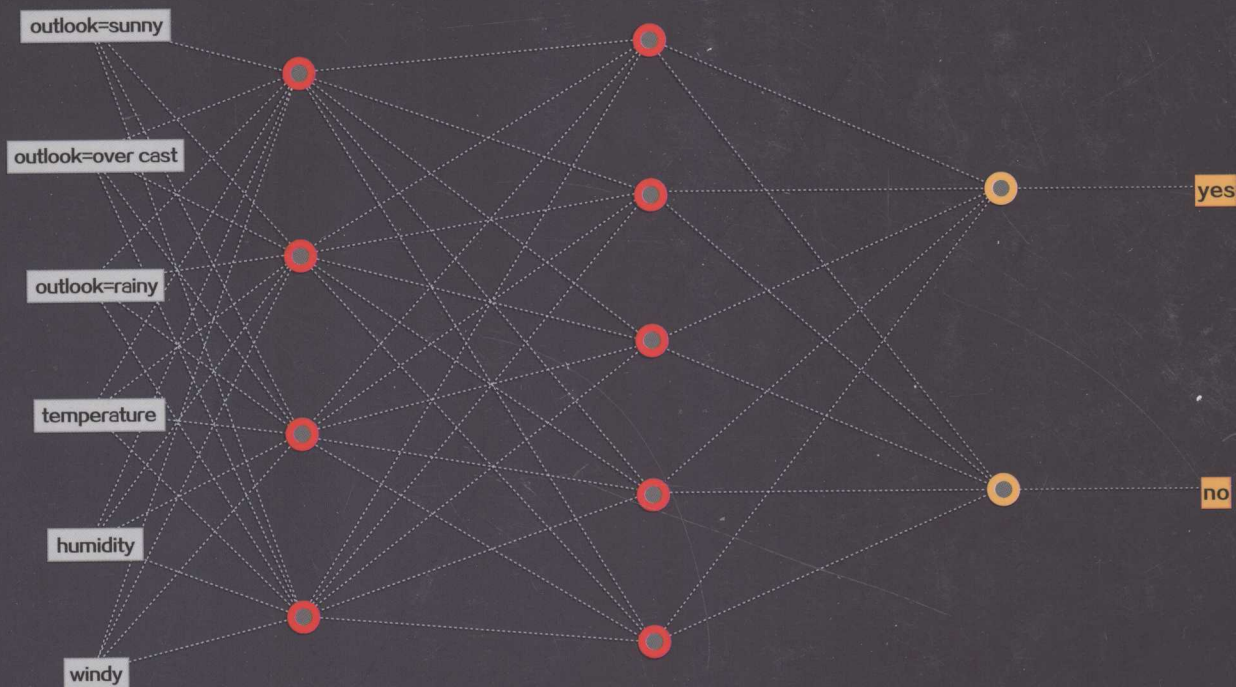


第二版  
版本更新，增实战案例

系统讲解数据挖掘机器学习工具Weka  
经典的开源挖掘工具、开放的Java环境  
初学者的入门首选、研究者的钻研利器



# 数据挖掘与机器学习

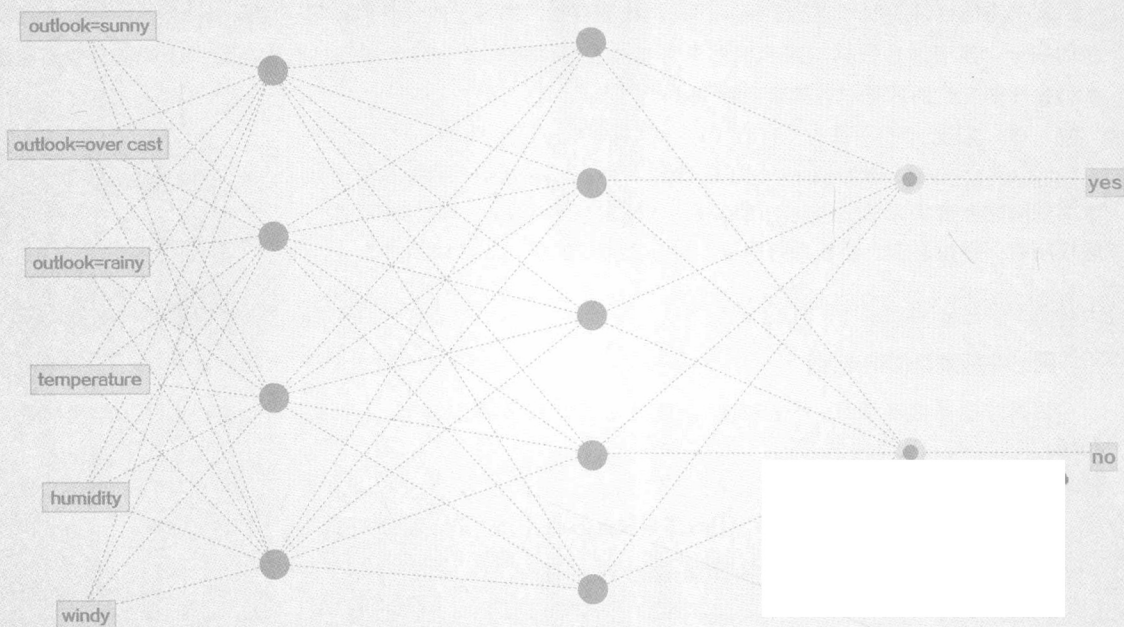
# WEKA

## 应用技术与实践 (第二版)

袁梅宇 编著



清华大学出版社



# 数据挖掘与机器学习 WEKA

应用技术与实践 (第二版)

袁梅宇 编著

清华大学出版社  
北京

## 内 容 简 介

本书借助代表当今数据挖掘和机器学习最高水平的著名开源软件 Weka, 通过大量的实践操作, 使读者了解并掌握数据挖掘和机器学习的相关技能, 拉近理论与实践的距离。全书共分 9 章, 主要内容包括 Weka 介绍、探索者界面、知识流界面、实验者界面、命令行界面、Weka 高级应用、Weka API、学习方案源代码分析和机器学习实战。

本书系统讲解 Weka 3.7.13 的操作、理论和应用, 内容全面、实例丰富、可操作性强, 做到理论与实践的统一。本书适合数据挖掘和机器学习相关人员作为技术参考书使用, 也适合用作计算机专业高年级本科生和研究生的教材或教学参考用书。

本书封面贴有清华大学出版社防伪标签, 无标签者不得销售。

版权所有, 侵权必究。侵权举报电话: 010-62782989 13701121933

### 图书在版编目(CIP)数据

数据挖掘与机器学习——WEKA 应用技术与实践/袁梅宇编著. —2 版. —北京: 清华大学出版社, 2016

ISBN 978-7-302-44470-1

I. ①数… II. ①袁… III. ①数据采集—软件工具 IV. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 171536 号

责任编辑: 魏 莹 郑期彤

封面设计: 杨玉兰

责任校对: 李玉萍

责任印制: 杨 艳

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座 邮 编: 100084

社 总 机: 010-62770175

邮 购: 010-62786544

投稿与读者服务: 010-62776969, [c-service@tup.tsinghua.edu.cn](mailto:c-service@tup.tsinghua.edu.cn)

质量反馈: 010-62772015, [zhiliang@tup.tsinghua.edu.cn](mailto:zhiliang@tup.tsinghua.edu.cn)

印 装 者: 清华大学印刷厂

经 销: 全国新华书店

开 本: 185mm×260mm

印 张: 34.25

字 数: 832 千字

版 次: 2014 年 8 月第 1 版 2016 年 8 月第 2 版

印 次: 2016 年 8 月第 1 次印刷

印 数: 1~3000

定 价: 79.00 元

---

产品编号: 065181-01

# 再版前言

自本书第一版出版到现在已经过去近两年。这段时间内，数据挖掘和机器学习领域快速发展，投入到相关领域研究的人员也越来越多，Weka 爱好者队伍也随之逐年发展壮大，Weka 学习讨论群所讨论内容的技术含量也日渐丰富。

第二版的修订工作以 Weka 3.7.13 版本为准，为此，全书重新截图，按照 Weka 新版本重新修订正文内容。此次再版修改了第一版中一些表述不清楚的陈述、前后不一致的术语，还新增了以下内容：第 1 章 1.3 节新增无法连接包管理器的解决办法，第 2 章 2.7 节新增边界可视化工具和代价/收益分析可视化及相关实验内容，第 4 章 4.2 节新增拆分评估器可视化参数内容，新增完整的第 9 章机器学习实战，丰富了 Weka 实践内容。

修订后的第二版共分 9 章。第 1 章介绍 Weka 的历史和功能、数据挖掘和机器学习的基本概念、Weka 系统安装，以及示例数据集；第 2 章介绍探索者(Explorer)界面的使用，主要内容包括图形用户界面、预处理、分类、聚类、关联、选择属性，以及可视化；第 3 章介绍知识流(KnowledgeFlow)界面，主要内容有知识流介绍、知识流组件、使用知识流组件，以及实践教程；第 4 章介绍实验者(Experimenter)界面，主要内容有实验者界面介绍、标准实验、远程实验，以及实验结果分析；第 5 章介绍命令行界面，主要内容有命令行界面介绍、Weka 结构、命令行选项、过滤器和分类器选项，以及 Weka 包管理器；第 6 章介绍一些 Weka 的高级应用，主要介绍 Weka 的贝叶斯网络、神经网络、文本分类和时间序列分析及预测；第 7 章介绍 Weka API，说明使用 Java 源代码来实现常见数据挖掘任务的基础知识，并给出一个展示如何进行数据挖掘的综合示例；第 8 章通过对 NaiveBayes 学习方案的源代码进行分析，深入研究 Weka 学习方案的工作原理，为开发人员提供实现学习算法的编码基础；第 9 章介绍如何使用 Weka 工具挖掘实际的大型数据集，以精选的两个 KDD 竞赛数据集为例，使读者能够快速进入实际的案例场景，应用所学数据挖掘知识来面对大数据的挖掘问题，考验自己完成难度较大的挖掘项目的动手能力。

第二版改动的内容较多，总体工作量很大，花费了很多时间。从酝酿第二版内容开始，至其杀青，历时超过一年。作者的感觉是：比编写第一版还要辛苦些。且不说 Weka 版本变动导致的修改，重新截图、重新梳理文字、重新改写 API 文档等，费时费力。因时间变化引起的一个小小的技术变动，就让人费力应对。例如，怀卡托大学后来不再提供包管理器元数据，导致第一版所述的解决办法不再有效，只能重新寻找解决包管理器无法连接的替代方法。又如，新版本 Weka 的 NaiveBayes 源代码有一些变动，作者不得不修订第 8 章的内容以适应新的版本变化。再如，第一版提供的网络链接有的已经不再有效，出版社编辑老师测试了所有的链接，保证了第二版提供的网络链接的正确性。当然，由于世界变化太快，无法保证在一两年后这些链接不会失效，这是无可奈何的事，作者只能保证书中叙述的方案在交稿时可行。

最耗费心力的应该是第 9 章的编写。早在第一版的写作中，曾经就有编写一个章节专门讲述 Weka 综合应用案例的设想，但苦于手上没有合适的实验对象。理想的应用案例必须满足如下要求：第一，难度适中。不能太简单，过于简单的小儿科案例会违背编写综合



应用的初衷；也不能太难，如果应用的技术方案太偏或难以理解，就达不到锻炼读者实际动手能力的意义。第二，领域不能太窄，应该让绝大多数人都能理解。第三，运算量不能太大，应该满足普通计算机能够处理的要求。这就限制了目标数据集文件大小为数十兆字节至数吉字节范围以内，实例总数在数十万条至数千万条之间，一台计算机能够在两周左右运行完毕。作者花费了很长时间寻找满足以上要求的案例，最后选中 KDD Cup 1999 和 KDD Cup 2010 竞赛数据集，前者共有 42 个属性，10%数据子集文件的大小为 45MB，样本数为 494021，完整的数据集文件大小为 743MB，样本数为 4898431；后者有两个数据集，本书选中的是较大的数据集，共有 21 个属性，训练数据集文件大小为 5.29GB，样本数为 20012498。认真的读者会发现，完成这两个案例的实验将会很辛苦，花费的精力和时间会远超预期。作者想象出这么一个画面：读者按照书中的实验方法工作至深夜，硬盘灯不停闪烁，CPU 利用率一直高居 95%，读者担心心爱的计算机会突然崩溃但仍然坚持，直至最终胜利。作者预先恭喜那些能够独立完成实验的读者，因为你们有足够的能力和毅力应付技术挑战，胜任要求极高的挖掘工作。

尽管在写作中付出了很多艰辛的劳动，但限于作者有限的能力和精力，书中肯定还存在一些缺陷，甚至错误，敬请各位读者批评指正。作者感谢修订工作的贡献者，昆明理工大学计算机系吴霖老师审阅了本书第 9 章内容，提出了很多建设性建议，感谢吴霖老师的贡献。昆明理工大学 2014 级研究生卫明同学参与了第 1 章和第 2 章的修订工作；光荣与梦想、弦月、Brady、海、\_\_末瞳.夫、不说再见！等网友对第一版提出了宝贵的建议，作者在第二版中采纳了这些建议，感谢这些朋友的贡献。第 9 章参考了昆明理工大学 2014 届计算机系吴泽恒同学本科毕业设计论文的部分内容，他是我指导过的最优秀的学生，感谢吴泽恒同学。感谢选择本书为高校教学参考书的教师在使用过程中提出的反馈意见和建议，作者学习到一些很有价值的思考方式。再次感谢清华大学出版社的编辑老师在出版方面提出的建设性意见和给予的无私帮助，编辑老师一丝不苟的工作态度给我留下深刻的印象。感谢购买本书的朋友，欢迎批评指正，你们的批评建议都会受到重视，并在再版中改进。希望第二版的发行能够吸引更多的读者和反馈建议。

编者

# 第一版前言

当代中国掀起了一股学习数据挖掘和机器学习的热潮，从斯坦福大学公开课“机器学习课程”，到龙星计划的“机器学习 Machine Learning”课程，再到加州理工学院公开课“机器学习与数据挖掘”课程，参加这些网络课程学习的人群日益壮大，数据挖掘和机器学习炙手可热。

数据挖掘是数据库知识发现中的一个步骤，它从大量数据中自动提取出隐含的、过去未知的、有价值的潜在信息。机器学习主要设计和分析一些让计算机可以自动“学习”的算法，这类算法可以从数据中自动分析获得规律，并利用规律对未知数据进行预测。数据挖掘和机器学习这两个领域联系密切，数据挖掘利用机器学习提供的技术来分析海量数据，以发掘数据中隐含的有用信息。

数据挖掘和机器学习这两个密切相关的领域存在一个特点：理论很强而实践很弱。众所周知，理论和实践是研究者的左腿和右腿，缺了一条腿的研究者肯定难以前行。有的技术人员花了若干年时间进行研究，虽然了解甚至熟悉了很多公式和算法，但仍然难以真正去面对一个实际挖掘问题并很好地解决手上的技术难题，其根本原因就是缺乏实践。

本书就是为了试图解决数据挖掘和机器学习的实践问题而编写的。本书依托新西兰怀卡托大学采用 Java 语言开发的著名开源软件 Weka，该系统自 1993 年开始由新西兰政府资助，至今已经历了 20 多年的发展，功能已经十分强大和成熟。Weka 集合了大量的机器学习和相关技术，受领域发展和用户需求所推动，代表了当今数据挖掘和机器学习领域的最高水平。因此，研究 Weka 能帮助研究者从实践去验证所学的理论，显然有很好的理论意义及实践意义。

本书共分 8 章。第 1 章介绍 Weka 的历史和功能、数据挖掘和机器学习的基本概念、Weka 系统安装，以及示例数据集；第 2 章介绍 Explorer 界面的使用，主要内容包括图形用户界面、预处理、分类、聚类、关联、选择属性，以及可视化；第 3 章介绍 KnowledgeFlow 界面，主要内容有知识流介绍、知识流组件、使用知识流组件，以及实践教程；第 4 章介绍 Experimenter 界面，主要内容有 Experimenter 界面介绍、标准实验、远程实验，以及实验结果分析；第 5 章介绍命令行界面，主要内容有命令行界面介绍、Weka 结构、命令行选项、过滤器和分类器选项，以及 Weka 包管理器；第 6 章介绍一些 Weka 的高级应用，主要介绍 Weka 的贝叶斯网络、神经网络、文本分类和时间序列分析及预测；第 7 章介绍 Weka API，说明使用 Java 源代码来实现常见数据挖掘任务的基础知识，并给出一个展示如何进行数据挖掘的综合示例；第 8 章通过对一个学习方案的源代码进行分析，深入研究 Weka 学习方案的工作原理，为开发人员提供编写学习算法的技术基础。

在阅读大量相关文献的过程中，作者深深为国外前辈们的理论功底和实践技能所折服，那些巨人们站在高处，使人难以望其项背。虽然得益于诸如网易公开课和龙星计划等项目，我们有机会和全世界站在同一个数量级的知识起跑线上，但是，这并不意味着能在将来的竞争中占据优势，正如孙中山先生所说“革命尚未成功，同志仍须努力”，让我们

一起共勉。

在本书的编写过程中，作者力求精益求精，但限于作者的知识和能力，且很多材料都难以获取，考证和去伪存真是一件时间开销非常大和异常困难的工作，因此书中肯定会有遗漏及不妥之处，敬请广大读者批评指正。

作者专门为本书设置读者 QQ 群，群号 245295017，欢迎读者加群，下载和探讨书中源代码，抒写读书心得，进行技术交流等。

本书承蒙很多朋友、同事的帮助才得以成文。特别感谢 Weka 开发组的全体人员，他们将自己 20 年心血汇聚的成果开源，对本领域贡献巨大；衷心感谢清华大学出版社的编辑老师在内容组织、排版，以及出版方面提出的建设性意见和给予的无私帮助；感谢昆明理工大学提供的宽松的研究环境；感谢昆明理工大学计算机系教师缪祥华博士，他为本书的成文提出了很多建设性的建议，对本书的改进帮助甚大；感谢昆明理工大学计算机系海归博士吴霖老师，他经常和作者一起讨论机器学习的技术问题，为本书的编写贡献了很多智慧；感谢昆明理工大学现代教育中心的何佳老师，他完成了本书部分代码的编写和测试工作；感谢国内外的同行们，他们在网络论坛和博客上发表了众多卓有见识的文章，作者从中学习到很多知识，由于来源比较琐碎，无法一一列举，感谢他们对本书的贡献；感谢理解和支持我的家人，他们是我写作的坚强后盾。感谢购买本书的朋友，欢迎批评指正，你们的批评建议都会受到重视，并在再版中改进。

编者

# 目 录

第 1 章 Weka 介绍.....	1	2.1.3 图像输出.....	37
1.1 Weka 简介.....	2	2.1.4 手把手教你用.....	37
1.1.1 Weka 历史.....	3	2.2 预处理.....	40
1.1.2 Weka 功能简介.....	3	2.2.1 加载数据.....	40
1.2 基本概念.....	5	2.2.2 属性处理.....	43
1.2.1 数据挖掘和机器学习.....	5	2.2.3 过滤器.....	44
1.2.2 数据和数据集.....	5	2.2.4 过滤器算法介绍.....	46
1.2.3 ARFF 格式.....	6	2.2.5 手把手教你用.....	52
1.2.4 预处理.....	7	2.3 分类.....	59
1.2.5 分类与回归.....	10	2.3.1 分类器选择.....	59
1.2.6 聚类分析.....	12	2.3.2 分类器训练.....	61
1.2.7 关联分析.....	12	2.3.3 分类器输出.....	62
1.3 Weka 系统安装.....	13	2.3.4 分类算法介绍.....	65
1.3.1 系统要求.....	13	2.3.5 分类模型评估.....	79
1.3.2 安装过程.....	14	2.3.6 手把手教你用.....	81
1.3.3 Weka 使用初步.....	16	2.4 聚类.....	98
1.3.4 系统运行注意事项.....	18	2.4.1 Cluster 标签页的操作.....	98
1.4 访问数据库.....	24	2.4.2 聚类算法介绍.....	99
1.4.1 配置文件.....	25	2.4.3 手把手教你用.....	101
1.4.2 数据库设置.....	26	2.5 关联.....	107
1.4.3 常见问题及解决办法.....	27	2.5.1 Associate 标签页的操作.....	107
1.5 示例数据集.....	28	2.5.2 关联算法介绍.....	108
1.5.1 天气问题.....	29	2.5.3 手把手教你用.....	111
1.5.2 鸢尾花.....	30	2.6 选择属性.....	117
1.5.3 CPU.....	31	2.6.1 Select attributes 标签页的 操作.....	118
1.5.4 玻璃数据集.....	32	2.6.2 选择属性算法介绍.....	119
1.5.5 美国国会投票记录.....	33	2.6.3 手把手教你用.....	120
1.5.6 乳腺癌数据集.....	33	2.7 可视化.....	128
课后强化练习.....	34	2.7.1 Visualize 标签页.....	128
第 2 章 探索者界面.....	35	2.7.2 边界可视化工具.....	131
2.1 图形用户界面.....	36	2.7.3 代价/收益分析可视化.....	133
2.1.1 标签页简介.....	36	2.7.4 手把手教你用.....	134
2.1.2 状态栏.....	37	课后强化练习.....	140



第3章 知识流界面 .....	143	5.2.4 其他包 .....	238
3.1 知识流介绍 .....	144	5.3 命令行选项 .....	238
3.1.1 知识流特性 .....	144	5.3.1 常规选项 .....	239
3.1.2 知识流界面布局 .....	145	5.3.2 特定选项 .....	241
3.2 知识流组件 .....	148	5.4 过滤器和分类器选项 .....	242
3.2.1 数据源 .....	148	5.4.1 过滤器选项 .....	242
3.2.2 数据接收器 .....	151	5.4.2 分类器选项 .....	245
3.2.3 评估器 .....	155	5.4.3 手把手教你用 .....	247
3.2.4 可视化器 .....	156	5.5 包管理器 .....	252
3.2.5 其他工具 .....	158	5.5.1 命令行包管理器 .....	252
3.3 使用知识流组件 .....	160	5.5.2 运行安装的算法 .....	254
3.4 手把手教你用 .....	162	课后强化练习 .....	255
课后强化练习 .....	181	第6章 Weka 高级应用 .....	257
第4章 实验者界面 .....	183	6.1 贝叶斯网络 .....	258
4.1 简介 .....	184	6.1.1 简介 .....	258
4.2 标准实验 .....	185	6.1.2 贝叶斯网络编辑器 .....	261
4.2.1 简单实验 .....	185	6.1.3 在探索者界面中使用贝叶斯 网络 .....	269
4.2.2 高级实验 .....	190	6.1.4 结构学习 .....	270
4.2.3 手把手教你用 .....	198	6.1.5 分布学习 .....	272
4.3 远程实验 .....	210	6.1.6 查看贝叶斯网络 .....	273
4.3.1 远程实验设置 .....	210	6.1.7 手把手教你用 .....	276
4.3.2 手把手教你用 .....	213	6.2 神经网络 .....	286
4.4 分析结果 .....	221	6.2.1 GUI 使用 .....	286
4.4.1 获取实验结果 .....	221	6.2.2 手把手教你用 .....	289
4.4.2 动作 .....	221	6.3 文本分类 .....	293
4.4.3 配置测试 .....	222	6.3.1 文本分类示例 .....	294
4.4.4 保存结果 .....	225	6.3.2 分类真实文本 .....	298
4.4.5 手把手教你用 .....	225	6.3.3 手把手教你用 .....	300
课后强化练习 .....	229	6.4 时间序列分析及预测 .....	306
第5章 命令行界面 .....	231	6.4.1 使用时间序列环境 .....	306
5.1 命令行界面介绍 .....	232	6.4.2 手把手教你用 .....	318
5.1.1 命令调用 .....	233	课后强化练习 .....	326
5.1.2 命令自动完成 .....	234	第7章 Weka API .....	327
5.2 Weka 结构 .....	235	7.1 加载数据 .....	328
5.2.1 类实例和包 .....	235	7.1.1 从文件加载数据 .....	328
5.2.2 weka.core 包 .....	236	7.1.2 从数据库加载数据 .....	329
5.2.3 weka.classifiers 包 .....	237		

7.1.3	手把手教你用.....	330	7.10.2	手把手教你用.....	392
7.2	保存数据.....	335	7.11	文本分类综合示例.....	395
7.2.1	保存数据至文件.....	335	7.11.1	程序运行准备.....	395
7.2.2	保存数据至数据库.....	335	7.11.2	源程序分析.....	396
7.2.3	手把手教你用.....	336	7.11.3	运行说明.....	403
7.3	处理选项.....	339	课后强化练习.....		404
7.3.1	选项处理方法.....	339	<b>第 8 章 学习方案源代码分析.....</b>		<b>405</b>
7.3.2	手把手教你用.....	340	8.1	NaiveBayes 源代码分析.....	406
7.4	内存数据集处理.....	341	8.2	实现分类器的约定.....	427
7.4.1	在内存中创建数据集.....	341	课后强化练习.....		429
7.4.2	打乱数据顺序.....	345	<b>第 9 章 机器学习实战.....</b>		<b>431</b>
7.4.3	手把手教你用.....	345	9.1	数据挖掘过程概述.....	432
7.5	过滤.....	349	9.1.1	CRISP-DM 过程.....	432
7.5.1	批量过滤.....	350	9.1.2	数据预处理.....	433
7.5.2	即时过滤.....	351	9.1.3	挖掘项目及工具概述.....	434
7.5.3	手把手教你用.....	351	9.2	实战 KDD Cup 1999.....	434
7.6	分类.....	355	9.2.1	任务描述.....	435
7.6.1	分类器构建.....	355	9.2.2	数据集描述.....	436
7.6.2	分类器评估.....	356	9.2.3	挖掘详细过程.....	438
7.6.3	实例分类.....	358	9.3	实战 KDD Cup 2010.....	447
7.6.4	手把手教你用.....	359	9.3.1	任务描述.....	447
7.7	聚类.....	370	9.3.2	数据集描述.....	447
7.7.1	聚类器构建.....	370	9.3.3	挖掘详细过程.....	450
7.7.2	聚类器评估.....	371	9.3.4	更接近实际的挖掘过程.....	459
7.7.3	实例聚类.....	373	课后强化练习.....		471
7.7.4	手把手教你用.....	373	<b>附录 A 中英文术语对照.....</b>		<b>472</b>
7.8	属性选择.....	379	<b>附录 B Weka 算法介绍.....</b>		<b>476</b>
7.8.1	使用元分类器.....	380	过滤器算法介绍.....		476
7.8.2	使用过滤器.....	380	分类算法介绍.....		498
7.8.3	使用底层 API.....	381	聚类算法介绍.....		526
7.8.4	手把手教你用.....	381	关联算法介绍.....		530
7.9	可视化.....	384	选择属性算法介绍.....		532
7.9.1	ROC 曲线.....	385	<b>参考文献.....</b>		<b>537</b>
7.9.2	图.....	385			
7.9.3	手把手教你用.....	386			
7.10	序列化.....	391			
7.10.1	序列化基本方法.....	391			



## 第 1 章

# Weka 介绍

Weka 是新西兰怀卡托大学用 Java 开发的数据挖掘著名开源软件，该系统自 1993 年开始由新西兰政府资助，至今已经历了 20 多年的发展，其功能已经十分强大和成熟。Weka 集合了大量的机器学习和相关技术，受领域发展和用户需求所推动，代表了当今数据挖掘和机器学习领域的最高水平。

## 1.1 Weka 简介

Weka 是怀卡托智能分析环境(Waikato Environment for Knowledge Analysis)的英文字首缩写, 官方网址为 <http://www.cs.waikato.ac.nz/ml/weka>, 在该网站可以免费下载可运行软件和源代码, 还可以获得说明文档、常见问题解答、数据集和其他文献等资源。Weka 的发音类似新西兰本土一种不会飞的鸟, 如图 1.1 所示, 因此 Weka 系统使用该鸟作为其徽标。

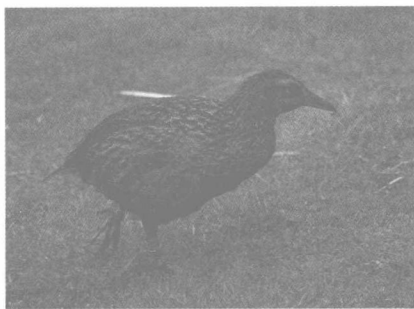


图 1.1 Weka(或 woodhen)鸟<sup>①</sup>

Weka 是一种使用 Java 语言编写的数据挖掘机器学习软件, 是 GNU 协议下分发的开源软件。Weka 主要用于科研、教育和应用领域, 还作为 Ian H. Witten、Eibe Frank 和 Mark A. Hall 三人合著的著名书籍 *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*<sup>②</sup> 的实践方面的重要补充, 该书于 2011 年由 Elsevier 出版。

Weka 是一套完整的数据处理工具、学习算法和评价方法, 包含数据可视化的图形用户界面, 同时该环境还可以比较和评估不同的学习算法的性能。

国内外很多著名大学都采用 Weka 作为数据挖掘和机器学习课程的实践工具。Weka 还有另外一个名字叫作 Pentaho Data Mining Community Edition(Pentaho 数据挖掘社区版), 此外, Pentaho 的网站(<http://weka.pentaho.com/>)还维护一个被称为 Pentaho Data Mining Enterprise Edition(Pentaho 数据挖掘企业版)的版本, 它主要提供技术支持和管理升级。另一个用 Java 编写的著名数据挖掘工具 RapidMiner 通过 Weka Extension(Weka 扩展)支持 Weka, 以充分利用 Weka 的“约 100 个额外的建模方案, 其中包括额外的决策树、规则学习器和回归估计器”, 参见网址 [https://marketplace.rapidminer.com/UpdateServer/faces/product\\_details.xhtml?productId=rmx\\_weka](https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_weka)。

有很多软件项目直接或间接使用 Weka, 包括能够处理 ARFF 格式的数据, 或者从其他编程环境中访问 Weka 的功能, 具体可参见网址 <http://www.cs.waikato.ac.nz/ml/weka/related.html>

① 来源: Weka\_a\_tool\_for\_exploratory\_data\_mining.ppt, [http://ncu.dl.sourceforge.net/project/weka/documentation/Initial%20upload%20and%20presentations/Weka\\_a\\_tool\\_for\\_exploratory\\_data\\_mining.ppt](http://ncu.dl.sourceforge.net/project/weka/documentation/Initial%20upload%20and%20presentations/Weka_a_tool_for_exploratory_data_mining.ppt)。

② 中译版书名为“数据挖掘: 实用机器学习工具与技术(原书第 3 版)”, 于 2014 年 5 月由机械工业出版社出版。



和 WekaWiki(<http://weka.wikispaces.com/Related+Projects>)。另外, Weka 还提供统计编程语言 R 以及 Python 的接口, 以及对分布式计算框架 Hadoop、Spark 的支持, 这使得 Weka 更具实用价值。

### 1.1.1 Weka 历史

怀卡托机器学习团队宣称: 我们团队的总体目标是要建立最先进的软件开发机器学习技术, 并将其应用于解决现实世界的数据挖掘问题。团队具体目标是: 使机器学习技术容易获得, 并将其应用到解决新西兰工业的重大实际问题中, 开发新的机器学习算法并推向世界, 为该领域的理论框架做出贡献。

1992 年年末, 新西兰怀卡托大学计算机科学系的 Ian H. Witten 博士申请基金, 1993 年获新西兰政府资助, 并于同年开发出接口和基础架构。次年发布了第一个 Weka 的内部版本。两年后, 在 1996 年 10 月, 第一个公开版本(Weka 2.1)发布。Weka 早期版本主要采用 C 语言编写, 1997 年年初, 团队决定使用 Java 重新改写, 并在 1999 年中期发布纯 Java 的 Weka 3 版本。选定 Java 来实现 Ian H. Witten 著作 *Data Mining* 的配套机器学习技术是有充分理由的, 作为一个著名的面向对象的编程语言, Java 允许用一个统一的接口来进行学习方案和方法的预处理和后处理。决定使用 Java 来替代 C++或其他面向对象的语言, 是因为 Java 编写的程序可以运行在绝大部分计算机上, 而无须重新编译, 更不需要修改源代码。已经测试过的平台包括 Linux、Windows 和 Macintosh 操作系统, 甚至包括 PDA。最后的可执行程序复制过来即可运行, 完全绿色, 不要求复杂安装。当然, Java 也有其缺点, 最大的问题是它在速度上有缺陷, 执行一个 Java 程序比对应的 C 语言程序要慢上好几倍。综合来看, 对于 Weka 来说, Java “一次编译, 到处运行”的吸引力远远超出对性能的渴望。

截止到 2016 年 2 月, Weka 最新的版本是 3.7.13, 这是 2015 年 9 月 11 日发布的稳定版, 本书第二版基于该版本。

### 1.1.2 Weka 功能简介

Weka 系统汇集了最前沿的机器学习算法和数据预处理工具, 以使用户能够快速灵活地将已有的成熟处理方法应用于新的数据集。它为数据挖掘的整个过程提供全面的支持, 包括准备输入数据、统计评估学习方案、输入数据和学习效果的可视化。Weka 除了提供大量学习算法(学习方案)之外, 还提供了适应范围很广的预处理工具, 用户通过一个统一界面操作各种组件, 比较不同的学习算法, 找出能够解决问题的最有效的方法。

Weka 系统包括处理标准数据挖掘问题的所有方法: 回归、分类、聚类、关联规则以及属性选择。分析要进行处理的数据是重要的一个环节, Weka 提供了很多用于数据可视化和预处理的工具。输入数据可以有两种形式, 第一种是通过以 ARFF 格式为代表的文件进行输入, 另一种是直接读取数据库表。

使用 Weka 的方式主要有三种: 第一种是将学习方案应用于某个数据集, 然后分析其输出, 从而更多地了解这些数据; 第二种是使用已经学习到的模型对新实例进行预测; 第三种是使用多种学习方案, 然后根据其性能表现选择其中的一种来进行预测。用户使用交互式界面菜单选择一种学习方案, 大部分学习方案都带有可调节的参数, 用户可通过属性

列表或对象编辑器修改参数, 然后通过同一个评估模块对学习方案的性能进行评估。

Weka 主界面被称为 Weka GUI 选择器(Weka GUI Chooser), 它通过右边的四个按钮提供四种主要的应用程序供用户选择, 如图 1.2 所示, 单击按钮即可进入相应的图形用户界面。

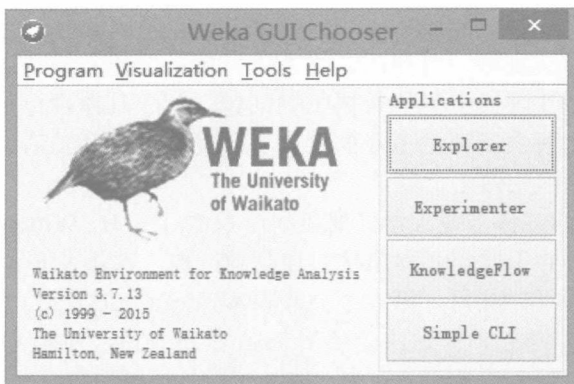


图 1.2 Weka 主界面

其中, Weka 系统提供的最容易使用的图形用户界面称为探索者(Explorer)。通过选择菜单和填写表单, 可以调用 Weka 的所有功能。例如, 用户仅仅单击几个按钮, 就可以完成从 ARFF 文件中读取数据集, 然后建立决策树模型的工作。Weka 界面十分友好, 能适时地将不宜用的功能选项设置为不可用状态; 将用户选项设计为表格方式以方便填写; 当鼠标移动到界面工具上短暂停留时, 会给出用法提示; 对算法都给出较为合理的默认值, 使得用户在进行配置时无须花费太多精力就可取得较好的效果等。

虽然探索者界面使用很方便, 但它存在一个缺陷, 即要求将所需数据全部一次读进内存, 一旦用户打开某个数据集, 就会批量读取全部数据。因此, 这种批量方式仅适合处理中小规模的问题。而知识流界面刚好能够弥补这一缺陷。

知识流(KnowledgeFlow)界面可以使用增量(分批)方式的算法来处理大型数据集, 用户可以定制处理数据流的方式和顺序。知识流界面允许用户在屏幕上任意拖曳代表学习算法和数据源的图形组件, 并以一定的方式和顺序组合在一起。也就是说, 按照一定顺序将代表数据源、预处理工具、学习算法、评估手段和可视化模块的各组件组合在一起, 形成数据流。如果用户选取的过滤器和学习算法具有增量学习功能, 那就可以实现大型数据集的增量分批读取和处理。

实验者(Experimenter)界面用于帮助用户解答实际应用分类和回归技术中遇到的一个基本问题: 对于一个已知问题, 哪种方法及参数值能够取得最佳效果? 通过 Weka 提供的实验者工作环境, 用户可以比较不同的学习方案。尽管探索者界面也能通过交互完成这样的功能, 但通过实验者界面, 用户可以让处理过程实现自动化。实验者界面更加容易使用不同参数去设置分类器和过滤器, 使之运行在一组数据集中, 收集性能统计数据, 实现重要的测试实验。

简单命令行(Simple CLI)界面是为不提供自己的命令行界面的操作系统提供的, 该界面用于和用户进行交互, 可以直接执行 Weka 命令。

## 1.2 基本概念

上节简要介绍了 Weka，读者也许迫不及待地想进一步深入了解并使用 Weka 来完成数据挖掘工作。但是，在此之前，有必要先了解数据挖掘和机器学习的一些基本概念，为进一步的学习打下基础。

### 1.2.1 数据挖掘和机器学习

数据挖掘和机器学习这两项技术的关系非常密切。机器学习方法构成数据挖掘的核心，绝大多数数据挖掘技术都来自机器学习领域，数据挖掘又向机器学习提出新的要求和任务。

数据挖掘就是在数据中寻找模式的过程。这个寻找过程必须是自动的或半自动的，并且数据总量应该具有相当大的规模，从中发现的模式必须有意义并能产生一定的效益。通常，数据挖掘需要分析数据库中的数据来解决问题，如客户忠诚度分析、市场购物篮分析等。

当今已进入海量数据时代。例如，全世界已经有约一万亿个网页；沃尔玛仅一个小时就有一百万的交易量，其数据库里的数据已有 2.5 拍(即  $2.5 \times 10^{15}$  字节)的信息，等等。

这些海量数据不可能采用手工方式进行处理，因此，迫切要求能进行数据分析的自动化方法，这些都由机器学习提供。

机器学习定义为能够自动寻找数据中的模式的一套方法，然后，使用所发现的模式来预测将来的数据，或者在各种不确定的条件下进行决策。

机器学习分为两种主要类型。第一种机器学习类型称为有监督学习，或称为预测学习，其目标是在给定一系列输入/输出实例所构成的数据集的条件下，学习输入  $x$  到输出  $y$  的映射关系。这里的数据集称为训练集，实例的个数称为训练样本数。第二种机器学习类型称为无监督学习，或称为描述学习，其目标是在给定一系列仅由输入实例构成的数据集的条件下，发现数据中的有趣模式。无监督学习有时候也称为知识发现，这类问题并没有明确定义，因为我们不知道需要寻找什么样的模式，也没有明显的误差度量可供使用。而对于给定的  $x$ ，有监督学习可以对所观察到的值与预测的值进行比较，得到明确的误差值。

### 1.2.2 数据和数据集

根据应用的不同，数据挖掘的对象可以是各种各样的数据，这些数据可以以各种形式进行存储，如数据库、数据仓库、数据文件、流数据、多媒体、网页等。既可以集中存储在数据存储库中，也可以分布在世界各地的网络服务器上。

通常将数据集视为待处理的数据对象的集合。由于历史原因，数据对象有多个别名，如记录、点、行、向量、案例、样本、观测等。数据对象也是对象，因此，可以用刻画对象基本特征的属性来进行描述。属性也有多个别名，如变量、特征、字段、维、列等。

数据集可以类似于一个二维的电子表格或数据库表。在最简单的情形下，每个训练输入  $x_i$  是一个  $N$  维的数值向量，表示特定事物的一些特征，如人的身高、体重。这些特征也

可以称为属性。有时  $x_i$  也可以是复杂结构的对象,如图像、电子邮件、时间序列、语句等。

属性可以分为四种类型:标称(nominal)、序数(ordinal)、区间(interval)和比率(ratio),其中,标称属性的值仅仅是不同的名称,即标称值仅提供区分对象的足够信息,如性别(男、女)、衣服颜色(红、黄、蓝)、天气(阴、晴、雨、多云)等;序数属性的值可以提供确定对象的顺序的足够信息,如成绩等级(优、良、中、及格、不及格)、职称(初级、中级、高级)、学生(本科生、硕士生、博士生)等;区间属性的值之间的差是有意义的,即存在测量单位,如温度、日历日期等;比率属性的值之间的差和比值都是有意义的,如绝对温度、年龄、长度、成绩分数等。

标称属性和序数属性统称为分类的(categorical)或定性的(qualitative)属性,它们的取值为集合,即使使用数值来表示,也不具备数的大部分性质,因此,应该像对待符号一样对待;区间属性和比率属性统称为定量的(quantitative)或数值的(numeric)属性,定量属性采用数值来表示,具备数的大部分性质,可以使用整数或连续实数值来表示。

大部分数据集都以数据库表和数据文件的形式存在,Weka 支持读取数据库表和多种格式的数据文件,其中,使用最多的是一种称为 ARFF 格式的文件。

### 1.2.3 ARFF 格式

ARFF 是一种 Weka 专用的文件格式,由 Andrew Donkin 创立,有传言说 ARFF 代表 Andrew's Ridiculous File Format(安德鲁的荒唐文件格式),但在 Weka 的正式文档中明确说明 ARFF 代表 Attribute-Relation File Format(属性—关系文件格式)。该文件是 ASCII 文本文件,描述共享一组属性结构的实例列表,由独立且无序的实例组成,是 Weka 表示数据集的标准方法,ARFF 不涉及实例之间的关系。

在 Weka 安装目录下的 data 子目录中,可以找到名称为 weather.numeric.arff 的天气数据文件,其内容如程序清单 1.1 所示。数据集是实例的集合,每个实例包含一定的属性,属性的数据类型包括如下几类:标称型(nominal),只能取预定义值列表中的一个;数值型(numeric),只能是实数或整数;字符串型(string),这是一个由双引号引用的任意长度的字符列表;另外还有日期型(date)和关系型(relational)。ARFF 文件就是实例类型的外部表示,其中包括一个标题头(header),以描述属性的类型,还包含一个用逗号分隔的列表所表示的数据部分(data)。

#### 程序清单 1.1 天气数据的 ARFF 文件

```
% This is a toy example, the UCI weather dataset.
% Any relation to real weather is purely coincidental.

@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
```



```

@data
sunny, 85, 85, FALSE, no
sunny, 80, 90, TRUE, no
overcast, 83, 86, FALSE, yes
rainy, 70, 96, FALSE, yes
rainy, 68, 80, FALSE, yes
rainy, 65, 70, TRUE, no
overcast, 64, 65, TRUE, yes
sunny, 72, 95, FALSE, no
sunny, 69, 70, FALSE, yes
rainy, 75, 80, FALSE, yes
sunny, 75, 70, TRUE, yes
overcast, 72, 90, TRUE, yes
overcast, 81, 75, FALSE, yes
rainy, 71, 91, TRUE, no

```

上述代码中，以百分号“%”开始的行称为注释行。与计算机编程语言类似，最前面的注释行应该写明数据集的来源、用途和含义。

`@relation` 行定义内部数据集的名称 `weather`，名称应简洁明了，尽可能容易理解。`relation` 也称为关系。

`@attribute outlook {sunny, overcast, rainy}` 行定义名称为 `outlook` 的标称型属性，有三个取值：`sunny`、`overcast` 和 `rainy`。按照同样的方式，`@attribute windy {TRUE, FALSE}` 行和 `@attribute play {yes, no}` 行分别定义 `windy` 和 `play` 两个标称型属性。要注意的是，最后一个属性默认为用于预测的类别变量，或称为目标属性。本例中，类别变量为标称型属性 `play`，它只能取两个值之一，使得天气问题成为二元(binary)的分类问题。

`@attribute temperature real` 行定义名称为 `temperature` 的数值型属性，`@attribute humidity real` 行定义名称为 `humidity` 的数值型属性。这两个属性都是实数型。

`@data` 标志后的各行构成数据集。每行为一个实例样本，由采用逗号分隔的值组成，顺序与由 `@attribute` 所定义属性的顺序一致。

本例没有使用字符串类型和日期类型，在将来的学习中会遇到这两种类型。

## 1.2.4 预处理

数据挖掘是在大量的、潜在有用的数据中挖掘出有用模式的过程。因此，源数据的质量直接影响到挖掘的效果，高质量的数据是进行有效挖掘的前提。但是，由于数据挖掘所使用的数据往往不是专门为挖掘准备的，期望数据质量完美并不现实，人的错误、测量设备的限制以及数据收集过程的漏洞都可能导致一些问题，如缺失值(由机械原因或人为原因造成的数据缺失)和离群值(数值与其他数值相比差异较大)。

由于无法在数据的源头控制质量，数据挖掘只能通过以下两个方面设法避免数据质量问题：①数据质量问题的检测与纠正；②使用能容忍低质量数据的算法。第一种方式在数据挖掘前检测并纠正一些质量问题，这个过程称为数据预处理；第二种方式需要提高算法的健壮性。

数据预处理是数据挖掘的重要步骤，数据挖掘者的大部分时间和精力都要花在预处理