

研究生教育书系
信息与电子学科

机器学习方法

Machine Learning Techniques

蒋艳凰 编著
赵强利



电子工业出版社

PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

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

研究生教育书系
信息与电子学科

机器学习方法

蒋艳凰 赵强利 编著

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书比较全面系统地介绍了机器学习的方法和技术,不仅详细阐述了许多经典的学习方法,还讨论了一些有生命力的新理论、新方法。全书共分为 13 章,分别介绍了机器学习的基本概念、最近邻规则、贝叶斯学习、决策树、基于事例推理的学习、关联规则学习、神经网络、支持向量机、遗传算法、集成学习、纠错输出编码、聚类分析、强化学习。各章对原理的叙述力求概念清晰、表达准确,突出理论联系实际,富有启发性,易于理解。

本书可作为高等院校计算机、自动化、电子和通信等专业研究生和高年级本科生的教材和参考书。本书内容对从事人工智能、机器学习、数据挖掘、模式识别等相关领域研究的科技人员具有较好的参考价值。

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

机器学习方法/蒋艳凰等编著. —北京:电子工业出版社,2009. 8
(研究生教育书系)
ISBN 978-7-121-09005-9

I. 机… II. 蒋… III. 机器学习—研究生—教材 IV. TP181

中国版本图书馆 CIP 数据核字(2009)第 091795 号

策划编辑:马 岚

责任编辑:秦淑灵

印 刷:北京市天竺颖华印刷厂

装 订:三河市鑫金马印装有限公司

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:787×980 1/16 印张:18.5 字数:384 千字

印 次:2009 年 8 月第 1 次印刷

印 数:4 000 册 定价:39.00 元

凡所购买电子工业出版社的图书,如有缺损问题,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 zlt@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线:(010) 88258888。

反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为；歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：(010) 88254396；(010) 88258888

传 真：(010) 88254397

E-mail : dbqq@phei.com.cn

通信地址：北京市万寿路 173 信箱

电子工业出版社总编办公室

邮 编：100036

前 言

自从1980年在卡内基-梅隆大学举行了第一届机器学习学术研讨会以来,机器学习的研究工作发展很快,不但在计算机科学的众多领域中尽显身手,而且成为一些交叉学科的重要支撑技术。目前,机器学习技术已广泛应用在各个方面,例如信息安全、搜索引擎、医学诊断、地质勘探、遥感图像识别、DNA序列测序、语音和手写识别、战略游戏、机器人运用、证券市场分析等。当前的热点研究方向数据挖掘,正是利用机器学习技术从海量数据中发现有用的知识。可以说,机器学习已成为人工智能的一个核心研究领域,而且已发展为整个计算机学科中最活跃、应用潜力最明显的领域之一。

根据有无经验知识,机器学习方法可以分为三类:监督学习、非监督学习和强化学习。监督学习的过程实际上是在以往数据的基础上概括总结经验的过程,学习系统的输入是有类别标识的训练数据集,学到的经验集成在学习所得到的分类器中,然后利用这些经验去解决以后遇到的问题。非监督学习是一种无经验的学习,既可以直接对无类别标识的数据进行识别,也可以利用聚类分析的结果预测新样本的类别。强化学习是从环境状态到行为映射的学习,以使系统行为从环境中获得的累积奖赏值最大,它是以环境反馈作为输入且适应环境的机器学习方法。

本书根据上述分类方式,比较全面系统地介绍了机器学习的方法与技术,其内容不仅包括经典的机器学习算法,如贝叶斯学习、决策树、神经网络、聚类分析等,而且涵盖了一些有生命力的新理论、新方法,如支持向量机、纠错输出编码、集成学习、强化学习等。作者在撰写本书时力图清楚地解释关键知识点,并避免冗长难懂的计算,并通过例题进一步加深读者对各种算法的理解。

书中作者的大部分研究工作是在杨学军教授的指导下完成的,在此对杨老师致以由衷的谢意!感谢殷建平教授、陈跃新教授、唐玉华教授、张拥军老师、詹浩博士、易勋博士等,他们为本书的编写和出版提供了许多帮助。感谢电子工业出版社的马岚编辑,她的建议大大提高了本书的质量。感谢国防科技大学计算机学院软件所和计算机所的同事们,在本书撰写期间,他们在工作和生活上给予我们许多关心与支持。

限于作者的水平,书中难免有不妥和谬误之处,敬请读者指正。

作 者

2009年7月

目 录

第1章 绪论	1
1.1 机器学习概念	1
1.2 机器学习系统	2
1.2.1 学习系统模型	2
1.2.2 机器学习系统结构	2
1.3 机器学习方法分类	3
1.3.1 监督学习	4
1.3.2 非监督学习	5
1.3.3 强化学习	6
1.4 一般性定理与规则	6
1.4.1 大多数原则	6
1.4.2 奥卡姆剃刀原理	7
1.4.3 无免费午餐定理	8
1.5 学习算法的评价	9
1.5.1 最短描述长度	9
1.5.2 预测精度分析	10
1.5.3 交叉验证法	11
1.6 本书各章概要	11
第2章 最近邻规则	14
2.1 最近邻分类	14
2.1.1 k -NN 规则的思想	14
2.1.2 k -NN 的距离度量	15
2.1.3 k -NN 算法应用示例	16
2.2 k -NN 算法的缺陷及其改进	17
2.2.1 k -NN 算法的计算复杂度	17
2.2.2 降维法	18
2.2.3 预建结构法	18
2.2.4 训练集裁减法	19
第3章 贝叶斯学习	20
3.1 概率论基础	20
3.1.1 随机事件	20

3.1.2	事件间的关系与运算	21
3.1.3	概率的定义与性质	22
3.1.4	统计概率	22
3.1.5	条件概率	23
3.1.6	概率密度	24
3.1.7	正态分布	24
3.2	贝叶斯定理	24
3.3	贝叶斯定理和概念学习	26
3.4	极大似然和最小误差平方假设	28
3.5	贝叶斯最优分类器	29
3.6	简单贝叶斯分类器	30
3.7	贝叶斯网络	31
3.7.1	贝叶斯网络基本概念	32
3.7.2	因果关系网	33
3.7.3	贝叶斯网络	33
3.7.4	联合概率	35
3.7.5	D分离	35
3.7.6	贝叶斯网络的推理模式	38
3.8	主观贝叶斯方法	40
3.8.1	规则的不确定性	41
3.8.2	证据的不确定性	43
3.8.3	推理计算	44
3.9	贝叶斯学习的优缺点	47
第4章	决策树	49
4.1	决策树的创建	49
4.1.1	分而治之的思想	49
4.1.2	决策树生成算法	50
4.2	分枝划分标准	51
4.2.1	测试条件	51
4.2.2	直接划分法	52
4.2.3	信息熵增益	53
4.2.4	增益比	57
4.2.5	Gini 系数	59
4.2.6	最短距离划分	60
4.2.7	最短描述长度	62
4.3	连续属性离散化	65

4.3.1	离散化问题描述	66
4.3.2	离散化方法分类	67
4.3.3	典型离散化处理过程	67
4.3.4	常用离散化方法介绍	68
4.3.5	离散化方法的一种归类模式	71
4.4	决策树剪枝	72
4.4.1	剪枝问题描述	72
4.4.2	预剪枝技术	73
4.4.3	后剪枝技术	74
4.5	动态离散化方法 RCAT	78
4.5.1	RCAT 算法思想	78
4.5.2	属性变换	79
4.5.3	区间离散化	80
4.5.4	区间边缘优化	81
4.6	二分决策树系统 Btrees	84
4.6.1	二分树林结构	84
4.6.2	多类问题扩展	85
4.6.3	决策树剪枝	86
4.6.4	实验结果	87
4.7	决策树规则化	94
4.7.1	问题提出	94
4.7.2	单个规则的泛化	96
4.7.3	规则集的分类行为	100
第 5 章	基于事例推理的学习	105
5.1	CBR 概念的引入	105
5.1.1	基于规则推理系统的缺陷	105
5.1.2	CBR 的概念	106
5.1.3	CBR 系统的特点	107
5.2	CBR 的描述性框架	108
5.2.1	CBR 循环	108
5.2.2	CBR 任务层次	109
5.3	CBR 方法描述	110
5.3.1	事例表示	110
5.3.2	事例检索	113
5.3.3	事例重用	114
5.3.4	事例修订	115

5.3.5	事例再生	116
5.4	演绎检索	116
5.4.1	基于决策树的事例表示	117
5.4.2	计算新事例索引值	118
5.4.3	选择事例	118
5.5	最近邻检索	120
5.5.1	基于加权的事例表示	120
5.5.2	相似度计算	120
第6章	关联规则学习	123
6.1	相关概念介绍	123
6.1.1	关联规则的基本概念	123
6.1.2	关联规则的分类	125
6.2	关联规则学习	125
6.2.1	Apriori 算法	126
6.2.2	FP-Growth 算法	130
6.3	多层关联规则学习	134
6.3.1	多层关联规则概念	134
6.3.2	多层关联规则学习	135
6.4	关联规则与相关性分析	138
6.4.1	问题的提出	138
6.4.2	相关性分析	139
6.4.3	相关规则学习	140
6.5	关联学习中的约束	140
第7章	神经网络	142
7.1	神经网络概述	142
7.1.1	神经网络基本模型	142
7.1.2	神经网络学习方法分类	144
7.2	常用神经网络介绍	145
7.2.1	感知机	145
7.2.2	BP 神经网络	146
7.2.3	径向基函数网络	147
7.2.4	Hopfield 网络	149
7.2.5	玻尔兹曼机	151
7.2.6	自组织特征映射网络	153
7.3	BPNN 的快速学习算法	155
7.3.1	动量项学习算法	155

7.3.2	RPROP 学习算法	156
7.4	BPNN 的结构化扩展	157
7.4.1	结构化神经网络 SNN	158
7.4.2	网络行为解释	159
第8章	支持向量机	163
8.1	核函数和特征空间	163
8.1.1	特征空间中的学习	163
8.1.2	核函数的定义	165
8.1.3	核函数的性质	166
8.2	构造核函数	168
8.2.1	由核函数构造核函数	168
8.2.2	由特征构造核函数	168
8.3	常用的核函数	169
8.4	VC 理论	170
8.4.1	PAC 学习模型	170
8.4.2	VC 理论简介	171
8.5	SVM 基本原理	172
8.5.1	SVM 方法的特点	173
8.5.2	线性情况	173
8.5.3	非线性情况	176
8.6	支持向量机算法扩展	177
8.6.1	多类别问题的扩展	177
8.6.2	海量样本数据的处理	178
8.7	支持向量机应用示例	179
第9章	遗传算法	182
9.1	基本遗传算法	182
9.1.1	编码	182
9.1.2	适应度函数	183
9.1.3	遗传算子	183
9.1.4	运行参数	187
9.2	执行过程及应用示例	187
9.2.1	遗传算法的执行过程	187
9.2.2	应用示例	188
9.3	遗传算法的数学基础	189
9.3.1	模式定理	189
9.3.2	建筑块假说	193

9.3.3 内在并行性	194
9.4 遗传算法的收敛性分析	195
9.5 遗传算法的局限性及其改进措施	195
9.5.1 遗传算法的局限性	195
9.5.2 遗传算法的改进措施	197
9.6 遗传算法的特点	199
9.7 遗传规划	200
9.7.1 遗传规划概念	200
9.7.2 遗传规划的特点	201
9.7.3 遗传规划的发展动向	202
9.8 基于遗传算法的机器分类	203
9.8.1 遗传进化与机器分类	203
9.8.2 遗传算法应用示例	204
9.8.3 遗传规划应用示例	205
第10章 集成学习	207
10.1 集成学习的概念	207
10.2 基本分类器的构造	208
10.2.1 Bagging	208
10.2.2 Adaboost	210
10.2.3 纠错输出编码	211
10.2.4 随机因素法	211
10.2.5 异构的学习方法	211
10.3 分类器集成方法	212
10.3.1 投票法	212
10.3.2 栈式集成法	213
10.3.3 级联组合法	214
10.3.4 算法相关法	215
10.4 选择性集成	215
10.4.1 选择性集成理论	216
10.4.2 GASEN 算法	217
10.4.3 前向选择算法	218
10.5 集成学习的性能	220
第11章 基于纠错编码的机器学习	222
11.1 纠错输出编码	222
11.1.1 纠错输出编码的思想	222
11.1.2 纠错输出码的特性分析	224

11.1.3	常用的编码方法及其特点	225
11.2	BCH 编码算法	227
11.2.1	相关定义	227
11.2.2	循环码及其构造方法	228
11.2.3	BCH 码的设计	229
11.3	搜索编码算法	231
11.3.1	搜索编码算法描述	231
11.3.2	搜索输出码的性质	232
11.4	基于纠错输出编码的监督分类技术	237
11.4.1	纠错输出编码与分类精度	237
11.4.2	基于搜索编码的监督分类技术	238
11.4.3	决策方法	239
11.5	应用示例	240
11.5.1	搜索编码法在简单贝叶斯分类中的应用	241
11.5.2	搜索编码法在 BP 神经网络中的应用	244
第 12 章	聚类分析	246
12.1	聚类分析概述	246
12.1.1	聚类分析的基本概念	246
12.1.2	相似性测量	246
12.1.3	聚类评价准则	248
12.2	类别数目未知时的聚类方法	249
12.2.1	自适应样本集构造法	250
12.2.2	Batchelor 和 Wilkins 算法	251
12.3	类别数目已知时的聚类方法	252
12.3.1	层次聚类法	252
12.3.2	k -均值聚类算法	253
12.3.3	ISODATA 算法	255
12.4	基于核构造的聚类法	257
12.5	模糊聚类	258
12.5.1	隶属度函数	258
12.5.2	普通的等价关系	259
12.5.3	基于隶属度的 FCM 算法	260
12.5.4	基于模糊等价关系的聚类算法	261
12.5.5	基于最大模糊支撑树的聚类算法	264
第 13 章	强化学习	266
13.1	强化学习的概念	266

13.2 强化学习的基本知识	266
13.2.1 问题假设	266
13.2.2 目标函数	267
13.2.3 问题类型	268
13.2.4 强化信号	269
13.2.5 状态值函数	269
13.3 TD 算法	270
13.4 Q -学习	271
13.4.1 Q -学习算法	271
13.4.2 应用示例	272
13.5 强化学习的局限性	274
附录 A 数据集描述	275
参考文献	278

第 1 章 绪 论

机器学习最初的研究动机是让计算机系统具有人的学习能力,以便实现人工智能。随着计算机技术的日益发展,人类收集数据、存储数据的能力有了极大提高,无论科学研究还是社会生活,各个领域都积累了大量的数据。如何对这些数据进行分析,从中发现蕴涵的规律,几乎成为所有领域的共同需求,这使得机器学习技术受到日益广泛的关注。目前,机器学习成为继专家系统之后人工智能应用的又一重要研究领域,且已发展为计算机学科中最活跃、应用潜力最明显的领域之一。

1.1 机器学习概念

一个人从婴幼儿开始就逐渐发展语言能力,这个过程包含了各种学习过程:记忆、模仿、归纳、类比和发现等。我们能够轻而易举地识别人脸,分辨语音,阅读文字,判定水果的成熟,这都是人类学习行为在起作用。

我们以一个例子来简单形象地描述一个学习过程。考虑一个孩子学习一种语言模式,首先,孩子听见妈妈说了一句话“爷爷已经吃完了”,这句话引起了他的注意;然后他逐渐发现大人说“…听完了”、“…说完了”、“…写完了”、“…算完了”等,过不了多久,小孩自己就能够说出“…看完了”、“…画完了”之类的话,这说明小孩已经会用“… <动词>完了”的语言模式来表示某一动作的结束。

图 1.1 给出了学习这一语言模式的过程。

学习是人类具有的一种重要智能行为,我们可以列出很多学习的例子,但是关于学习的明确定义,长期以来却众说纷纭,社会学家、逻辑学家和心理学家各有不同的看法。按照人工智能大师 H. A. Simon 的观点,“学习”就是系统在不断重复的工作中对本身能力的增强或者改进,结果是系统在下次执行同样任务或类似任务时,比现在做得更好或效率更高。简而言之,“学习”指系统改进其性能的任何过程。

正是由于人类具有强大的学习能力,人们才会很自然地想到去设计和制造一台具有这种能力的机器,然而,貌似简单的识别行为背后却是非常复杂的处理机制,这种处理机制就是机器学习理论。

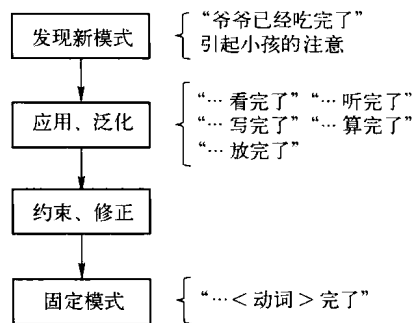


图 1.1 学习新语言模式的过程

什么叫做机器学习？这一概念至今也没有统一的定义。一种广义的定义是：机器学习是研究如何使用机器来模拟人类学习活动的一门学科。稍微严格的定义是：机器学习是一门研究机器获取新知识和新技能，并识别现有知识的学问。

1.2 机器学习系统

1.2.1 学习系统模型

一个问题求解系统可以定义为三元组的形式： (I, T, O) ，其中， I 为问题的输入空间； O 为任务的求解空间； T 为问题的求解机制，从某种意义上讲，它是一个 $I \rightarrow O$ 的函数。

按照 Tom M. Mitchell 的观点，给定某类任务 T 、与 T 相关的性能 P 和经验 E ，计算机程序从经验 E 中学习，用学习的结果改善性能 P ，从而实现自我完善，则称该程序具有学习能力。因此，一个机器学习系统就是一个问题求解系统，但是又有自身的特殊性，机器学习系统可描述为四元组的形式：

$$(E_m, T_m, O_m, V)$$

其中，

- E_m ：训练经验，是学习系统的输入，可表示为 (I, O) ， I 为问题求解系统的输入， O 为其输出；
- T_m ：学习任务，与问题求解任务密切相关，如学习问题求解系统的控制参数、问题求解策略、分类规则和一般假设等。
- O_m ：学习系统的输出，问题求解执行部件所使用的假设。
- V ：关于问题求解系统性能 P 的评估， P 可以是全局的或局部的。学习系统的目标是将 O_m 用于问题求解系统时，性能 P 得到改善。

例如，对于西洋跳棋学习问题， T 为西洋跳棋的走子策略， E 为对弈时的走子序列及输赢结果， P 为比赛中击败对手的百分比。对于手写字符识别的学习问题， T 为图像与字符集的映射（分类规则）， E 为已知分类的手写字符图像的数据库， P 为分类的正确率。

1.2.2 机器学习系统结构

图 1.2 给出了机器学习系统的基本结构。环境向系统的学习部分提供某些信息作为训练数据，学习部分利用选取的学习算法对获得的信息进行学习，学习过程中可能对环境产生影响，并从环境中获得新的训练数据。学习结束后，获得的知识根据不同的学习算法表示成不同的形式，如规则库、表达式、表格或神经网络等，我们统称其

为知识库。执行部分主要是利用学习的结果对待识别数据进行处理,获得相应的识别结果,同时把执行结果反馈给学习部分,以便进一步学习,改进系统性能。

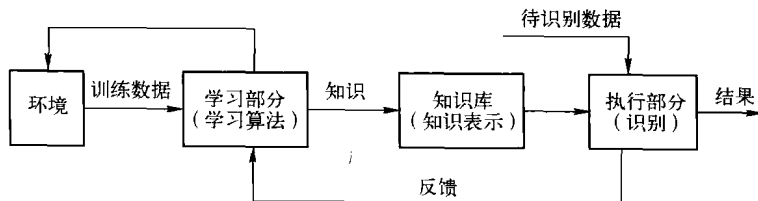


图 1.2 机器学习系统的基本结构

影响学习系统设计的最重要因素是环境向系统提供的信息,即学习模型中的 E_m , E_m 的质量直接影响学习结果。知识库里存放的是指导执行部分动作的一般原则,但环境向学习系统提供的信息却是各种各样的。如果信息的质量比较高,与一般原则的差别比较小,则学习部分比较容易处理;如果向学习系统提供的是杂乱无章的指导执行动作的具体信息,则学习系统需要在获得足够数据之后,删除不必要的细节,进行总结推广,形成指导动作的一般原则,这样学习部分的任务就比较繁重,设计起来也较为困难。

由于学习系统获得的知识往往是不完全的,它总结出来的规律可能正确,也可能不正确,这要通过执行效果加以检验。正确的学习结果能使系统的效能提高,应予保留;不正确的学习结果应予修改或从知识库中删除。

知识库是影响学习系统设计的第二个因素。知识的表示有多种形式,这些表示各有各的特点,在选择表示方式时要兼顾以下 4 个方面:①表达能力强;②易于推理;③容易修改知识库;④知识表示易于扩展。知识库的表示与具体学习算法相关。

执行部分是整个学习系统的核心,它除了包括学习结果的具体应用外,还涵盖了对学习结果 O_m 的评价 V 。与执行部分有关的问题有 3 个:复杂性、反馈能力和透明性。复杂性是指执行部分的复杂度,例如,一个简单的任务可能是按照单一的概念或规则进行预测,比较复杂的任务可能涉及多个概念和上百条规则。反馈能力是指如何评价学习部分获取的假设,通常是由教师(见第 1.3.1 节)提出外部执行标准,观察执行部分的结果与该标准的差异,并将相应结果返回给学习部分,以便进一步完善学习假设。透明性即要求根据执行部分的结果可以很容易地对知识库进行评价,例如,下完一盘棋后根据最后的输赢来判断每一步棋的优劣是很困难的,但若记录了每一步棋后的局势,根据局势来判断每一步棋的优劣则相对较容易和直观。

1.3 机器学习方法分类

根据输入的训练数据中包含的经验情况,可以将机器学习方法分为以下三类:监督学习 (supervised learning)、非监督学习 (unsupervised learning) 和强化学习 (rein-

forcement learning)。监督学习的训练数据中包括样本的真实输出,因此可以利用真实输出来评价学习结果的优劣,如医学诊断问题中已知训练样本的疾病类别,市场价格预测问题中已知训练样本的真实价格等。非监督学习的训练数据中没有经验知识,学习过程中需要建立对输出的评价方法。强化学习的经验知识来自学习过程中环境的反馈,有时有一定的延时。下面对这三种学习方法进行简单描述。

1.3.1 监督学习

监督学习又称为有教师学习,所谓“教师”即对一组给定的输入提供应有的输出结果,也就是对训练样本集中的每个样本都能提供其真实输出,有时还提供预测错误的代价。监督学习是最重要的机器学习方法,已经产生了许多经典的学习算法,如决策树、人工神经网络、贝叶斯网络和支持向量机等。

我们将样本 X 表示为属性向量的形式,即

$$X = (x_1, x_2, \dots, x_l)$$

其中,元素 $x_j (1 \leq j \leq l)$ 为样本 X 在第 j 个属性上的值; l 为属性的个数,各属性可以是离散或连续属性。另外,定义样本的属性空间为 \mathbf{X} ,监督学习的求解空间为 \mathbf{Y} , \mathbf{N} 为自然数集合。根据上述符号定义,可将监督学习问题描述为:给定一个已知类别的样本集合 $LS = \{(X_i, y_i) \mid X_i \in \mathbf{X}, y_i \in \mathbf{Y}, i \in \mathbf{N}\}$, LS 确定了属性向量 X 与其输出 y 之间的某种未知函数关系 $y = f(X)$,利用监督学习算法 \mathfrak{S} 对 LS 中的样本进行学习,得到学习器 $\tilde{f} = \tilde{f}(LS)$,用于近似表示该未知函数。

监督学习可以用于分类问题和回归问题。对于分类问题,输出空间 \mathbf{Y} 是一个类别的集合 $CS = \{c_1, c_2, \dots, c_m\}$, m 表示类别的个数,每个元素 c_i 对应样本可能所属的某个类别,例如,医学诊断问题就是一个分类问题;对于回归问题,输出空间 \mathbf{Y} 中的元素一般是一个实数值,如市场价格预测问题。由于这两类问题的求解方法基本相同,在本书的讨论中,对各种监督学习算法的描述主要针对分类问题展开,并将其学习结果称为分类器(classifier)。监督学习结束后,可利用该分类器对未知类别的样本进行类别预测。

监督学习系统通常涉及如下几个步骤:数据采集、属性选择和提取、模型选择、监督训练、分类器评价,如图 1.3 所示。由于每个步骤都会影响到监督学习的性能,故这些步骤可能需要重复,以便获得更好的学习结果。

对于一个学习系统而言,为了保证系统工作时具有良好的性能,必须采集足够多的原始样本数据。数据采集与具体的应用领域相关,它整个系统总费用中占的比重是相当惊人的。

根据监督学习问题领域的性质,选择具有明显区分意义的属性,是系统设计过程中非常关键的问题。在属性选择和提取过程中,我们总是希望发现那些容易提取,对不相关变形保持不变,对噪声不敏感,以及对区分不同类别的样本很有效的属性集。