

[美] J. 斯克兰斯基 著
G. N. 瓦塞尔

模式分类器和可训练机器



科学出版社

73.82
659

模式分类器和可训练机器

【美】 J. 斯克兰斯基 著
G. N. 瓦塞尔

阎平凡 等 译



科学出版社

1987

8710363

DT83/13
内 容 简 介

本书是系统论述可训练模式分类器的基本理论及设计方法的专著。全书共分七章，主要内容包括训练理论，线性和非线性分类器，局部训练分段线性分类器，对策论方法，随机逼近方法，马尔可夫链训练模型和训练方法，连续状态模型等。本书理论严谨，取材精粹，思路新颖，条理清晰。既注重基本概念又有不少具体实例。为便于读者进一步理解书中的内容，每章后都附有大量习题和参考文献。

本书可作为模式识别、人工智能等专业的研究生和高年级学生的教学参考书，也可供上述领域及应用数学、生物医学工程和计算机应用方面的科技人员参考。

Jack Sklansky, Gustav N. Wiesel

PATTERN CLASSIFIERS AND TRAINABLE
MACHINES

Springer-Verlag New York Inc., 1981

模式分类器和可训练机器

〔美〕J. 斯克兰斯基 G. N. 瓦塞尔 著

阎平凡 等 译

责任编辑 鞠丽娜

科学出版社出版

北京朝阳门内大街137号

中国科学院印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

1987年5月第一版 开本: 787×1092 1/32

1987年5月第一次印刷 印张: 13 1/2

印数: 0001—2,400 字数: 304,000

统一书号: 15031·805

本社书号: 4168-15-8

定 价: 3.15 元

108017-

译者的话

模式识别是六十年代以来随着计算机的发展而发展起来的一门学科。短短的几十年内在理论、方法和应用方面都取得了有意义的成果。随着人们对信息科学及智能机器的深入研究,模式识别必将得到越来越多的人们的重视。近十几年来虽然出版了一些专门论述模式识别方法的书籍,但象J. 斯克兰斯基教授等所写的这本从学习的观点系统地讨论可训练分类器的原理和设计方法方面的书还很少见到。本书中的许多内容是作者自己的研究成果,也是作者长期从事这一领域教学工作的总结。因而它既是一本有相当学术水平的专著,又是一本研究可训练分类器方法的很好的教科书。

本书可作为从事模式识别方面工作的研究生及高年级大学生的参考书,也可作为这一领域的研究人员及实际工作者的参考书。作者在分析问题时使用了较广泛的数学工具,因而它对应用数学工作者也有一定参考价值。同时由于作者长期从事模式识别在生物医学中的应用方面的研究,所举例子很多是结合生物医学方面的问题,因而从事生物医学工程方面研究工作的工程技术人员也会对本书发生兴趣。

参加本书翻译工作的有阎平凡(第一、二、三章,序言及附录),高林(第四、五章),汤之永(第六、七章)。全书由边肇祺教授校阅。翻译过程中得到常迥教授的热情鼓励和帮助,在此表示感谢。

对原书公式中的一些错误，我们在翻译过程中尽量作了订正。

由于译者水平有限，错误和不当之处在所难免，欢迎广大读者批评指正。

译 者

1984年12月

原 序

本书是加州大学伊尔温分校(UOI)从1966年开始的一个研究计划及研究生课程的产物。也被作为研究生教材用于设在波莫纳的加州州立工业大学(Cal. Poly. Pomona)。这一研究计划是 UOI 的模式识别任务的一部分,主要是关于可训练分类器的设计。研究生课程则范围要广些,它还包括了下面一些内容,如:特征选择、聚类分析、数据组选择以及概率密度估计。

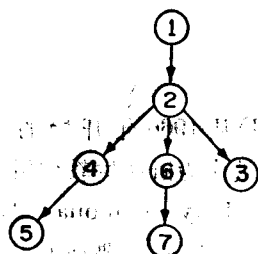
为了尽量减少与其它有关模式识别及分类理论的书在内容上重复,本书选择了一些专门的问题作为重点作较深入的讨论。其中有些材料是至今尚未发表过的。本书可以作为模式分类设计者的指南,也可用作工程或计算机专业研究生的教科书。尽管本书主要是为工程师及计算机科学家们写的,但是心理学家、生物学家、医学科学家、社会科学家都可能对本书发生兴趣。

本书对线性可训练分类器及把这种“线性机器”扩展到非线性分类器给予了特殊注意。这种分类器的设计技术首先在二十世纪五十年代末及六十年代初开始发展,这些技术的某些不足之处促使我们去发现新的设计概念——如加窗训练及最相近的不同类别样本对——并且发展马尔可夫过程的连续状态模型。除了上述结果外本书还包括了至今尚未充分发表的一些属于模式识别与可训练机器领域的课题。

各章的相互关系用下边的图表示。图中每一节点上圆圈中数目是章数,由节点 i 指向节点 j 的一个支表示学习第 j

章要求先学第 6 章，因此对一个较短的课程可以选读部分内容。例如第一、二、四、五章。要进一步削减则可再去掉某

章中的一部分，例如若按一、二、四、五章的顺序，则 2.2 节、2.3 节和 2.4 节可以省略。



每章后面附有为学生用的习题。

感谢加州大学伊尔温分校工程学院的前后三个院长：Robert M. Saunders 博士，James H. Mulligan, Jr. 博士和 Allen R. Stubberud 博士，以及加州州立工业大学教务长 Beamont Davison 博士。由于他们对研究和教学工作的支持和鼓舞，促使我写出了这本书。

我们感谢 Leo Michelotti 博士对第三章叙述的局部训练分段线性分类器所做的贡献，以及他对本书手稿特别是第二章所提的意见。我们对 Phil Merryman 博士表示感谢，第七章的大部分内容是基于他的关于连续状态模型的博士论文。感谢 Lee Kilday 和他的讨论得出了我们所写的选择最小损失工作点的方法，感谢 Ramalingam Chellappa，他仔细阅读了原稿并提出了建设性意见，特别对贝叶斯学习与充分统计量这两部分的贡献。感谢 Carolyn Kimme-Smith, Walter Blume 和 Gloria Frankl 博士对第五章中胸部组织分类器的设计和编程所作的贡献。

感谢我们的学生和合作研究的同事们对本教科书的意见和建设性的建议。

感谢加州大学伊尔温分校、国家科学基金会、空军科学研究所、国家一般医学科学研究所以及 Spectra 研究系统对研究工作的支持，本书中的一些结果就是由于上述研究得出的。

这些支持大部分是根据下述资助: NSF 资助 GK-4226 号, NSF 科学学会成员 60196, 美国空军资助 69-1813 号以及美国公用事业资助 GM-17632.

感谢 Audrey Bennett, Frances Candelori, Deborah S. Germain, Edna Nemetz, Mary Phillips, Dorothy Shearer, Martha Sue Spence Campbell, 以及 Suzanne Costelloe 为本书打字. 感谢为本书画图的 Arlene Sanders.

目 录

第一章 概述	1
1.1 基本定义	2
1.2 可训练分类器与训练理论	3
1.3 假定与符号	5
1.4 训练过程的例子	6
1.5 线性判别函数	8
1.6 特征空间的扩张	14
1.7 二值输入分类器	15
1.8 权空间与特征空间	18
1.9 统计模型	19
1.10 性能评价	23
第二章 线性可分的类别	35
2.1 引言	35
2.2 凸集, 可和性与线性可分性	36
2.3 符号与术语	46
2.4 感知器与比例增量训练方法	49
2.5 固定比值训练方法	53
2.6 一种多类训练方法	55
2.7 用对策论综合	58
2.8 简化技术	69
2.9 例子	72
2.10 梯度下降法	76
2.11 保证收敛的条件	79

2.12 用于设计分类器的梯度下降法	80
2.13 Ho-Kashyap 法	83
第三章 非线性分类器	93
3.1 引言	93
3.2 Φ 分类器	94
3.3 贝叶斯估计: 参数训练	95
3.4 平滑技术: 非参数训练	112
3.5 条形图	113
3.6 Parzen 窗与位函数	121
3.7 节省存储容量	124
3.8 固定基区条形图	128
3.9 样本集与标本	131
3.10 近对立标本对	135
3.11 局部训练分段线性分类器	137
第四章 损失函数和随机逼近	146
4.1 引言	146
4.2 比例增量法损失函数	147
4.3 样本梯度	153
4.4 先验知识的利用	155
4.5 一些重要训练方法的损失函数和梯度	157
4.6 损失函数的比较	176
4.7 不等代价的分类决策	178
4.8 随机逼近	180
4.9 几种不同组成密度的梯度和超平面	191
4.10 结论	198
第五章 用于非可分类别的线性分类器	204
5.1 梯度下降的修正	204
5.2 规范化、原点选择和初始向量	212

5.3	窗训练法	227
5.4	最小均方误差训练法	254
5.5	等错误训练法	264
5.6	不等代价的考虑	273
5.7	应用	274
5.8	总结	276
第六章	关于不可分类别的马尔可夫链训练模型	280
6.1	引言	280
6.2	分析一个随机差分方程的问题	281
6.3	单特征分类器的实例	283
6.4	具有常增量训练的单特征分类器	285
6.5	学习动态特性的基本性质	287
6.6	各态历经性和大范围稳定性	298
6.7	训练-工作进程: 双方式类	305
6.8	最优有限记忆的学习	319
6.9	多维特征空间	337
第七章	连续状态模型	343
7.1	引言	343
7.2	形心方程式	346
7.3	在 $np \leq t < \infty$ 的情况下证明 $\Sigma(n) = O(\rho) \mathbf{U}$	350
7.4	协方差方程式	352
7.5	学习曲线和方差曲线	358
7.6	关于 t 的规格化问题	360
7.7	例题	361
7.8	单特征分类器中学习曲线的形状	367
7.9	等错误和最小错误点间的接近程度如何?	371
7.10	渐近的大范围稳定性	374
附录 A	向量与矩阵	380

A.1 向量不等式及其它向量符号	380
A.2 置换矩阵	381
附录 B 窗训练方法收敛性的证明	385
附录 C 等错误训练法收敛性的证明	390
C.1 $E(\ Z\ ^2 V) < \infty$ 及 $\ E(Z V)\ ^2 < \infty$ 的证明	390
C.2 定理 5.3 的证明	392
参考文献	396
汉英名词对照索引	402

第一章 概 述

正如飞机的发明促进了空气动力学的研究一样，现代数字计算机促进了对智能及学习过程的研究。

智能行为的一种经常出现形式是对数据的拣选及分类。鉴别食物中是否存在有毒物的过程就是这种行为的一个例子。分类过程的特别高级的形式是科学家把对自然界的观测变换为自然界的规律。

由于许多实际分类工作比较困难，因此这一智能行为很自然地常常要依靠一个学习过程。例如，一个放射学家把X光片上的损伤部分区分为良性或恶性的准确度在很大程度上依赖于事先对他的训练。从二十世纪五十年代后期以来，数字计算机技术的发展既推动了模式分类机的技术，也促进了简单形式的人和机器学习过程的数学理论的发展。

模式分类是一个信息转换过程。也就是说一个分类器可以把一组较大且难以理解的数据转换为一组较小且有用的数据，因此计算机就象有生命的有机体那样具有检测和区分模式的能力。已经做成并有效地使用着这种机器(包括少数已商品化了的)的例子有：血球分类器、染色体分析器、空中照片分析器、语音分析器、邮政编码识别机、指纹分析器以及X光片分析器。

本书将叙述几种人工模式分类器设计的数学方法，重点放在学习技术，还要讨论连续状态及马尔可夫链学习模型的数学理论。这两种学习模型对人工分类器及人的决策过程都是适用的。

本章先给出一些基本概念及符号，再从较广泛的意义上讨论一下机器学习的效用，我们先从基本定义开始。

1.1 基本定义

分类器是能把数据分为不同的范畴或类别的一种装置或过程。

可训练的分类器是能根据所接受到的信息不断改进自己性能的一种分类器。令 $\mathcal{C}$ 表示这种分类器， I 表示接受到的信息。 $\mathcal{C}$ 可以是一台机器、一个生物机体或人，也可以是一个生物物质、人机系统、一个商业机构或国家经济体系。 I 经常是 $\mathcal{C}$ 以前性能的数学函数，有时仅仅是一些观测数值及正确分类的特殊序列。通常它是下面三者的组合：一组特殊的观测结果、相应的分类以及 $\mathcal{C}$ 的某些或全部过去性能的函数值。

训练是指按照 I 来调节 $\mathcal{C}$ 的参数这一过程（如 $\mathcal{C}$ 是人，则这些参数通常是学习过程的心理学模型中的一部分），学习方法则是实现训练过程的算法（一般是计算机算法）。

学习是指一个系统的有效性（或性能）从一水平到另一水平的运动，当运动方向使有效性增加时，则学习是正确的（本定义意味着有效性或性能的定量计算方法是存在的）。学习通常伴随有反馈，并能提供使人可以控制自己的技术。机器可以自动适应变化的环境，以及种族可以延续的手段。

智能机器学习的社会应用在以下方面特别明显：

1. 根据学习模型可以建立能够学习用户目的的机器，即使这个目的经过一段时间后改变了，它仍可再学习。正是由于这种反馈过程，使用者可以确信这种机器总是按用户的目的的工作并能避免错误。

2. 根据学习模型可以建立能够学习克服机器自身结构不完善或零部件的故障的一种机器, 这种自组织和自修复的能力对设计并操纵一个复杂的智能机器具有重大意义。

3. 人的学习模型为建立能分担人类学习任务的机器提供基础, 在这里我们设想一个包含两个学与教环节的循环, 在第一个环节中机器去学习总任务中最适于机器去做的方面, 并把学到的内容教给人。在第二个学与教环节中人去学习人做最有效的那部分任务, 基于所学内容再去教给机器。

1.2 可训练分类器与训练理论

在本节中我们要进一步明确对模式分类的观点, 并对训练理论给以定义和讨论。前边把分类器定义为一种能把数据分类的装置, 这些数据通常可看作是特征空间中的一些向量, 此特征空间中的每一点叫做一个特征向量。通常将特征向量 $\mathbf{x}$ 的各分量 x_i 称为特征, 属性或所分析对象的性质。例如分类器可以去分析人的单个细胞中的染色体, 并将染色体分为 23 对(分组)。每一染色体的特征向量可由以下分量组成: 中心部分宽度(腰宽), 臂的平均长度, 凸壳¹⁾重心与整体质心间距离^[3]等。

通常假定在某种意义上某类的特征向量与本类所有特征向量之间的距离比它与其它类全部或大多数特征向量间距离要小, 这就是紧致性假定^[4]。

同一类特征向量在特征空间中占据一个区域, 叫做类区。通常假定所有类区都是有界的, 另一共同假定是各类区互不

1) 一个平面区 $\mathcal{R}$ 的凸壳是包含 $\mathcal{R}$ 的最小凸区, 可形象化地想象先沿 $\mathcal{R}$ 的边界准确地将 $\mathcal{R}$ 切下来, 然后用一伸长的弹性橡皮带包围所切下区的边界, 此橡皮带刚好就处在凸壳的边界。

交叠(不过,对多数实际问题来说,某种程度的交叠是存在的)。当类区不交叠时称这些类是可分的,并具有可分性。如果对每一类区都可用一超平面把它与其余类区分开,则称这些类是线性可分。许多模式分类理论的早期工作都是关于线性可分类别的,近来大量工作则是关于不可分(交叠)及线性不可分类别的。

分类器用一组决策超平面把每一特征向量归属为特征空间中一个特定决策区 $\mathcal{R}_j$ (图1.1)。这种归属可能对应于也可能并不对应于正确的或所要求的分类。一个可训练的分类器可按照观测到的特征向量序列 $\{\mathbf{x}(n)\}$ 来调整决策区 $\{\mathcal{R}_j\}$ 以使被错分的个数变小甚至没有。这时称上述观测量是在学习或训练期得到的。

除特征向量外观测值中也可能包括能把这些特征向量正确分类的信息。如果观测值中包含了正确类别的数据,则称训练是监督的或有教师的。这些正确类别的数据有时叫做补强数据。如果数据中不包含补强数据,则训练是非监督的或无教师的。根据放射学家的诊断结果得知,用计算机把X光片上反映出来的各种心脏扩大现象分类的过程则是监督训练

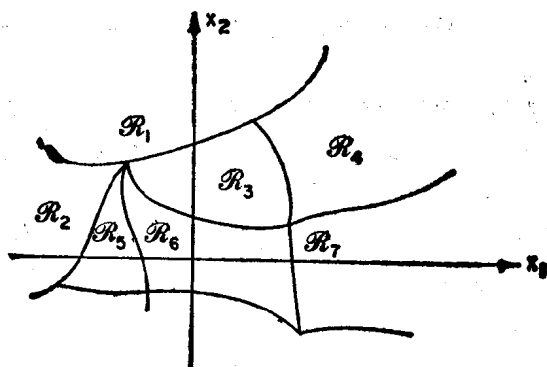


图 1.1 二维特征空间中的一组决策区 $\{\mathcal{R}_j\}$

的一个例子^[5]。而植物学家学习如何把新的植物分为科、属、种等则是非监督训练的例子。

根据所观测的特征向量对决策区族 $\{\mathcal{R}_i\}$ 中的各部分进行调整的过程或算法叫训练方法。每次调整对应于一个或几个特征向量。按照心理学家及精神生理学家的名词，每一次调整连同与此调整有关的观测及补强数据一起叫做一次试验。试验的次数是训练过程长短的指标。

分类器经训练后就用于分析未分类的数据，这一工作阶段叫工作期，而处于训练状态的阶段叫训练期。在训练期作为输入数据的一组特征向量或观测值叫做训练集。有时训练期与工作期可以交叠甚至重合，非监督训练就经常是这种情况。

分类器的决策面通常由以下判别函数 $\{g_i(\mathbf{x})\}$ 确定：

$$\mathcal{R}_i = \{\mathbf{x} | g_i(\mathbf{x}) \geq g_i(\mathbf{x}), \text{ 对所有 } i\} \quad (1.1)$$

当 $g_i(\mathbf{x})$ 取如下形式

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad (1.2)$$

时，则 $g_i(\mathbf{x})$ 是线性判别函数，对应的分类器叫线性分类器。式中 $\mathbf{w}_i^T$ 表示权向量 $\mathbf{w}_i$ 的转置。线性分类器的训练步骤比较简单，对它的了解也较多，在训练期中要调整的权 w_{ij} 的数目也较少，这是它比非线性分类器优越的地方。

训练理论研究训练过程及学习的关系，它包括期望性能^[6]、训练长度、大量训练的稳定性^[7] 以及特征空间的维数之间的关系。例如用此理论的某些部分可以预测可训练分类器的平均性能与训练长度之间的函数关系^[8]。

1.3 假定与符号

本书假定进入分类器的数据是一些统计独立的 d 维随机