习过程中样本的类别归属是否被标志, 可以将模式识别划分为以下两大类。

1. 模式分类

模式分类是指, 在模式识别中, 利用已知类别的训练样本进行学习, 产生各类决策域的判决界面, 依据待识别样本所处的子区域而确定其类别。模式分类是模式识别系统的核心内容之一。从机器学习角度考虑, 模式分类属于有监督学习。

2. 模式聚类

模式聚类是指, 在模式识别中, 如果学习过程中样本的类别未知, 而是采用一定的距离测度测量样本的“距离远近”, 并基于“物以类聚”的思想, 将样本自然分类的识别过程。模式聚类属于无监督学习。

需要特别指出的是, 本书所定义的模式分类是一种狭义上的定义。有些文献上的模式分类则是一种广义上的定义, 即将模式识别与模式分



类看做同一个概念,而把有训练样本情况下的模式识别称为有监督模式识别(即本书所定义的模式分类),把没有训练样本情况下的识别称为非监督模式识别(即本书所定义的模式聚类)。在没有特殊说明的情况下,本书所指的模式分类概念,均是指狭义上的定义。

1.1.2 模式分类的基本方法

按照模式分类的使用原理和方法划分,模式分类方法可以归结为如下4类。

1. 统计模式分类

目前,统计模式分类的理论体系已经相当完善,方法也很多,如 Bayesian 方法、Fisher 方法、K-近邻方法等。从根本上来看,统计模式分类都是利用各类的分布特征,即直接利用各类的概率密度函数、后验概率等,或者隐含地利用上述概念进行分类识别。如 Fisher 方法就是首先用已知类别的训练样本产生判别函数,然后根据待分类模式代入判别函数后所得值的情况判定其类别。判别函数所提供的相邻两类判决域的界面,实质上相当于在一些设定下两类概率密度函数之差。

2. 结构模式分类

结构模式分类主要立足于分析模式的结构信息。结构模式分类是将对象分解为若干个基本单元(即基元),用这些基元及其结构关系来描述对象,基元及其结构关系可以用字符串或图形来表示;然后,运用形式语言理论进行句法分析,依据其是否符合某一类的文法而决定其类别。

3. 模糊模式分类

模糊模式分类是基于模糊数学的分类方法。现实世界中存在许多界限不分明、难以精确描述的事物或现象,而模糊数学则可以用数学的方法研究和处理这类具有“模糊性”的事物或现象。模糊数学的出现,使得人们可以模拟人类神经系统的活动,描述模式属于某类的程度。因此,



模糊数学在模式分类中得到了很好的应用。目前,模糊模式分类方法较多,如模糊 k -近邻、模糊最小/最大神经网络等。该类方法的有效性主要在于对象类的隶属函数是否良好。

4. 神经网络模式分类

神经网络模式分类是利用神经元网络中出现的神经计算模式来进行分类的。大部分神经元网络都有某种训练规则,如基于现有模式调节连接权重。换句话说,神经元网络直接对例子进行学习,得出其结构特征,进而推广,就像孩子从狗的例子中认识狗一样。

人工神经网络可以超越传统的基于计算机的模式分类系统的能力。人们可以利用计算机或神经元网络进行模式分类。计算机利用传统的数学算法来检测给定的模式是否跟现有模式相匹配。这是一个简单易行的方法。但该方法只能进行是或非的判断,且不允许模式有噪声。神经元网络则可以允许模式有噪声,而且如果训练得当,神经元网络会对未知模式的类别做出正确的响应。虽然神经元网络不能创造奇迹,但若采用合适的结构,对好的数据进行正确的训练,神经元网络就可以给出令人惊异的结果。比如,BP 神经网络直接从观测数据(训练样本)学习,非常简便、有效,因而获得了广泛应用,但它是一种启发式技术,缺乏指定工程实践的坚实的理论基础。

上述分类方法中,统计模式分类与结构模式分类是模式分类中的经典性和基础性技术;而模糊分类与神经网络分类则是最近发展起来的新方法,是信息科学与人工智能的重要组成部分。上述 4 种分类方法各有其特点与应用范围,它们不能相互取代,只能相互共存,相互促进、借鉴、渗透与融合。一个好的分类方法可能综合利用了上述各类识别方法的观点、概念和技术,如本书涉及的许多核方法,从本质上讲应该属于统计模式分类,但其具体的实现过程也涉及神经网络、信息熵等一些有关内容;同样,本书所提出的累积量随机学习方法,理论上属于非参数概率密度估计范畴,但却是借助多层感知器来实现的。



总之,模式分类作为一门学科,尽管它已经经历了 40 多年的历史,建立了丰富的理论体系,但由于问题的复杂性不断提高,以及应用领域范围的不断扩大,决定了它应该是开放的、发展的学科。随着相关学科新成果的不断涌现,模式分类的新方法与新理论也在不断涌现,如基于统计学习理论的支撑向量机,基于小波分析、马尔可夫随机模型的图像分类技术等都是典型的例子。

1.1.3 模式分类系统

一般而言,模式分类系统都由两个过程组成,即设计和实现。设计是指用一定数量的样本进行分类器设计;实现是指用所设计的分类器将待识别的样本进行分类决策。图 1-1 所示为一个模式分类系统的简单框图,它主要由信息获取,预处理,特征(或基元)提取与选择,分类器设计及分类决策 5 部分组成^[4]。

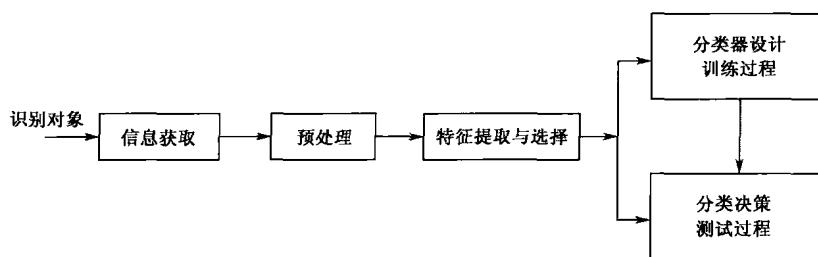


图 1-1 模式分类系统的基本构成

1. 信息获取

信息获取是指利用各种传感器把被研究对象的各种信息转换为计算机可以接受的数值或符号(串)。习惯上,称这种由数值或符号(串)所组成的空间为模式空间。

2. 预处理

预处理的目的是去除信息获取中掺入的干扰和噪声,人为地加强有用的信息,并对种种因素造成的退化现象进行复原。



3. 特征（或基元）提取与选择

由信息获取部分获得的原始数据量一般是相当大的。为了有效地实现分类识别，要对原始数据进行选择或映射，从而得到最能反映分类本质的特征，把这些特征组成一个向量（称为特征向量）。此过程即为特征特取与选择过程。通过特征提取与选择，不仅减少了处理时间，而且也减少了分类错误。

4. 分类器设计

为了把待识别的模式分配到各自的模式类中，必须设计出一套分类判别规则。其基本做法是，用一定量的训练样本确定某个判别规则，使得按照这种判别规则对被识别对象进行分类所造成的错误识别率最小或引起的损失最小。

5. 分类决策

分类器按已确定的判别规则对待识模式进行判别，输出分类结果，这就是模式分类的实现过程，也称为分类决策。

上述每一阶段设计的好坏，都会对整个系统的性能产生严重的影响，所以每个阶段都应该力争尽善尽美。在模式分类系统中，第1部分和第2部分一般属于数字信号处理和图像处理领域所研究的课题，而且大都与所研究的具体问题有关。本书重点讨论核方法与累积量随机学习方法在第3部分至第5部分上的应用。

1.1.4 模式分类系统性能评价

对于一个具体的模式分类系统，可以从以下4个方面去评测它的性能^[8, 9]，并依据评测结果来决定是否有必要改进系统的组成部件。

1. 识别率

识别率包括训练样本的识别率及测试样本的识别率。无论对于训练数据还是对于测试数据，一个好的分类系统都应该保持较高的识别率。

如果一个系统单纯对训练样本能获得完美的表现,但对测试样本不能令人满意,即系统存在“过适应”现象,则不能称为完美的分类系统。

2. 计算复杂度

有些模式分类问题尽管理论上可以用某种算法“解决”,但很不切合实际。例如,在光学字符识别中,对 20×20 的二值点阵图像的所有可能情况都进行分类标志,然后用“查表法”的方式对输入样本分类。尽管从理论上说,确实可以达到无错误的识别结果,但是由于需要处理 $2^{20} \times 2^{20} = 2^{40}$ 个模式,需要花费的类别标志时间和存储量要求大得惊人,甚至根本无法实现。因此,考虑算法的计算资源消耗和计算复杂度有着重要的实践意义。

3. 鲁棒性

在模式分类中,鲁棒性一般是指系统对输入的训练集的微小改变不敏感,或者对分类器中所涉及的某些参数的微小改变不敏感。一个好的模式分类系统,应该能够忍受少量噪声的影响。也就是说,当输入的样本加入少量的噪声,其输出的结果应该变化不大。鲁棒性的要求主要是考虑在实际的分类问题中,所采集到的数据往往受到噪声的影响,或者存在不可避免的人为误差。

4. 统计稳定性

统计稳定性是指如果使用同一个分布的不同样本集训练分类器,其识别结果应该变化不大。也就是说,模式系统应该反映训练集背后所隐含的分布规律,而不是具体的训练集本身。

1.2 模式分类中的核方法概要

1.2.1 核方法的基本思想

近年来,受统计学习理论尤其是支撑向量机 (Support Vector