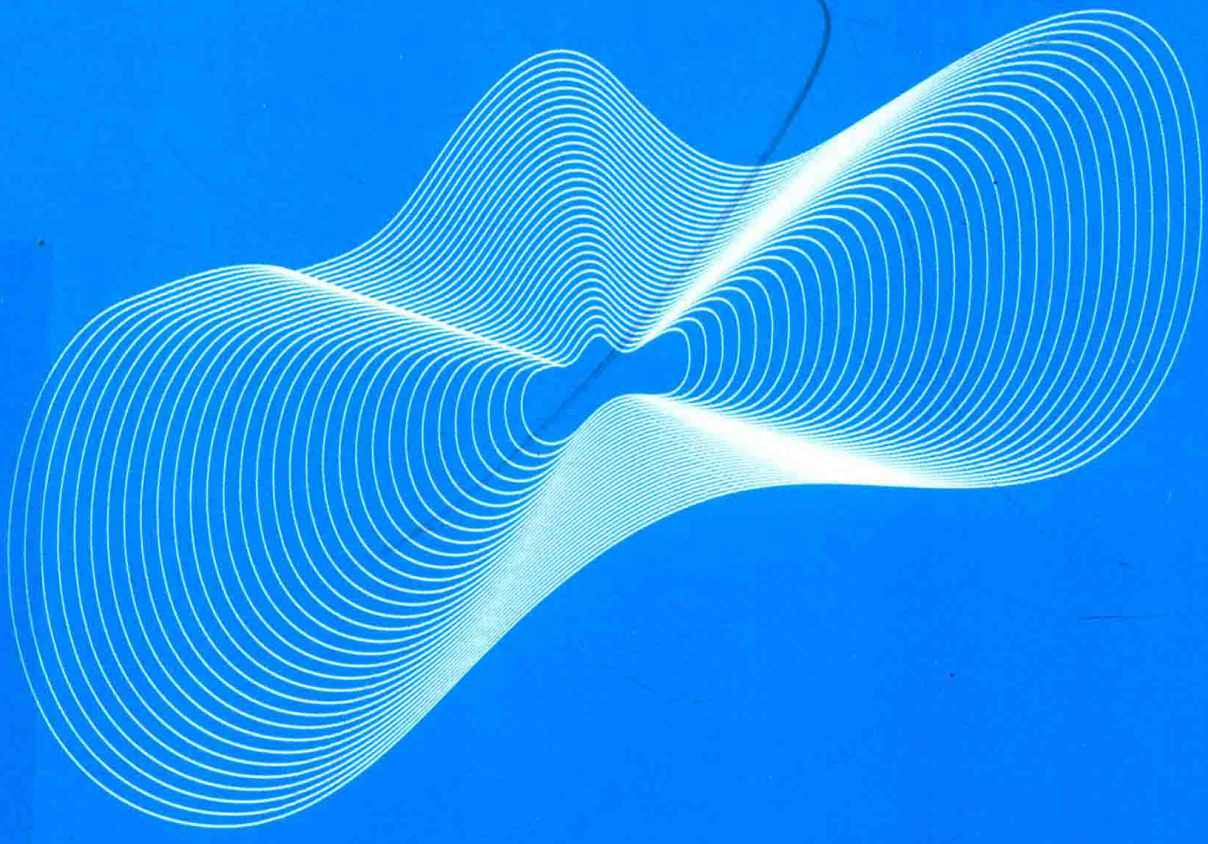
高等院校信息类新专业规划教材
大数据和人工智能技术丛书

EVOLUTIONARY
MACHINE LEARNING

演化机器学习

徐　华◎编著



北京邮电大学出版社
www.buptpress.com

高等院校信息类新专业规划教材
大数据和人工智能技术丛书

演化机器学习

徐 华 编著

北京邮电大学出版社
www.buptpress.com

内 容 简 介

近年来，演化计算作为计算智能中传统的优化技术，已经广泛应用于求解各种数据挖掘问题，形成了一种基于遗传的机器学习新范式学习分类器。一方面，在真实场景中采集的原始数据不可避免地包含着冗余乃至噪声属性的信息，这些不相关的特征将对学习分类器算法的学习性能与计算效率造成负面影响。另一方面，学习分类器以显式规则表示目标概念，在监督学习或强化学习机制的基础上，利用演化算法对规则空间进行搜索，从而完成学习任务。规则空间的有效搜索是影响学习分类器性能的关键。针对上述问题，本书的主要探讨内容：一是学习分类器与特征选择方法，重点是做两者的整合研究，将学习分类器的分类模型构建过程与特征选择的特征子集搜索过程统一集成在基于遗传的机器学习框架下，同时改善分类算法的预测性能与运行效率；二是从提高规则空间的搜索质量出发，着眼于分类问题，介绍了基于分布估计算法的学习分类器。

本书可作为大数据及人工智能等相关专业的教材与参考用书。

图书在版编目(CIP)数据

演化机器学习 / 徐华编著. -- 北京：北京邮电大学出版社，2021.4

ISBN 978-7-5635-6330-2

Ⅰ. ①演… Ⅱ. ①徐… Ⅲ. ①机器学习 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2021)第 027485 号

策划编辑：姚 顺 刘纳新 **责任编辑**：刘 颖 **封面设计**：七星博纳

出版发行：北京邮电大学出版社

社 址：北京市海淀区西土城路 10 号

邮政编码：100876

发 行 部：电话：010-62282185 传真：010-62283578

E-mail：publish@bupt.edu.cn

经 销：各地新华书店

印 刷：北京玺诚印务有限公司

开 本：787 mm×1 092 mm 1/16

印 张：15

字 数：314 千字

版 次：2021 年 4 月第 1 版

印 次：2021 年 4 月第 1 次印刷

ISBN 978-7-5635-6330-2 定价：45.00 元

大数据和人工智能技术丛书

顾问委员会

编委会

前　言

伴随着信息技术和移动互联网技术的发展以及社会经济数字化进程的加快，以计算机、智能硬件为代表的数字化移动设备已经渗透到社会领域的方方面面，正在深刻地改变着传统的工作和生活方式。一方面，网络移动互联技术的普及为信息化革命注入了强大活力，覆盖全球的移动互联网为人们提供了一个方便快捷又廉价的信息发布平台，从而带来了数据规模的爆炸性增长。另一方面，云存储技术与数据库技术的发展也为收集并存储这些海量数据提供了便利。然而，虽有海量的数据，若缺乏强有力的数据分析工具则只能带来信息的缺位。时至今日，面对数据的汪洋大海，人们已经将自己置身于“数据泛滥但信息贫乏”的窘境当中。20 世纪 80 年代末，一门致力于在实现数据分析自动化的交叉学科应运而生，这就是至今方兴未艾的数据挖掘（Data Mining，DM）。简言之，数据挖掘指的是从大规模数据中提取有用的信息（知识），因此也被称为从数据中发现知识（Knowledge Discovery from Data，KDD）。作为一门交叉学科，数据挖掘涉及多个领域，既有负责海量数据存储的数据库技术，也有提供强大的运算能力的高性能云计算技术，不过其中最为核心的是进行知识抽取的数据分析技术。从早期的统计学到模式识别（Pattern Recognition，PR），直至当下更为盛行的机器学习（Machine Leaning，ML）、深度学习（Deep Learning），数据挖掘从多个领域汲取养分，增强了本学科海量数据分析的实用性和有效性。

现实世界中各个行业、各个部门对于数据挖掘提出的应用需求五花八门，总的说

来，数据挖掘任务可以分为两大类：描述性（Descriptive）挖掘与预测性（Predictive）挖掘。前者主要挖掘数据集的一般性质，对应于机器学习领域的无监督学习（Unsupervised Learning）；后者则根据当前数据进行推断并做出预测，其中主要用到的是有监督学习（Supervised Learning）技术。作为一种典型的预测性任务，分类（Classification）旨在找出描述和区分数据类别（或称概念）的分类模型，以便使用该模型预测类别标签值未知的数据实例，模型构建过程中主要的可利用信息是已知类别标签值的训练数据集（Training Dataset）。现实中人们感兴趣的预测值有很大一部分都可以抽象为离散型的类别信息（如文字识别、故障诊断、入侵检测等），故而分类成为一种主要的分析手段。为此，在过去的几十年中，研究人员也提出了许多应用广泛的分类算法，包括决策树（Decision Tree）、朴素贝叶斯分类器（Naive Bayesian Classifier）、k-近邻算法（k-Nearest Neighbor，k-NN）、人工神经网络（Artificial Neural Network，ANN）、支持向量机（Support Vector Machine，SVM）和深度神经网络（Deep Neural Network，DNN）。

近年来，一种新的名为学习分类器（Learning Classifier System，LCS）的机器学习范式正吸引越来越多研究者的注意力。总的来说，学习分类器基于规则归纳的思想，主要致力于解决分类问题。在学习分类器中，处于中心地位的规则学习单元通常是演化计算（Evolutionary Computaion，EC）中的遗传算法（Genetic Algorithm，GA），故而在一部分文献中又被称为基于遗传的机器学习（Genetic-Based Machine Learning，GBML）。通过将学习性能指标定义为优化目标函数，学习分类器实质上将作为学习问题的分类任务转化为传统的优化问题进行求解，继而基于遗传算法的全局优化搜索能力确保规则知识表示的假设空间中一定强度的随机化搜索，以期在合理的运算时间内收敛到较优性能的问题解，从而更好地平衡了算法性能与计算效率这对矛盾。

在实际的分类问题中，一方面，在真实场景中采集的原始数据不可避免地包含着冗余乃至噪声的属性信息，这些不相关的特征将对学习分类器算法的学习性能与计算效率造成负面影响。另一方面，学习分类器以显式规则表示目标概念，在监督学习或强化学习机制的基础上，利用演化算法对规则空间进行搜索，从而完成学习任务。规

则空间的有效搜索是影响学习分类器性能的关键。针对上述问题，本书从内容体例上分为上、下两篇，各自自成体系，分别以相对独立的方式对学习分类器中的重要内容加以介绍，以便于读者根据需要分别研读。上篇重点介绍了学习分类器与特征选择方法，重点是介绍两者的整合研究内容，将学习分类器的分类模型构建过程与特征选择的特征子集搜索过程统一集成在基于遗传的机器学习框架下，同时改善分类算法的预测性能与运行效率。下篇从提高规则空间的搜索质量出发，着眼于分类问题，介绍了基于分布估计算法的学习分类器。本书的相关内容主要来自作者所在清华大学团队近年来在演化机器学习研究方面的工作积累，力图以比较全面的视角对学习分类器及相关重要问题的研究做一个深入的介绍和探讨。在本书完成之际感谢清华大学计算机系人工智能研究团队成员温赟、杨甲东、袁源、王勃、黄嘉宇、余文梦、孟繁阳等同学为此付出的辛勤努力。

本书的相关工作受国家自然科学基金项目（No. 61673235，61175110，60875073，60575057）的持续资助支持，本书的编写工作得到清华大学教育教学改革项目的支持。本书可作为人工智能领域、机器学习领域的学生、研究人员和工程技术人员的教材和参考用书。由于本书相关的内容也处于研究与探讨阶段，不当之处请读者给予指正（请用电子邮件联系：xuhua@tsinghua. edu. cn）。同时我们持续性的相关研究工作与成果会公开在共享网站：https：//github. com/thuiar。

徐　华
于清华大学

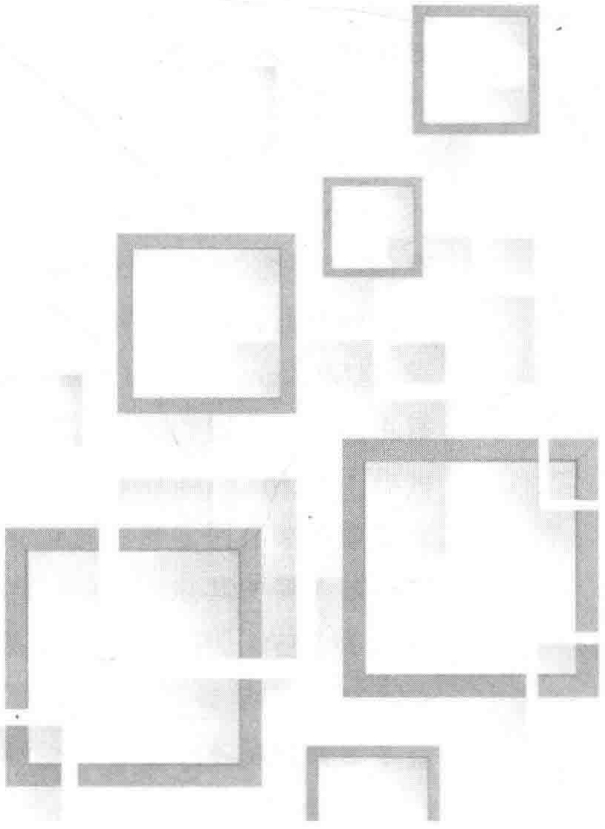
目 录

上篇 演化机器学习——内嵌特征选择的学习分类器

上篇

演化机器学习

——内嵌特征选择的学习分类器



第1章

上篇引言

1.1 研究背景

伴随着信息技术和移动互联网技术的发展以及社会经济数字化进程的加快，以计算机、智能移动硬件为代表的数字化设备已经渗透到日常生活的方方面面，正在深刻地改变着传统的工作和生活方式。一方面，网络移动互联技术的普及为信息化革命注入强大活力，覆盖全球的 Internet 为人们提供了一个方便快捷又廉价的信息发布平台，带来了数据量的爆炸性增长。另一方面，云存储技术与大规模分布式数据库技术的发展也为收集并存储这些海量数据提供了便利。然而，虽有海量的数据，若缺乏强有力的数据分析工具却只能带来信息的缺位或者数据孤岛，时至今日，面对数据的汪洋大海，人们已经将自己置身于“数据泛滥但信息贫乏”的窘境当中[1]。20 世纪 80 年代末，一门致力于在实现数据分析自动化的交叉学科应运而生，这就是至今方兴未艾的数据挖掘(Data Mining，DM)。简言之，数据挖掘指的是从大规模数据中提取有用的信息(知识)，因此也被称为从数据中发现知识(Knowledge Discovery from Data，KDD)[1]。作为一门交叉学科，数据挖掘涉及多个领域，既有负责海量数据存储的数据库技术，也有提供强大的运算能力的高性能计算技术，不过其中最为核心的是进行知识抽取的数据分析技术。从早期的统计学[2]到模式识别(Pattern Recognition，PR)[3]，直至当下更为盛行的机器学习(Machine Leaning，ML)[4]，数据挖掘从多个领域汲取养分，增强了本学科海量数据分析的实用性和有效性。