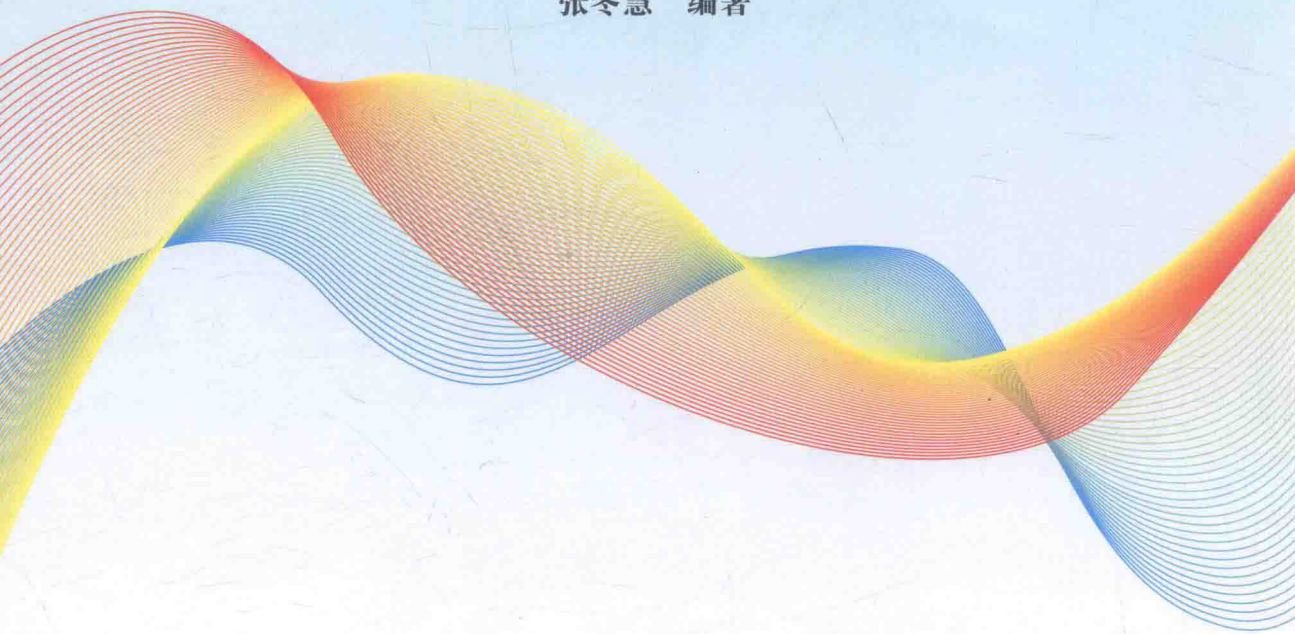
基于Rattle的可视化数据挖掘技术

张冬慧　编著



清华大学出版社

基于Rattle的可视化数据挖掘技术

张冬慧　编著

清华大学出版社

内 容 简 介

数据挖掘技术近年来发展异常迅猛，已成为大数据时代最热门的技术和研究热点，不仅产生了大量不同类型、功能强大的数据挖掘算法，而且推动了众多数据挖掘工具软件的发展。在这些软件中，R 语言是数据挖掘领域最重要的软件之一。Rattle 是一种用于数据挖掘的 R 语言的图形交互界面，或称为可视化数据挖掘工具。Rattle 给出了从数据整理到模型评价的完整解决方案。

本书主要介绍如何用 Rattle 包进行数据挖掘，全书共 9 章，通过大量精选实例，循序渐进、全面系统地讲述数据挖掘过程。

本书不仅是从事数据挖掘和大数据分析工程技术人员开发相关系统的技术资料，也可作为学习数据挖掘和大数据分析等课程的参考用书。

图书在版编目(CIP)数据

基于 Rattle 的可视化数据挖掘技术/张冬慧编著. —北京：清华大学出版社，2017
ISBN 978-7-302-47432-6

Ⅰ. ①基… Ⅱ. ①张… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 129477 号

责任编辑：龙启铭　张爱华
封面设计：何凤霞
责任校对：徐俊伟
责任印制：李红英

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,010-62795954
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185mm×230mm　　**印　　张**：11.75　　**字　　数**：256 千字
版　　次：2017 年 8 月第 1 版　　**印　　次**：2017 年 8 月第1 次印刷
印　　数：1～2000
定　　价：39.00 元

产品编号：072766-01

前言

数据挖掘是指从大量数据中通过各种算法挖掘知识的过程，是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解模式的过程。近年来，数据挖掘技术发展异常迅猛，不仅产生了大量不同类型、功能强大的数据挖掘算法，而且推动了众多数据挖掘工具软件的发展。在这些软件中，R 语言已悄然成为数据挖掘领域最重要的软件之一。R 语言是一个包含众多学科、工程统计的庞大系统，是目前世界上流行的统计软件之一。R 语言既是用于统计计算和统计制图的优秀工具，又是大数据分析和挖掘的重要工具。

R 语言得到全球顶级的统计学家支持，并实现了数据挖掘的所有关键算法。本书介绍了基于 R 语言开发的自由开源软件包 Rattle 进行数据挖掘的基本过程。Rattle 包的源代码对每个人都是可见的，没有限制，任何人都可以扩展它。

本书将引导读者通过 Rattle 包提供的各种选项完成数据挖掘任务。许多例子深入到 R 语言编程，目的是鼓励读者直接使用 R 语言作为脚本语言，通过 R 脚本实现数据挖掘所需要的基本技能。

本书程序很少依赖对某些计算机编程语言的熟练程度。即使没有计算机编程经验，也能从本书受益。不过，还是鼓励所有读者熟悉使用某种计算机编程语言来处理和分析数据的方法。

对大多数读者而言，本书容易理解，不需要很深的计算机和统计学背景知识。本书也会介绍一些较为复杂的统计、数学和程序设计概念，但主要的原则是保持简单，这意味着简化概念，且在不失去概念内涵的前提下保证概念的准确性。

Rattle 易学易用，不要求很多的 R 语言基础，被广泛应用于数据挖掘实践和教学之中。即使对 R 语言不是很了解的用户，也可以通过简单的鼠标点击来读入、转换、探索数据。而且，用户可以在 Log 中了解 Rattle 所使用的 R 语言命令记录。

全书共 9 章，内容包括绪论、入门指南、数据准备、数据理解、数据检验、数据变换、数据建模、模型评估、模型部署。

由于篇幅限制，本书并不能涵盖数据挖掘全部内容，读者可以通过《数理统计和数据分

析》扩展统计学方面的知识；通过《线性回归分析导论》扩展线性回归方面的知识；通过《统计建模与 R 语言》扩展 R 在统计中应用方面的知识。若想更深入地了解 R 语言在数据挖掘中的应用，推荐参考《R 语言与数据挖掘——最佳实践和经典案例》；若想了解数据挖掘可视化 R 语言实现，推荐参考《R 数据可视化手册》。

感谢南通大学程显毅教授在资料整理过程中所做的工作。感谢北京信息科技大学计算中心给予的支持。

R 语言是正在蓬勃发展的编程语言，其在数据挖掘领域的应用还有一些有价值的新内容来不及收入本书。加之编者知识水平和实践经验有限，书中难免存在不足之处，敬请读者批评指正。

编　者

2017 年 3 月于北京

目 录

第1章

绪　论

1.1　数据挖掘的认识

1.1.1　为什么要进行数据挖掘

在大数据时代下，数据挖掘已渗透到各行各业。大数据不仅仅是一个概念、一个话题，而且正在成为人们生活中的一部分：用大数据预测疾病，用大数据助力企业的商业决策，用大数据分析客户心理。

所谓数据挖掘(Data Mining，DM)，是指从大量不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、有用的信息和知识的过程。其表现形式为概念、规则、模式等。

谈到发现模式与规则，其实就是一项业务流程，为业务服务。人们要做的是让业务做起来更简单，或直接帮助客户提升业务，在大量的数据中找到有意义的模式和规则。在大量数据面前，数据的获得不再是一个障碍，而是一个优势。现在很多技术在大数据集上比在小数据集上的表现更好——可以用数据产生智慧，也可以用计算机来完成其最擅长的工作。

数据挖掘是一种决策支持过程，主要基于人工智能、机器学习、模式识别、统计学、数据库、可视化技术等，高度自动化地分析企业的数据，做出归纳性的推理，从中挖掘出潜在的模式，帮助决策者调整市场策略，减少风险，做出正确的决策。

1.1.2　数据挖掘过程

一个完整的数据挖掘过程包括以下几个步骤(见图 1.1)。

1. 数据准备

数据挖掘的处理对象是数据，这些数据一般存储在数据库系统中，是长期积累的结果。不是在任何数据上都能进行挖掘。首先要清除数据噪声和与挖掘主题明显无关的数据；其

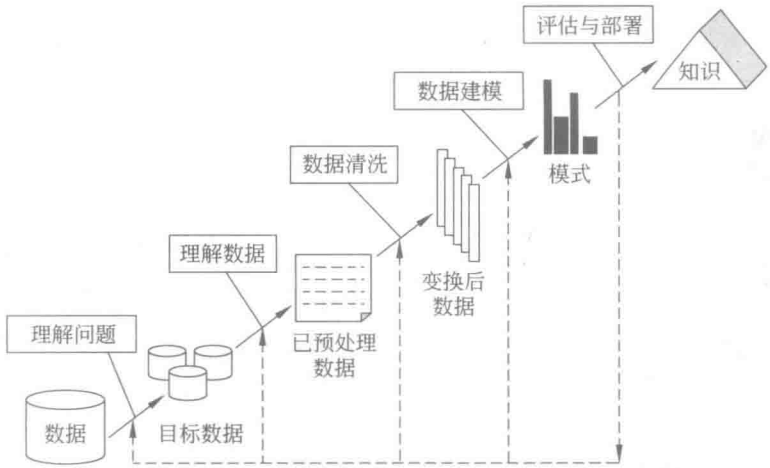


图 1.1 数据挖掘过程

次将来自多数据源中的相关数据组合及合并；然后将数据转换为易于进行数据挖掘的数据存储形式：这就是数据准备。数据准备是数据挖掘的第一步，它是整个过程中很重要的一步。数据准备是否合适将影响到数据挖掘的效率、准确率以及最终模式的有效性。

数据准备可细分为三个步骤：

(1) 理解问题(熟悉业务，搞清楚要解决的业务问题)。

(2) 理解数据(定义业务问题的变量，提取所需的数据)。

(3) 数据清洗(导入数据，并对数据做量纲调整、缺失值补充的处理)。

2. 数据建模

数据挖掘就是根据数据挖掘的目标，选取相应算法及参数，分析准备好的数据，产生一个特定的模式或数据集，从而得到可能形成知识的模式模型。

图 1.2 给出了一些常用数据挖掘模型。

3. 评估与部署

由挖掘算法产生的模式，存在无实际意义或无实用价值的情况，也存在不能准确反映数据的真实意义的情况，甚至在某些情况下与事实相反，因此，需要对其进行评估，从挖掘结果中筛选出有意义的模式。在此过程中，为了取得更为有效的知识，可能会返回前面的某一处理步骤中进行反复提取，从而提取出更有效的知识。模型评估的目标是完成对知识的一致性检查，确保发现的知识与已知可信的知识不发生抵触。

发现知识的目的是运用(模型部署)。运用知识有两种方法：一种是直接运用知识来决策；另一种是要求对新的数据运用知识，由此可能产生新的问题，从而需要对知识做进一步

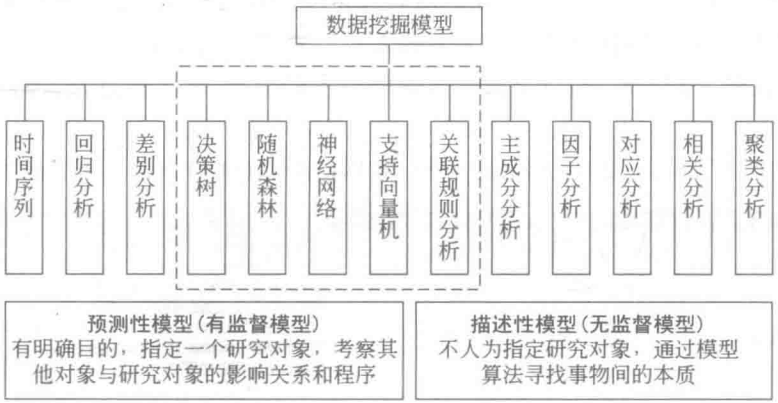


图 1.2　常用数据挖掘模型

的优化。

1.1.3　数据挖掘九大定律

数据挖掘九大定律如下：

(1) 目标律：业务目标是所有数据解决方案的源头。

(2) 知识律：业务知识是数据挖掘过程每一步的核心。

(3) 准备律：数据预处理比数据挖掘其他任何一个过程都重要。

(4) 试验律：对于数据挖掘者来说，天下没有免费的午餐，一个正确的模型只有通过试验才能被发现。

(5) 模式律：数据中总含有模式。

(6) 洞察律：数据挖掘增大了对业务的认知。

(7) 预测律：预测提高了信息泛化能力。

(8) 价值律：数据挖掘结果的价值取决于模型的稳定性或预测的准确性。

(9) 变化律：所有的模式均因业务变化而变化。

1.2　R 与 Rattle

1.2.1　R 语言

R 是用于统计分析、绘图的语言和操作环境。R 由来自新西兰奥克兰大学的 Ross Ihaka 和 Robert Gentleman 开发(也因此称为 R)，现在由 R 开发核心团队负责。R 是基于 S 语言的一个 GNU 项目，所以也可以当作 S 语言的一种实现。通常，用 S 语言编写的代码不做修改就可以在 R 环境下运行。

R 是一套完整的数据处理、计算和制图软件系统。其功能包括：数据存储和处理系统；数组运算工具(其向量、矩阵运算功能尤其强大)；完整连贯的统计分析工具；优秀的统计制图功能；简便而强大的编程语言(可操纵数据的输入和输出，可实现分支、循环，用户可自定义功能)。

图 1.3 列出了 2012—2014 年数据分析、数据挖掘、数据科学常用语言排行榜。

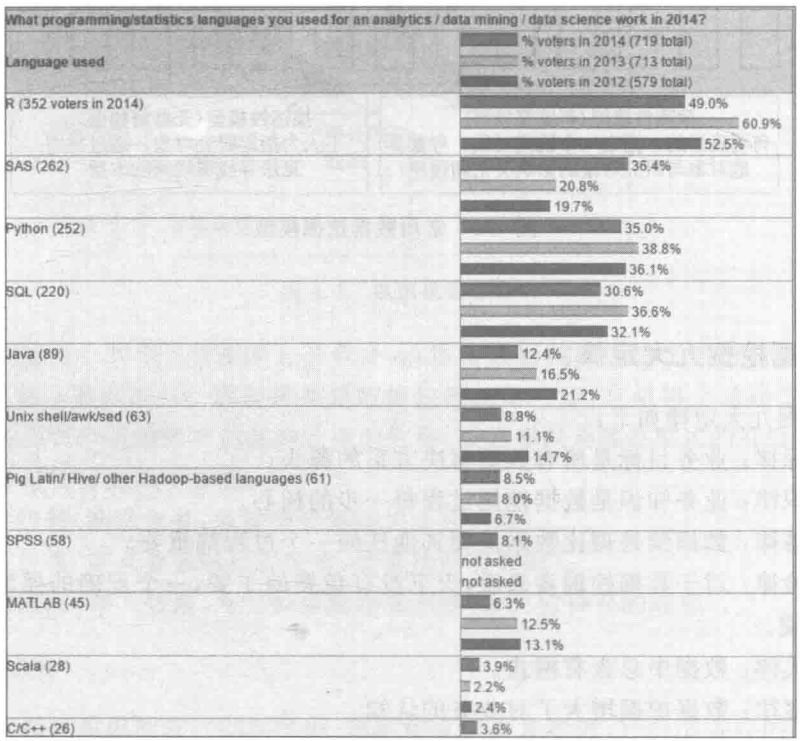


图 1.3 常用语言排行榜

从图 1.3 可以看出，R 语言受到了越来越多数据科学研究者的认可。

1.2.2 R 语言的基本语法

1. 基本概念

(1) 赋值：<-(也可用＝、->代替)。

例如，c <-a＋b，c ＝a＋b，a＋b->c 是等价的。

(2) 注释：＃ABC。

(3) 变量：无须定义，但区分大小写，例如 China 和 china 是不同的。

2. 对象

(1) 向量(Vector)：一系列元素的组合。

产生向量：V_1<-c(1,2,2)；V_2<-c("a","a","b","b","c")。

求向量长度：length()。例如：length(V_1)⇒3　length(V_2)⇒5。

删除向量：vector[-n]，即删除第 n 个向量。例如：V_1[−1]⇒2,2。

纵向合并：rbind()，即向量元素都作为一行。例如：rbind(V_1,V_2)⇒$\begin{matrix}V_1\\V_2\end{matrix}$。

横向合并：cbind()，即向量元素都作为一列。例如：cbind(V_1,V_2)⇒V_1,V_2。

等差数列构成的向量：seq(from，to，by ＝间隔)。例如：seq(2、8,2)⇒2,4,6,8。

由模式构成的向量：rep(mode,time)，产生 mode 重复 time 次的向量。例如：rep(V_1,2)⇒1、2、2,1、2、2。

常用计算函数：

- mean(x)，求均值。
- sum(x)，求和。
- min(x)，求最小值。
- max(x)，求最大值。
- var(x)，求方差。
- sd(x)，求标准差。
- cov(x)，求协方差。
- cor(x)，求相关度。
- prod(x)，求所有值相乘的积。例如：prod(V_1)⇒4
- which(x 的表达式)，which. min(x)等价于 which(max(x))，which. max(x)等价于 which(min(x))。
- rev(x)，用于反转。例如：rev(V_1)⇒2,2,1。
- sort(x)，用于排序。

(2) 矩阵(Matrix)：二维的数据表，是数组的一个特例。

```
x <-1:12 ;  dim(x)<- c(3,4)
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

等价于 x<- matrix(1:12,nrow＝3,byrow＝F)，如果 byrow＝T，则列优先。

行列命名：

- colnames(matrix)＝c("","",…)。

- rownames(matrix)=c("","",…)。

矩阵运算：

- 矩阵相乘：A%*%B。
- t(matrix)，矩阵转置。
- diag(matrix)，矩阵的对角(向量)。
- solve(matrix)，矩阵求逆。
- eigen(matrix)，求矩阵的特征值和特征向量。
- svd(matrix)，奇异值分解，返回 matrix 包含属性 U、d、V。

(3) 数据框(Dataframe)：由一个或几个向量和(或)因子构成，它们必须是等长的，但可以是不同的数据类型(见图 1.4)。

样方	物种数	科数	属数	海拔	坡度	类型	⇦列名names
样方1	40	15	22	600	25	山顶	
样方2	51	12	26	350	30	山坡	
样方3	46	11	20	390	45	山坡	⇦每行作为一个Entry
样方4	38	12	24	260	20	低地	
样方5	49	10	25	220	33	低地	

⇧行名index　⇧每列可看作带名称的名量　⇧字符串、因子

图 1.4　数据框示例

数据框引用：

- df$物种数　　#引用第 2 列
- df[,1];　　#引用第 1 列
- df[5,];　　#引用第 5 行
- df[5,1];　　#引用第 1 列，第 5 行
- df[1:5,]　　#引用 1～5 行
- df[-c(1,3),]　　#引用除 1、3 行的数据。
- subset(df,物种数>40)　　#取数据的子集。

(4) 数组(Array)：数组是 k 维的数据表($k=1,2,\cdots,n$，n 为正整数)。$n=1$ 表示向量，$n=2$ 表示矩阵，$n\geqslant 3$ 表示高维数组。

(5) 列表(List)：列表可以包含任何类型的对象。例如，列表成员可以包含向量、矩阵、高维数组，也可以包含列表。

(6) 因子(Factor)：因子用水平来表示所有可能的取值。

① 创建(转换)因子：

- factor(v,level=vl)，level 如不指定，则默认取 v 中所有值。