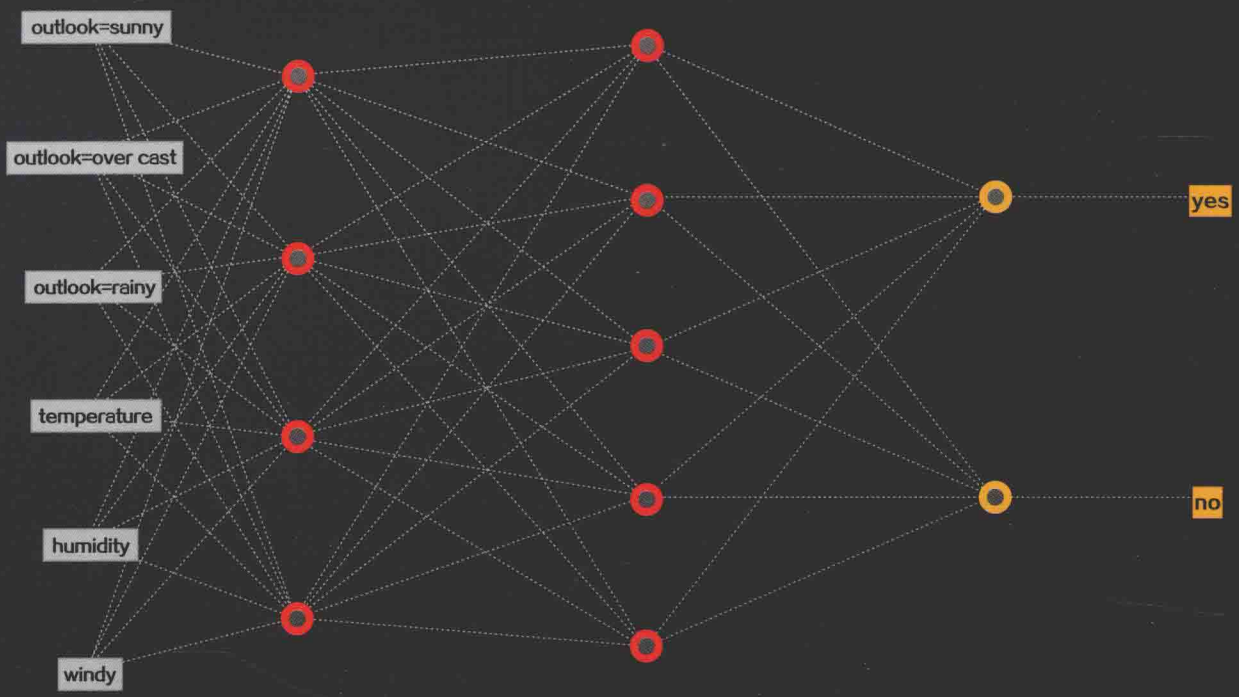
系统讲解数据挖掘机器学习工具Weka
经典的开源挖掘工具、开放的Java环境
初学者的入门首选、研究者的钻研利器



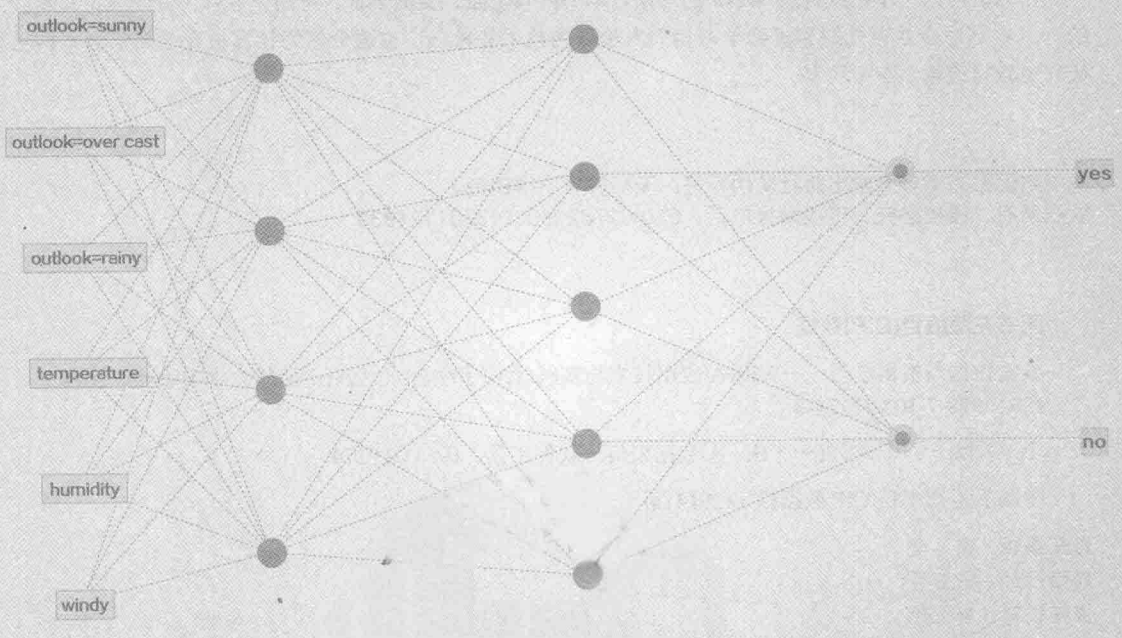
数据挖掘与机器学习

WEKA

应用技术与实践

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

袁梅宇 编著

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 8 章，主要内容包括 Weka 介绍、Explorer 界面、Knowledge Flow 界面、Experimenter 界面、命令行界面、Weka 高级应用、Weka API 和学习方案源代码分析。

作为国内第一本系统讲解 Weka 的书籍，本书内容全面、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，也适合作为计算机专业高年级本科生和研究生教材或教学参考用书。

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. —北京：清华大学出版社，2014
ISBN 978-7-302-37174-8

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(2014)第152046号

责任编辑：魏　莹
封面设计：杨玉兰
责任校对：李玉萍
责任印制：沈　露

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
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn,010-62791865

印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：29.25　　字　　数：708 千字
版　　次：2014 年 7 月第 1 版　　印　　次：2014 年 7 月第1 次印刷
印　　数：1～3000
定　　价：52.80 元

产品编号：051213-01

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，其算法是一类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。机器学习和数据挖掘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，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 年的发展，它的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或实际意义。

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 Knowledge Flow 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，以及分析实验结果；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，介绍如何使用 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；最后一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，为开发人员提供一个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，因此肯定会有所遗漏及不妥之处，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，他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

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 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[①]

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、Frank Eibe 和 Mark A. Hall 三人合著的著名书籍 ——《Data Mining — Practical Machine Learning Tools and Techniques, Third Edition》(数据挖掘：实用机器学习工具与技术，第 3 版)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 http://rapid-i.com/content/view/202/206/。

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

① 来源：Weka_a_tool_for_exploratory_data_mining.ppt。http://sourceforge.net/projects/weka/files/documentation/Initial%20upload%20and%20presentations/Weka_a_tool_for_exploratory_data_mining.ppt/download?use_mirror=ncu

易获得，并将其应用到解决新西兰工业的重大实际问题，开发新的机器学习算法并推向世界，为该领域的理论框架作出贡献。

1992年末，新西兰怀卡托大学计算机科学系Ian 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 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2013年2月，Weka最新的版本是3.7.8，这是2013年1月24日发布的稳定版，本书基于该版本。

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，以便用户能够快速灵活地将已有的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，第一种是以ARFF格式为代表的文件；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器，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中选择一种学习方法，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主界面称为Weka GUI选择器，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，用鼠标单击按钮进入到相应的图形用户界面。

其中，Weka系统提供的最容易使用的图形用户接口称为探索者(Explorer)。通过选择菜单和填写表单，可以调用Weka的所有功能。例如，用户用鼠标仅仅单击几个按钮，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选；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，这

样，帮助用户尽量少花精力进行配置就可取得较好的效果等。

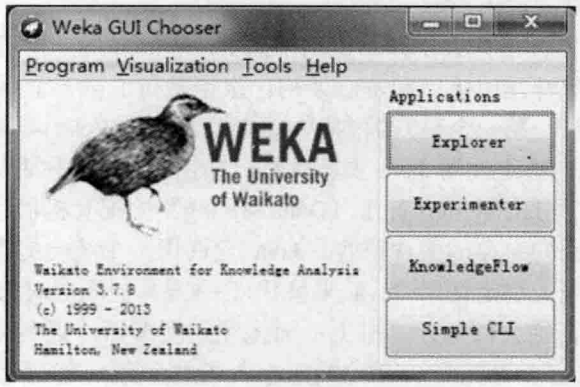


图 1.2　Weka 主界面

虽然探索者界面使用很方便，但它也存在一个缺陷，要求它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。知识流刚好能够弥补这一缺陷。

知识流(KnowledgeFlow)界面可以使用增量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构件，并以一定的方式和顺序组合在一起。也就是，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构件组合在一起，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，该简单命令行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，为进一步地学习打下基础。

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，并且数据总量应该是具有相当大的规模，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，等等。

当今已进入海量数据时代。例如，全世界已经有约 1 000 000 000 000 个网页；沃尔玛仅一个小时就有一百万的交易量，其数据库里数据已有 2.5 拍(即 2.5×10^{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，在给定一系列仅由输入实例构成的数据集的条件下，其目标是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。

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存储，如数据库、数据仓库、数据文件、流数据、多媒体、网页，等等。即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(初职、中职、高职)、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，可以使用整数值或连续值来表示。

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；数字型(numeric)，只能是实数或整数；字符串(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 一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，最后一个属性缺省为用于预测的类别变量。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 定义名称为 temperature 的数值型属性，@attribute humidity real 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，如缺失值和离群值。

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阶段。Weka 专门提供若干过滤器进行预处理，还在探索者界面中提供选择属性标签页专门处理属性的自动选择问题。数据预处理涉及的策略和技术非常广泛，主要包括如下技术。

1) 聚集

聚集(Aggregation)就是将两个或多个对象合并为单个对象。一般来说，定量数据通常通过求和或求平均值的方式进行聚集，定性数据通常通过汇总进行聚集。聚集通过数据归约来减少数据量，所导致的较小数据集只需要较少内存和处理时间的开销，因此，可以使用开销更大的数据挖掘算法。另外，聚集使用高层数据视图，起到了范围或度量转换的作用。虽然站在很高的角度去检视问题容易避免只见树木不见森林的问题，但也可能导致有趣细节的丢失。

2) 抽样

如果处理全部数据的开销太大，数据预处理可以使用抽样，只选择数据对象的子集进行分析。使用抽样可以压缩数据量，因此，能够使用效果更好但开销较大的数据挖掘算法。由于抽样是一个统计过程，好的抽样方案就是确保以很高的概率得到有代表性的样本，即：样本近似地具有原数据相同的性质。

抽样方式有多种，最简单的抽样是选取每一个数据行作为样本的概率都相同，这称为简单随机抽样，又分为有放回抽样和无放回抽样两种形式，前者是从 N 个数据行中以概率 $1/N$ 分别随机抽取出 n 个数据行，构成样本子集；后者与有放回抽样的过程相似，但每次都要删除原数据集中已经抽取出来的数据行。显然，有放回抽样得到的样本子集有可能重复抽取到相同的数据行。

当整个数据集由差异较大的数据行构成时，简单随机抽样可能无法抽取到不太频繁出现的数据行，这会导致得到的样本不具代表性。分层抽样(Stratified Sampling)尽量利用事先掌握的信息，充分考虑保持样本结构和总体结构的一致性以提高样本的代表性。其步骤是，先将数据集按某种特征分为若干不相交的“层”，然后再从每一层中进行简单随机抽样，从而得到具有代表性的抽样数据子集。

3) 维度归约

维度是指数据集中属性的数目。维度归约(Dimension Reduction)是指创建新属性，通过数据编码或数据变换，将一些旧属性合并在一起以降低数据集的维度。

维度归约可以删除不相关的属性并降低噪声，维度降低会使许多数据挖掘的算法变得更好，还能消除了维灾难带来的负面影响。维灾难是指，随着维度的增加，数据在它所占的空间越来越稀疏，对于分类问题，这意味着可能没有足够的数据对象来创建模型；对于聚类问题，点之间的密度和距离的定义失去意义。因此，对于高维数据，许多分类和聚类等学习算法的效果都不理想。维度归约使模型的属性更少，因而可以产生更容易理解的模型。

4) 属性选择

除维度归约外，降低维度的另一种方法是仅只使用属性的一个子集。表面看来似乎这种方法可能丢失信息，但很多情况下，数据集存在冗余或不相关的属性。其中，冗余属性是指某个属性包含了其他属性中的部分或全部信息，不相关属性是指对于手头数据挖掘任务几乎完全没有用处的信息。属性选择是指从数据集中选择最具代表性的属性子集，删除冗余或不相关的属性，从而提高数据处理的效率，使模型更容易理解。

最简单的属性选择方法是使用常识或领域知识，以消除一些不相关或冗余属性，但是，选择最佳的属性子集通常需要系统的方法。理想的属性选择方法是：将全部可能的属性子集作为数据挖掘学习算法的输入，然后选取能产生最好结果的子集。这种方法反映了对最终使用的数据挖掘算法的目的和偏爱。但是，由于 n 个属性的子集的数量多达 2^n 个，大部分情况下行不通。因此，需要考虑三种标准的属性选择方法：嵌入、过滤和包装。

嵌入方法(Embedded Approach)将属性选择作为数据挖掘算法的一部分。在挖掘算法运行期间，算法本身决定使用哪些属性以及忽略哪些属性。决策树算法通常使用这种方法。

过滤方法(Filter Approach)在运行数据挖掘算法之前，使用独立于数据挖掘任务的方法进行属性选择，即：先过滤数据集产生一个属性子集。

包装方法(Wrapper Approach)将学习算法的结果作为评价准则的一部分，使用类似于前文介绍的理想算法，但通常无法枚举出全部可能的子集以找出最佳属性子集。

根据属性选择过程是否需要使用类别信息，属性选择可分为有监督属性选择和无监督属性选择。前者通过度量类别信息与属性之间的相互关系来确定属性子集，后者不使用类别信息，使用聚类方法评估属性的贡献度，根据贡献度来确定属性子集。

5) 属性创建

属性创建就是通过对数据集中旧的属性进行处理，创建新的数据集，这样能更有效的获取重要的信息。由于通常新数据集的维度比原数据集少，因此，可以获得维度归约带来的好处。属性创建有三种方法：属性提取、映射数据到新空间和属性构造。

属性提取是指由原始数据创建新的属性集。例如，对照片数据进行处理，提取一些较高层次的特征，诸如与人脸高度相关的边和区域等，就可以使用更多的分类技术。

映射数据到新空间，是指使用一种完全不同的视角挖掘数据可能揭示重要而有趣的特征。例如，对时间序列实施傅立叶变换，转换为频率信息，可能检测到其中的周期模式。

当原始数据集的属性含有必要信息，但其形式不适合数据挖掘算法的时候，可以使用属性构造，将一个或多个原来的属性构造出新的属性。

6) 离散化和二元化

有的数据挖掘算法，尤其是某些分类算法，要求数据是分类属性的形式。发现关联模式的算法要求数据是二元属性的形式。因此，需要进行属性变换，将连续属性转换为分类属性称为离散化(Discretization)，将连续和离散属性转换为一个或多个二元属性称为二元化(Binarization)。

连续属性离散化为分类属性分为两个子任务：决定需要多少个分类值，以及如何确定将连续属性值映射到这些分类值中。因此，离散化问题就是决定选择多少个分割点，以及确定分割点的位置。利用少数分类值标签替换连续属性的值，从而减少和简化原来的数据。

根据是否使用类别信息，可以将离散化技术分为两类：使用类别信息的称为有监督的离散化，反之则称为无监督的离散化。

等宽和等频离散化是两种常用的无监督的离散化方法。等宽(Equal Width)离散化将属性的值域划分为相同宽度的区间，区间的数目由用户指定。这种方式常常会造成实例分布不均匀。等频(Equal Frequency)离散化也称为等深(Equal Depth)离散化，它试图将相同数量的对象放进每个区间，区间的数目由用户指定。

7) 变量变换

变量变换(Variable Transformation)也称为属性变换，是指用于变量的所有值的变换。下面讨论两种重要的变量变换：简单函数变换和规范化。

简单函数变换是使用一个简单数学函数分别作用于每一个值。在统计学中，使用平方根、对数变换和倒数变换等变量变换常用于将不具有高斯分布的数据变换为具有高斯分布的数据。

变量的标准化(Standardization)是使整个值的集合具有特定的性质。例如，假如$\bar{x}$是某个属性的均值，S_X是其标准差，则变换公式$x'=(x-\bar{x})/s_x$创建一个具有均值 0 和标准差 1 的新的变量。由于均值和标准差受离群点的影响较大，因此，常常修正上述变换。例如，用中位数(Median)替代均值，用绝对标准差替代标准差，等等。