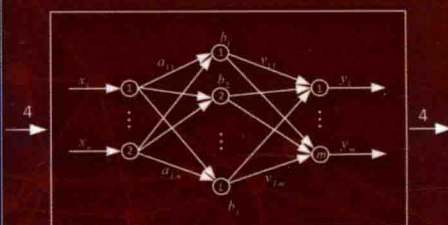


基于神经网络的 监督和半监督学习方法与 遥感图像智能解译

韩敏 编著



JIYU SHENJING WANGLUO DE
JIANDU HE BANJIANDU XUEXI FANGFA YU
YAOGAN TUXIANG ZHINENG JIEYI



中国水利水电出版社
www.waterpub.com.cn

基于神经网络的 监督和半监督学习方法与 遥感图像智能解译

韩敏 编著



中国水利水电出版社
www.waterpub.com.cn

内 容 提 要

本书从机器学习的理论与应用出发,重点对基于人工神经网络的学习方法进行了阐述,并对监督和半监督两种学习形式分别进行了深入的探讨。在理论研究的基础之上,针对遥感图像解译的实际问题,给出了多个仿真实例,进而说明本书理论在实际应用中的有效性。

本书可供从事神经网络、机器学习等方面的高校师生和科研工作者参考,同时也可以为从事遥感图像处理与应用、智能信息处理等领域的工程技术人员提供帮助。

图书在版编目(CIP)数据

基于神经网络的监督和半监督学习方法与遥感图像智能解译 / 韩敏编著. — 北京: 中国水利水电出版社, 2015. 12

ISBN 978-7-5170-3965-5

I. ①基… II. ①韩… III. ①机器学习—研究 IV. ①TP181

中国版本图书馆CIP数据核字(2015)第317471号

书 名	基于神经网络的监督和半监督学习方法与 遥感图像智能解译
作 者	韩敏 编著
出版发行	中国水利水电出版社 (北京市海淀区玉渊潭南路1号D座 100038) 网址: www.waterpub.com.cn E-mail: sales@waterpub.com.cn 电话: (010) 68367658 (发行部)
经 售	北京科水图书销售中心(零售) 电话: (010) 88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排 版	中国水利水电出版社微机排版中心
印 刷	北京嘉恒彩色印刷有限责任公司
规 格	170mm×240mm 16开本 19印张 362千字
版 次	2015年12月第1版 2015年12月第1次印刷
印 数	0001—1000册
定 价	58.00元

凡购买我社图书,如有缺页、倒页、脱页的,本社发行部负责调换

版权所有·侵权必究

前 言

机器学习是人工智能的一个核心领域，目前已经成为整个计算机领域中最活跃、最具应用潜力的领域之一，且扮演着日益重要的角色。近年来，机器学习领域的研究在理论和应用上都取得了巨大的发展，计算学习理论、概率统计方法、神经网络理论以及模式识别技术等众多原本各自独立的研究领域和问题逐渐融合成一个统一的理论方法体系；同时，机器学习技术的应用也得到了扩展，在解决传统的语音辨识、手写汉字识别以及人脸识别等问题之外，还被广泛应用于包括数据库知识挖掘问题、语言处理问题、优化组合问题、机器人控制问题在内的许多新问题。

本书在学习问题的多种假设空间表示方法中，重点对基于人工神经网络的学习方法进行了阐述，对监督和半监督两种学习形式分别进行了深入的讨论，并将理论方法应用于遥感图像的智能解译，有效解决了土地利用或覆盖研究中的诸多实际问题。作者长期致力于人工神经网络及其学习算法的研究，本书正是作者多年来研究积累的结晶，同时收集了国内外众多学者的相关著作和研究成果，力求理论完整、层次清晰、内容深入浅出。希望本书能为从事人工神经网络、机器学习、遥感图像处理等方面研究的科学工作者提供帮助。

本书内容共分为8章。第1章在机器学习概念下对监督和半监督两种学习形式进行了简单介绍，并重点对基于人工神经网络的遥感图像分类方法进行了阐述。第2章介绍神经网络的基本概念以及几种主流的神经网络方法，包括径向基函数神经网络、支持向量机、极限学习机和ARTMAP神经网络的结构和学习算法。其余6章可分为两大部分，第3~第5章为监督学习方法的理论分析与应用，第6~第8章为半监督学习方法的深入讨论与应用介绍。其中，第3章阐述了集成径向基函数网络的监督学习方法，基于径向基函数分别

给出随机森林、轮转森林、反向标定训练数据以及混合集成方法的描述和遥感图像仿真实例。第4章为支持向量机网络监督学习方法的介绍，阐述了支持向量机相关的理论基础和基本原理，在此基础上分别给出改进的支持向量机方法和优化的支持向量机算法。第5章介绍了极限学习机的监督学习方法，在基本原理的基础上给出一种改进的极限学习机方法以及该改进方法在实际应用中的具体分析，进而介绍了其他几种改进模型。第6章阐述基于ARTMAP自适应网络的半监督学习方法，同时以遥感图像分类问题为例构建案例推理ARTMAP分类系统。第7章利用直推式学习理论将支持向量机模型扩展到半监督学习领域，并对仿真结果作了详细的分析。第8章以极限学习机为半监督学习的具体算法进行展开，并以仿真实例作为辅助说明。

在本书出版之际，作者要感谢参与本书相关内容的研究、书稿编写及核对工作的学生们，他们是唐晓亮博士、姚为博士、王新迎博士、朱新荣、刘贲、杨雪、张檬、杨亚飞、边茂松、张雅美、张成坤。全书由周阳做了认真细致的统稿。借此机会，谨向长期以来国内外所有给予作者关心、支持和鼓励的人士及家人致以最诚挚的谢意。

本书由大连市人民政府、国家自然科学基金（61374154）、国家重点基础研究发展规划项目（2013CB430403）资助出版，在此表示感谢。

由于作者水平有限，书中难免存在错误和不足之处，热诚期待读者们的批评指正。

作 者

2015年6月

目 录

前言

第1章 概述	1
1.1 监督学习方法	2
1.1.1 监督学习概述	2
1.1.2 监督学习方法简介	3
1.2 半监督学习方法	9
1.2.1 半监督学习概述	9
1.2.2 半监督学习方法简介	11
1.3 基于神经网络的遥感图像分类	17
1.3.1 遥感图像分类技术概述	17
1.3.2 基于神经网络的遥感图像分类概述	19
1.4 本书结构安排	29
1.5 小结	30
参考文献	31
第2章 常用神经网络基础知识	40
2.1 神经网络	40
2.1.1 神经网络概念和基本结构	40
2.1.2 神经网络的发展	41
2.1.3 神经网络的基本模型和特点	42
2.2 径向基函数神经网络	43
2.2.1 径向基函数神经网络概述	44
2.2.2 模式可分性的 Cover 定理	45
2.2.3 径向基函数神经网络数学描述	46
2.3 支持向量机	48
2.3.1 支持向量机理论基础	48
2.3.2 常用支持向量机学习方法	54
2.4 极限学习机	58
2.4.1 极限学习机的运行机制	58

2.4.2 极限学习机算法存在的问题	60
2.5 ARTMAP 神经网络	62
2.5.1 ARTMAP 神经网络	62
2.5.2 模糊 ARTMAP 神经网络	64
2.5.3 贝叶斯 ARTMAP 神经网络	68
2.6 小结	71
参考文献	71
第3章 基于径向基函数神经网络的集成监督学习方法	75
3.1 集成监督学习方法	75
3.1.1 集成监督学习方法	75
3.1.2 常用的集成监督学习方法	77
3.2 基于径向基函数神经网络的随机森林监督学习方法	79
3.2.1 随机森林集成监督学习方法	80
3.2.2 基于径向基函数神经网络的随机森林监督学习方法	81
3.2.3 仿真实例	83
3.3 基于径向基函数神经网络的轮转森林监督学习方法	88
3.3.1 轮转森林集成监督学习方法	88
3.3.2 基于径向基函数神经网络的轮转森林监督学习方法	90
3.3.3 仿真实例	91
3.4 基于径向基函数神经网络的反向标定训练数据监督学习方法	94
3.4.1 反向标定合成训练数据集成监督学习方法	95
3.4.2 基于径向基函数神经网络的反向标定训练监督学习方法	95
3.4.3 仿真实例	98
3.5 基于径向基函数神经网络的混合集成监督学习方法	102
3.5.1 混合集成监督学习方法	102
3.5.2 基于径向基函数神经网络的混合集成监督学习方法	102
3.5.3 仿真实例	105
3.6 小结	108
参考文献	108
第4章 基于支持向量机的监督学习方法	112
4.1 改进支持向量机监督学习方法	112
4.1.1 C 支持向量机 (C-SVM) 及其改进算法	113
4.1.2 ν 支持向量机 (ν -SVM) 及其改进算法	116
4.1.3 其他支持向量机监督学习方法	118

4.1.4	支持向量机监督学习方法分析与比较	122
4.2	求解支持向量机的优化方法	124
4.2.1	常用的支持向量机求解算法	125
4.2.2	求解支持向量机的四重序列解析优化算法	129
4.2.3	求解支持向量机的分块序列最小化优化方法	133
4.3	基于近邻协同的支持向量机高光谱遥感图像分类	136
4.3.1	基于近邻协同的支持向量机分类方法	136
4.3.2	NC-SVM 高光谱遥感图像仿真试验	139
4.4	小结	143
	参考文献	143
第5章	基于极限学习机的监督学习方法	146
5.1	基于极限学习机的监督学习改进方法	146
5.1.1	分段训练极限学习机	146
5.1.2	迭代-解析极限学习机	149
5.1.3	集成极限学习机	151
5.1.4	正余弦极限学习机	152
5.2	极限学习机及其改进方法的仿真实例	156
5.2.1	极限学习机的仿真试验	156
5.2.2	分段训练极限学习机的仿真试验	159
5.2.3	集成极限学习机和迭代-解析极限学习机的仿真试验	160
5.2.4	正余弦极限学习机的仿真试验	162
5.3	基于局部 Lanczos 双对角化的极限学习机算法	163
5.3.1	局部 Lanczos 双对角化方法	163
5.3.2	基于局部 Lanczos 双对角化的最小二乘近似解	164
5.3.3	基于局部 Lanczos 双对角化的极限学习机学习算法	166
5.3.4	基于局部 Lanczos 双对角化的极限学习机算法的分析	166
5.4	基于局部 Lanczos 双对角化极限学习机算法的应用实例	170
5.4.1	遥感图像分类的应用实例	171
5.4.2	遥感图像的特征参数提取	173
5.4.3	基于特征融合的分类方法	175
5.4.4	基于分类结果的土地利用变化分析	182
5.5	小结	183
	参考文献	183

第 6 章 基于 ARTMAP 神经网络的半监督学习方法	186
6.1 基于贝叶斯 ARTMAP 神经网络的半监督学习方法与应用	186
6.1.1 基于贝叶斯 ARTMAP 神经网络的半监督学习方法	186
6.1.2 基于贝叶斯 ARTMAP 神经网络的半监督学习方法的收敛性	191
6.1.3 仿真实例	192
6.2 基于 ARTMAP 神经网络的案例推理广义半监督学习方法与应用	198
6.2.1 基本结构及实现流程	198
6.2.2 学习方法及操作步骤	202
6.2.3 仿真实例	209
6.3 小结	213
参考文献	213
第 7 章 基于支持向量机的半监督学习方法与应用	216
7.1 基于直推式支持向量机的半监督学习方法	216
7.1.1 直推式半监督学习方法	216
7.1.2 直推式支持向量机	217
7.1.3 仿真实例	219
7.2 基于渐进直推式支持向量机的半监督学习方法与应用	222
7.2.1 基于渐进直推式支持向量机的半监督学习方法	222
7.2.2 仿真实例	227
7.3 基于改进渐进直推式支持向量机的半监督学习方法与应用	231
7.3.1 基于改进渐进直推式支持向量机的半监督学习方法	232
7.3.2 基于置信度的渐进直推式支持向量机半监督学习方法	235
7.3.3 仿真实例	239
7.4 半监督最小二乘支持向量机方法与应用	244
7.4.1 基于最小二乘支持向量机的半监督学习方法	244
7.4.2 仿真实例	250
7.5 小结	253
参考文献	254
第 8 章 基于极限学习机的半监督学习方法与应用	257
8.1 基于极限学习机的半监督学习方法	257
8.1.1 基于极限学习机的半监督学习方法理论基础	257
8.1.2 基于极限学习机的半监督学习方法实现过程	260
8.1.3 仿真实例	262

8.2 基于协同训练的极限学习机半监督学习方法	265
8.2.1 在线序列极限学习机	265
8.2.2 协同训练方法	269
8.2.3 基于协同训练的极限学习机半监督学习方法	270
8.2.4 仿真实例	271
8.3 基于三重可逆训练的极限学习机增量式半监督学习方法	272
8.3.1 可逆极限学习机	272
8.3.2 三重训练方法	276
8.3.3 三重可逆极限学习机半监督学习算法	278
8.3.4 仿真实例	280
8.4 小结	290
参考文献	290

第 1 章

概 述

机器学习 (Machine Learning) 是研究如何使计算机程序随着经验积累自动提高性能的一门学科, 其目标是使计算机程序利用样本数据或以往的经验来解决给定的问题^[1]。它是人工智能 (Artificial Intelligence) 的一个核心领域, 并已经成为整个计算机领域中最活跃、应用潜力最明显的领域之一。机器学习在解决传统的语音辨识、手写汉字以及人脸识别等问题之余, 还被广泛应用于包括遥感图像智能分类、数据库知识挖掘问题、自然语言处理问题、机器人控制问题等在内的许多新问题^[2,3]。而神经网络以其自组织性、自学习性、并行性、容错性、高度非线性、高度鲁棒性、对任意函数的任意精度逼近能力和不透明映射性能, 成为机器学习领域最热的焦点, 可以说神经网络是机器学习研究、开发和应用最为活跃的分支之一。

根据输入的训练样本中包含的经验情况, 机器学习可分为监督学习 (Supervised Learning)、无监督学习 (Unsupervised Learning) 和半监督学习 (Semi-Supervised Learning)。其中监督学习是最常见的机器学习方法, 也是训练神经网络常使用的一项技术, 监督学习过程中所使用的样本均是有标记的 (Labeled), 这里的“标记” (Label) 是指样本具有与输入明确对应的输出。神经网络依赖于这些确定的标记信息判断网络的误差, 然后据此调整网络权值以达到适应系统的目的。随着数据收集和存储技术的飞速发展, 未标记 (Unlabeled) 样本的收集已变得相当容易, 而有标记样本的获取仍相对较为困难。例如, 遥感图像类标的获取通常需要实地考察和通过经验进行目视解译来对样本进行标记, 因此往往需要大量的人力和物力^[4]。无监督学习虽然不需要标记样本, 但是缺乏先验知识的有效引导, 模型的准确性难以保证。此外, 只使用少量的有标记样本所训练出的神经网络往往泛化能力不强, 而且仅对少量有标记样本加以利用, 而忽略大量未标记样本, 是对数据资源的极大浪费。因此, 在有标记样本较少时, 利用大量的未标记样本来改善学习性能的半监督学习已成为当前机器学习研究中备受瞩目的内容之一。在半监督学习中, 采用神经网络作为学习器有助于提升半监督学习方法的自适应调节能力, 并降低其对先验知识的依赖程度^[5]。

本书从监督学习方法、半监督学习方法和神经网络的基本概念出发, 分析当前实际问题的需要, 以神经网络作为基础, 以监督学习和半监督学习方法为



工具,介绍基于神经网络的监督和半监督学习的最新理论研究成果,并着重介绍其在遥感解译领域的应用实例,试图为机器学习和遥感图像解译研究领域引入新思路和新方法。

1.1 监督学习方法

监督学习也称有教师学习,是最重要也最流行的机器学习方法之一,它可以从有标记的训练数据集中学习出一个函数,并依此函数对新的数据进行预测。监督学习的训练数据集由输入(特征)和有标记的输出(目标)组成的样本对构成。其中,有标记的输出(目标)也称为监督信号(Supervised Signal),是监督学习方法得名的由来。本节将首先对监督学习方法进行概述,然后简单介绍几种主流和热门的监督学习方法。

1.1.1 监督学习概述

在监督学习问题中,训练样本集中每个样本都有确定的输出标记,机器使用这些有标记的样本,根据输入的属性数据和对应的输出标记进行学习,产生模型(Model),该模型可对测试样本的标记进行预测。这里样本集可以表示为 $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, 其中 m 为样本个数, $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ ($i=1, 2, \dots, d$) 表示样本的属性,每个样本均由 d 个属性进行描述, x_{ij} 对应第 i 个样本的第 j 个属性,这些属性既可以是离散的也可以是连续的; $y_i \in Y$ ($i=1, \dots, m$) 表示样本 x_i 的预期输出或称为监督信号, Y 为输出空间,用函数 $y=f(x)$ 表示,期望输出 y 与属性向量 x 之间存在的非线性映射关系,该函数的输出可以是一个分类标记(此类问题称为分类问题),也可以是一个连续的值(此类问题称为回归问题)。也就是说,监督学习既可以用于分类问题又可以用于回归问题。对于分类问题,输出空间 Y 是一个类别的集合 $C=\{c_1, c_2, \dots, c_l\}$, 其中 c_i ($i=1, 2, \dots, l$) 为样本可能所属的类别, l 表示类别数,最后分类学习获得的模型称为分类器(Classifier);对于回归问题来说,输出空间 Y 一般为实数值。分类和回归问题的求解方法基本相同,在本书后续内容中,对各种方法的讨论主要针对分类问题展开。

监督学习有两种模型。一般常用的模型是监督学习产生的全局模型,即将输入映射到期望输出。而另一种模型则是将这种映射作为一个局部模型(如案例推理及最近邻算法)。为解决一个给定的监督学习问题,可分为以下5个步骤进行:

- 1) 确定训练样本数据。在进行其他步骤前,首先应根据学习任务确定要

使用哪种样本数据。例如,为解决遥感图像分类问题,可能选择多光谱或高光谱遥感图像数据。

2) 收集训练样本数据。这些样本数据通常须具有表征真实事物的特征。一般可以由专家确定或者由机器和传感器测量得到输入和相应的输出数据。

3) 确定学习函数输入特征的表示方法。输入数据的表示形式在很大程度上决定着学习函数的准确度。传统上,输入数据会被转换成一个特征向量,该特征向量包含了许多描述数据的特征。因为维数灾难的关系,特征的个数不宜太多,但也要足够大,才能准确地预测输出。

4) 确定要学习的函数及其对应的学习算法所使用的学习器类型,如可以选择神经网络。

5) 完成设计。在步骤 2) 搜集到的样本数据上运行学习算法。可以通过将算法运行在样本数据的子集(称为验证集)上或采用交叉验证(Cross-Validation)来调整学习算法的参数。参数调整后,算法可以运行在不同于训练集的测试集上。

目前可用于监督学习的通用算法不胜枚举,其中使用最广泛的方法有神经网络、最近邻算法、遗传算法、贝叶斯学习方法及决策树方法等。著名的 Statlog 研究项目^[6]将 18 种算法用于 8 种大型实际问题,其中有卫星图像、手写数字图像、Karhunen-Loeve (KL) 数字图像、车辆图像、心脏损伤、心脏病、信誉风险和移动控制问题等应用。结果表明,各种监督学习方法都各有优缺点,算法性能在很大程度上与待处理的样本数据的特性有关,没有哪一种算法可以在所有给定的问题上都表现最好。目前,常常需要借助一些经验方法来对监督学习方法的性能进行比较,并且寻找决定监督学习方法性能的样本数据特性。

1.1.2 监督学习方法简介

近年来,随着科技的发展,互联网、计算机、机器学习等多个领域相互融合,对监督学习产生了更多的需求,目前监督学习的发展主要着眼于如何提升分类精度与计算效率,因此而提出的集成分类器和大规模数据分类器成为当前监督学习研究领域的两个热点^[7,8]。为此,本小节将对 K -最近邻算法、遗传算法、贝叶斯算法、决策树算法这 4 种主流监督学习算法以及集成分类器和大规模数据分类器进行简单的介绍。

(1) K -最近邻算法

K -最近邻算法^[9](K -Nearest Neighbors, KNN)是将在特征空间中最接近的训练样本进行分类的监督学习方法。 K -最近邻算法最初由 Cover 和 Hart 于 1967 年提出,其思路非常简单直观,易于快速实现,错误率较低。



K -最近邻算法的基本思想为：根据距离函数计算待分类样本 x 和每个训练样本的距离，选择与待分类样本 x 距离最小的 K 个样本作为 x 的 K 个最近邻，最后根据 x 的 K 个最近邻判断 x 的类别。该算法没有单独的学习阶段，是一种在分类过程中实现学习的监督学习方法。

K -最近邻算法为非参数学习算法，在处理未知分布和非正态分布的数据时分类准确率较高，主要优点包括以下 4 点：

1) 算法简单直观，易于快速实现。

2) 无需额外数据对规则进行描述，训练数据本身就是算法的规则，而且对数据的一致性问题没有要求，具有一定的抗噪声能力。

3) K -最近邻算法虽然是以极限定理为理论依据，但在类别判别过程中，决策结果只取决于极少数的相邻样本数据，故 K -最近邻算法可以较好地处理样本数量的不平衡问题。

4) 在分类过程中， K -最近邻算法直接利用样本数据间的关系，降低了类别特征选择不正确对分类结果造成的不利影响，可以最大限度地减少分类过程中的误差项。

这就使得 K -最近邻算法在处理一些类别特征不明显样本时，充分体现出其分类规则独立性的优势，进而使分类的自学习成为可能。

K -最近邻算法是基于实例的惰性学习方法，在分类时，这种学习法的运行需要非常大的存储空间。 K -最近邻算法有以下几点不足：

1) 分类速度慢。 K -最近邻算法需要存储全部的训练样本，待进行分类时，算法须计算待分样本与训练样本库中每一个样本的相似度，才能求得与其最近的 K 个样本。对于样本集规模较大或者高维样本来说，该方法时间和空间复杂度较高，时间复杂度为 $O(mn)$ ，其中 m 为向量空间中特征维数， n 为训练样本数。

2) 属性等同权重影响了准确率。 K -最近邻算法中每个属性具有相同的权重，但实际在这些特征中，有些特征与分类相关性较强，有些特征与分类相关性较弱，还有一些特征与分类不相关。因此，如果在计算相似度的时候，按所有特征等同权重来计算样本相似度会影响分类准确度。

3) 样本集规模依赖性较强。 K -最近邻算法在实际应用中受到很大的限制，因为某些类别可能无法提供充足的训练样本，无法满足 K -最近邻算法所需要的相对均匀的特征空间条件，使类别判定误差较大。

4) K 值难以确定， K 值选择不当则分类精度不能保证。

为克服上述不足，很多学者从降低计算复杂度提高算法执行效率、优化相似度度量方法和决策策略、 K 值选取等方面对 K -最近邻算法进行改进，有效提高了分类器的性能。



(2) 遗传算法

遗传算法 (Genetic Algorithm, GA)^[10]起源于 20 世纪 60 年代美国密歇根大学 Holland 教授对自然和人工自适应系统的研究, Bagley 发明“遗传算法”一词并发表了第一篇有关遗传算法应用的论文。遗传算法的基本思想为:模拟达尔文生物进化论的自然选择和 Mendel 遗传学机理的生物进化过程,将解空间中每一个点都编码为二进制位串,称为染色体,并对应一个适应度值,适应度值按概率决定个体性质遗传到下一代中的机会,在每一代中使用选择、交叉和变异等作用机制获得新的种群,若干代后,种群中包含的个体具有更高的适应度,直到满足某种收敛指标为止。遗传算法作为一种快捷、简便的算法具有搜索面广、适应性强、容错能力强等特点。它采用随机搜索方法,几乎不需要所求问题的任何信息而仅需要目标函数的信息,并可不受搜索空间是否连续或可微的限制就可找到最优解,具有较强的适应能力,且便于并行计算。

遗传算法理论方面主要包含以下内容^[11]:分析遗传算法的编码策略、全局收敛性和搜索效率的数学基础、遗传算法的新结构探索、基因操作策略及其性能分析、遗传算法参数的选择以及与其他算法的综合比较等。

以遗传算法理论为基础的遗传学习系统是一种增强式学习系统,该学习系统的理论框架由 Holland^[10]奠定,一般由产生式规则组成候选种群,和外界环境进行交互式学习以完成特定的任务。Holland^[10]发展了认知系统一号 (Cognitive Systems - I, CS - I), 该系统是第一个遗传学习系统,又称为分类器系统 (Classifier System), 并为后续研究提供了模板。

目前,遗传学习已被应用于诸多领域,其本身的发展也处于不断进化过程中^[12]。

(3) 贝叶斯算法

自从 20 世纪 60 年代,贝叶斯学派形成后,关于贝叶斯算法的研究经久不衰。自 20 世纪 90 年代以来,贝叶斯算法一直是机器学习研究的重要方向之一。贝叶斯算法提供了一种概率手段,可用于确定给定数据下最可能的假设。贝叶斯算法的基本思想为:假设待考察的样本遵循某种概率分布,基于这些先验和数据观测假定进行推理,获得观测数据的后验概率,以此作出最优决策。贝叶斯算法能够方便地处理不完全数据,能够学习变量间的因果关系,同时贝叶斯网络与贝叶斯统计相结合,能够充分利用领域知识和样本数据的信息。

近年来,贝叶斯方法以其独特的不确定性知识表达形式、丰富的概率表达能力、综合先验知识的增量学习特性等成为当前众多机器学习方法中最为引人注目的焦点之一。其中贝叶斯密度估计探讨如何根据样本的数据信息和专家的先验知识获得对未知变量(向量)的分布及其参数的估计。贝叶斯网络是处理不确定信息最有效的方法之一,是表示变量间概率分布及关系的有向无环图,

应用于有条件地依赖多种控制因素的决策，可以从不完全、不精确或不确定的知识或信息中作出推理，在多个领域中获得广泛应用。朴素贝叶斯 (Naive Bayesian) 模型假定特征向量的各分量相对于决策变量是独立的，也就是说各分量独立地作用于决策变量。尽管这一假定在一定程度上限制了朴素贝叶斯模型的适用范围，然而在实际应用中，不仅指数级的降低了贝叶斯网络构建的复杂性，而且即使在违背这种假定的条件下，朴素贝叶斯也表现出相当的健壮性和高效性，它已经成功地应用到分类、聚类及模型选择等数据挖掘的任务中。目前，许多研究人员正致力于放松特征变量间条件独立性的限制，以使它适用于更大的范围。

(4) 决策树算法

决策树^[13]算法起源于 20 世纪 60 年代的概念学习系统 (Concept Learning System)，是目前应用最广泛的监督学习算法之一。决策树算法的基本思想为：采用树结构表示数据特征与目标之间的映射关系。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直至到达叶子节点，将叶子节点存放的类别作为决策结果。

决策树算法能够逼近离散函数，对噪声数据有很好的鲁棒性，该方法对完整表示的假设空间进行搜索，从而避免了受限假设空间的不足，能够学习解析表达式。它采用自顶向下不回溯策略，能保证找到一个简单的树。决策树算法通常包括两个部分，即树的生成和树的修剪。树的生成从一棵空的决策树开始，不断加入新的节点，根节点包含所有数据，根据设定的标准选择测试属性，用不同的测试属性递归进行数据分割，直到生成的决策树能够对所有训练样本正确分类为止。树的修剪过程是去掉一些可能是噪声或异常的数据。决策树建立的关键在于如何确定内部节点的分支。而在确定内部节点的分支过程中，选择不同的字段值将划分出不同的样本子集。

对于基本决策树算法，目前已有许多扩展算法，其中主要包括以下几方面：后修剪的方法；处理实数值的属性；容纳缺少属性值的训练样本；当有了新的训练实例时递增精化决策树；使用信息增益之外的其他属性选择度量；考虑与实例属性关联的代价。1979 年，由 J. Ross Quinlan 提出了 ID3 算法，此算法能够减少树的深度，但是缺少对叶子数的探讨。在 ID3 算法的基础上发展了 C4.5 决策树算法，C4.5 决策树能够处理连续属性，此外对于预测变量的缺值处理、剪枝技术、派生规则等方面都作了较大改进。

(5) 集成学习分类器

集成学习是使用一系列学习器进行学习，并使用某种规则把各个学习结果



进行整合,从而获得比单个学习器更好的学习效果的一种监督学习方法^[14]。分类问题是集成学习研究的基本问题。集成学习分类器的基本思想是:将多个弱分类器按特定方式集成起来,得到一个更强大的分类器。这种集成是有条件的,要想实现对整体精度的提升,要求每个子分类器的错误率都小于50%,且各分类器产生的错误相互独立。实现集成学习分类器的两个关键步骤是:构造足够多有差异的子分类器;对这些子分类器进行合理的集成。

构造子分类器的方法主要包括以下几种:

1) 样本子集法。对训练样本集进行处理,即采用同一学习算法对不同样本子集进行独立学习,以生成多个子分类器。常使用这种方法构建基于决策树、神经网络、规则学习等算法的子分类器,这些算法中训练集微小改动便会造成分类器参数的较大变化,样本子集法对这类稳定性较差的算法,具有较好效果。Bagging^[15]、Boosting^[16]以及交叉检验^[17]均属于样本子集法。

2) 属性选择法。当样本属性较多时对训练数据的输入属性进行一定操作可以让每个子分类器分别学习不同的属性,之后再对学习结果进行整合。该方法在构造子分类器时同样存在一定的局限性,当样本输入属性的数目比较少的时候,通过属性选择法构造子分类器并综合其结果,最终的分类精度不会提高反而有所降低。

3) 类别编码法。此类方法以错误校正输出编码法^[18]为代表,首先对类别编码,将每个类别都对应到 L 位的二进制位串,从而得到一个 $K \times L$ 的二维编码阵(K 表示类别数);接着对编码矩阵按列进行学习,得到对应各列的二分类器。对于编码的第 i ($1 \leq i \leq L$)位为0的类别,二分类器对应该类别的输出为0;反之则输出为1。预测过程中,每个二分类器均对新样本进行分类,从而得到码长 L 的位串。利用最小汉明距离(Minimum Hamming Distance)法确定与该位串最接近的类别编码,从而实现对新样本类别的判断。类别编码法能将多类问题简化成为二类判别问题,因而引起广泛的关注。

4) 随机因素法。引入随机因素实现多分类器。以BP神经网络分类器组为例,对相同样本集、相同结构的网络随机采用不同的权值初值,从而生成不同的子分类器。在决策树方法中,叶节点的扩展属性也可以作为一种随机因素,即对相同的内部节点随机选取不同扩展属性,从而得到有差别的决策树。

当得到子分类器后,可采用简单投票、贝叶斯投票、Winner-Take-All以及证据推理等方法进行集成。

集成学习的主要目标是提高分类精度,但也会造成存储量和计算负担增加的问题。集成学习需要更大的空间来保存各个子分类器,相应也需要更多计算时间来完成各个子分类器的学习过程。集成学习存在的问题是目前机器学习中的一个热门的研究方向,如如何提高集成学习的学习效果、如何提高集成学习