

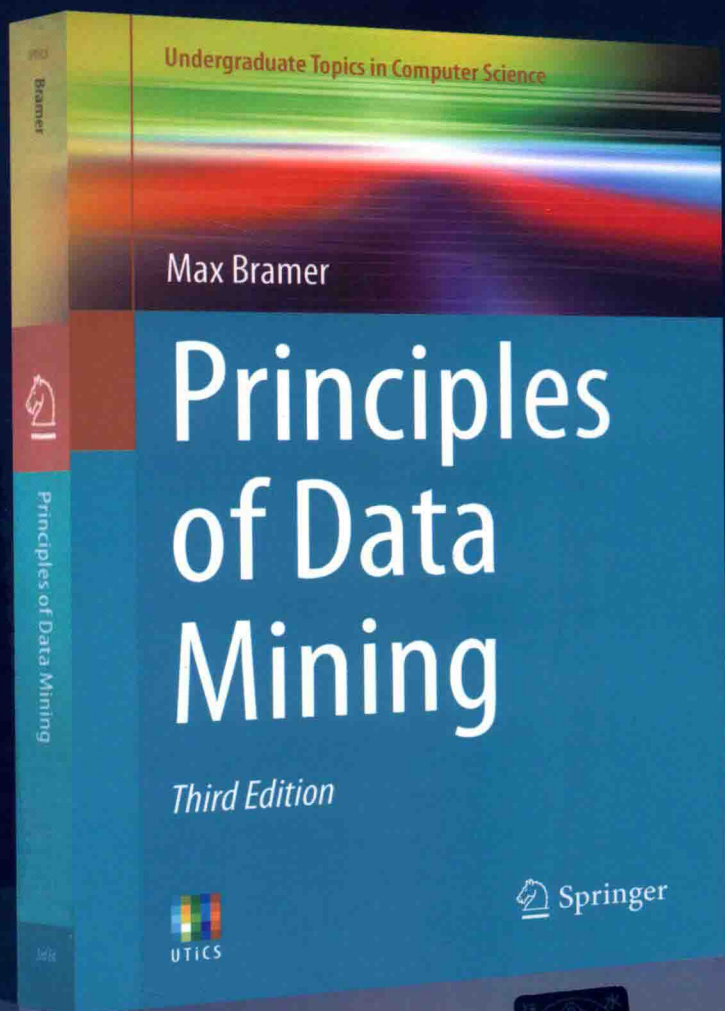
国外计算机科学经典教材

Principles of Data Mining, Third Edition

# 数据挖掘原理

(第3版)

[英] 麦克斯·布拉默(Max Bramer) 著  
王 净 译



 Springer



清华大学出版社

国外计算机科学经典教材

# 数据挖掘原理

(第3版)

[英] 麦克斯·布拉默(Max Bramer) 著

王 净

译

清华大学出版社

北 京

北京市版权局著作权合同登记号 图字：01-2018-7422

Translation from English language edition: Principles of Data Mining, Third Edition by Max Bramer Copyright © 2016, Springer-Verlag Berlin London is a part of Springer Science+Business Media.

All Rights Reserved.

本书中文简体字翻译版由德国施普林格公司授权清华大学出版社在中华人民共和国境内(不包括中国香港、澳门特别行政区和中国台湾地区)独家出版发行。未经出版者预先书面许可,不得以任何方式复制或抄袭本书的任何部分。

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

数据挖掘原理/(英)麦克斯·布拉默(Max Bramer)著;王净译.—3版.—北京:清华大学出版社,2019

(国外计算机科学经典教材)

书名原文:Principles of Data Mining, Third Edition

ISBN 978-7-302-52681-0

I. ①数… II. ①麦… ②王… III. ①数据采集—高等学校—教材 IV. ①TP274

中国版本图书馆CIP数据核字(2019)第057435号

责任编辑:王军 韩宏志

装帧设计:孔祥峰

责任校对:成凤进

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](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:三河市少明印务有限公司

经 销:全国新华书店

开 本:170mm×240mm 印 张:27.5 字 数:537千字

版 次:2019年9月第1版 印 次:2019年9月第1次印刷

定 价:79.80元

产品编号:081479-01

# 译者序

数据挖掘(Data Mining)是从不完全的、有噪声的、模糊的大量随机数据中提取出隐含的、人们事先不知道但潜在有用的信息和知识的过程。随着信息技术的高速发展,商业和科研领域积累的数据量急剧增长,动辄以 TB 计,如何从海量数据中提取有用信息成为当务之急。为顺应这种趋势,数据挖掘技术应运而生。数据挖掘是知识发现的关键。

本书共分 22 章,另有 5 个附录,系统介绍数据挖掘与应用的基本算法,探讨每种算法的基本原理,并列举大量示例加以演示,这种理论与实践相结合的方式极大地方便了读者对抽象算法的理解和掌握。第 1~3 章主要介绍数据挖掘的基本概念和分类,重点分析朴素贝叶斯算法和最近邻算法。第 4~15 章是全书的重点,浓墨重彩地描述决策树,包括属性选择的不同标准、预测分类器的精度、避免过度拟合、连续属性离散化、熵、归纳分类的模块化规则、集成分类等。第 16~18 章主要介绍关联规则挖掘的相关内容。第 19 章和第 20 章讲述聚类和文本挖掘。第 21 章和第 22 章讲述如何对流数据进行分类。为巩固学习效果,每章结尾处提供若干自我评估练习,以便读者自查,附录 E 给出练习题答案。此外,与其他许多书籍不同,在学习本书的过程中,不需要太多数学知识即可理解相关内容。附录 A 介绍相关数学内容和符号,附录 B 详细列出书中所用的数据集,附录 C 提供其他一些关于数据挖掘的资源,附录 D 列出术语表。

本书可作为本科生或硕士研究生的数据挖掘教科书,适用的学科广泛,包括计算机科学、商业研究、市场营销、人工智能、生物信息学和法医学等。对于那些希望进一步提高自身能力的技术或管理人员来说,本书也是一本优秀的自学书籍。

参与本书翻译的人员有王净、范园芳、胡训强和晏峰,最终由王净负责统稿。

译者在翻译过程中,尽量保持原书特色,并对书中出现的术语和难句进行了认真推敲和研究。但毕竟有少量技术是译者在自己的研究领域中所不曾遇到过的,所以疏漏和争议之处在所难免,望广大读者提出宝贵意见。

最后,希望广大读者能多花些时间细细品味这本凝聚了作者和译者大量心血的书籍,为自己将来的职业生涯奠定良好基础。

王净  
作于广州

首先感谢我的女儿 Bryony，她帮我绘制了许多复杂的图表并提出设计建议。其次感谢 Frederic Stahl 博士，他就第 21 章和第 22 章给出了许多宝贵建议，最后要感谢我的妻子 Dawn，她对本书草案给出了相当宝贵的意见。不过，最终版本中的任何错误仍然由我负责。

Max Bramer

# 关于第3版的说明

---

自第1版以来，可用于数据挖掘的数据量大幅增加。第1章所引用的数字现在看来相当保守。根据IBM于2016年所做的统计，每天从各种传感器、移动设备、在线交易和社交网络生成的数据量高达2.5YB，仅过去两年就创建了世界上90%的数据。现在，每天的数据流记录超过100万条。因此，第3版新增了两章，用于详细解释对流数据进行分类的算法。

# 前 言

本书面向计算机科学、商业研究、市场营销、人工智能、生物信息学和法医学专业的学生，可用作本科生或硕士研究生的入门教材。同时，对于那些希望进一步提高自身能力的技术或管理人员来说，本书也是一本极佳的自学书籍。本书所涉及的内容远超一般的数据挖掘入门书籍。与许多其他书籍不同的是，在学习过程中你不需要拥有太多的数学知识即可理解相关内容。

数学是一种可以表达复杂思想的语言。遗憾的是，99%的人都无法很好地掌握这门语言；很多人很早就开始在学校学习一些基础知识，但学习过程往往充满曲折。

本书涉及数学公式较少，将重点介绍相关概念。但遗憾的是，完全不使用数学符号是不可能的。附录 A 给出开始学习本书需要掌握的所有内容。对于那些在学校学习数学的人来说，这些内容应该是非常熟悉的。掌握这些内容后，其他内容就较好理解了。如果觉得某些数学符号难以理解，通常可放心地忽略它们，只需要关注结果和给出的详细示例即可。而对于那些希望更深入理解数据挖掘的数学基础知识的人来说，可参考附录 C 中列出的内容。

过去，没有一本关于数据挖掘的入门书可使你具备该领域的研究水平——但现在，这样的日子已经过去了。本书的重点是介绍基本技术，而不是展示当今最新的数据挖掘技术，因为大多数情况下，当拿到一本书时，书中介绍的技术可能已被其他更新的技术取代了。一旦掌握了基本技术，你可通过多种渠道来了解该领域的最新进展。附录 C 列出一些常用资源，而其他附录包括有关本书示例中使用的主要数据集的信息，供你在自己的项目中使用。此外附录 D 包括技术术语表。

为便于检查对所学知识的掌握情况，每章都包含自我评估练习。参考答案见附录 E。

另外说明一下，本书涉及大量数据集、属性和值，也涉及不少数学公式，字母繁多，格式复杂。为保证全书的科学性和严谨性，中文书中，字母的正斜体与英文原书基本保持统一。



# 目 录

第1章 数据挖掘简介	1	3.2 朴素贝叶斯分类器	18
1.1 数据爆炸	1	3.3 最近邻分类	24
1.2 知识发现	2	3.3.1 距离测量	26
1.3 数据挖掘的应用	3	3.3.2 标准化	28
1.4 标签和无标签数据	4	3.3.3 处理分类属性	29
1.5 监督学习：分类	4	3.4 急切式和懒惰式学习	30
1.6 监督学习：数值预测	5	3.5 本章小结	30
1.7 无监督学习：关联规则	6	3.6 自我评估练习	30
1.8 无监督学习：聚类	7	第4章 使用决策树进行分类	31
第2章 用于挖掘的数据	9	4.1 决策规则和决策树	31
2.1 标准制定	9	4.1.1 决策树：高尔夫示例	31
2.2 变量的类型	10	4.1.2 术语	33
2.3 数据准备	11	4.1.3 degrees 数据集	33
2.4 缺失值	13	4.2 TDIDT 算法	36
2.4.1 丢弃实例	13	4.3 推理类型	38
2.4.2 用最频繁值/平均值替换	13	4.4 本章小结	38
2.5 减少属性个数	14	4.5 自我评估练习	39
2.6 数据集的 UCI 存储库	15	第5章 决策树归纳：使用熵	
2.7 本章小结	15	进行属性选择	41
2.8 自我评估练习	15	5.1 属性选择：一个实验	41
第3章 分类简介：朴素贝叶斯和		5.2 替代决策树	42
最近邻算法	17	5.2.1 足球/无板篮球示例	42
3.1 什么是分类	17	5.2.2 匿名数据集	44

5.3 选择要分裂的属性:	7.4 方法3: $N$ -折交叉验证	70
使用熵	7.5 实验结果 I	71
5.3.1 lens24 数据集	7.6 实验结果 II: 包含缺失值	
5.3.2 熵	的数据集	73
5.3.3 使用熵进行属性选择	7.6.1 策略1: 丢弃实例	73
5.3.4 信息增益最大化	7.6.2 策略2: 用最频繁值/	
5.4 本章小结	平均值替换	74
5.5 自我评估练习	7.6.3 类别缺失	75
<b>第6章 决策树归纳: 使用频率表</b>	7.7 混淆矩阵	75
进行属性选择	7.8 本章小结	77
53	7.9 自我评估练习	77
6.1 实践中的熵计算	<b>第8章 连续属性</b>	79
6.1.1 等效性证明	8.1 简介	79
6.1.2 关于零值的说明	8.2 局部与全局离散化	81
6.2 其他属性选择标准:	8.3 向 TDIDT 添加局部	
多样性基尼指数	离散化	81
6.3 $\chi^2$ 属性选择准则	8.3.1 计算一组伪属性的	
6.4 归纳偏好	信息增益	82
6.5 使用增益比进行属性选择	8.3.2 计算效率	86
6.5.1 分裂信息的属性	8.4 使用 ChiMerge 算法进行	
6.5.2 总结	全局离散化	88
6.6 不同属性选择标准生成的	8.4.1 计算期望值和 $\chi^2$	90
规则数	8.4.2 查找阈值	94
6.7 缺失分支	8.4.3 设置 minIntervals 和	
6.8 本章小结	maxIntervals	95
6.9 自我评估练习	8.4.4 ChiMerge 算法: 总结	96
<b>第7章 估计分类器的预测精度</b>	8.4.5 对 ChiMerge 算法的	
7.1 简介	评述	96
7.2 方法1: 将数据划分为	8.5 比较树归纳法的全局离	
训练集和测试集	散化和局部离散化	97
7.2.1 标准误差	8.6 本章小结	98
7.2.2 重复训练和测试	8.7 自我评估练习	98
7.3 方法2: $k$ -折交叉验证		

<b>第 9 章 避免决策树的过度拟合</b> ..... 99	
9.1 处理训练集中的冲突..... 99	
9.2 关于过度拟合数据的更多规则..... 103	
9.3 预剪枝决策树..... 104	
9.4 后剪枝决策树..... 106	
9.5 本章小结..... 111	
9.6 自我评估练习..... 111	
<b>第 10 章 关于熵的更多信息</b> ..... 113	
10.1 简介..... 113	
10.2 使用位的编码信息..... 116	
10.3 区分值..... 117	
10.4 对“非等可能”的值进行编码..... 118	
10.5 训练集的熵..... 121	
10.6 信息增益必须为正数或零..... 122	
10.7 使用信息增益来简化分类任务的特征..... 123	
10.7.1 示例 1: genetics 数据集..... 124	
10.7.2 示例 2: bcst96 数据集..... 126	
10.8 本章小结..... 128	
10.9 自我评估练习..... 128	
<b>第 11 章 归纳分类的模块化规则</b> ..... 129	
11.1 规则后剪枝..... 129	
11.2 冲突解决..... 130	
11.3 决策树的问题..... 133	
11.4 Prism 算法..... 135	
11.4.1 基本 Prism 算法的变化..... 141	
11.4.2 将 Prism 算法与 TDIDT 算法进行比较..... 142	
11.5 本章小结..... 143	
11.6 自我评估练习..... 143	
<b>第 12 章 度量分类器的性能</b> ..... 145	
12.1 真假正例和真假负例..... 146	
12.2 性能度量..... 147	
12.3 真假正例率与预测精度..... 150	
12.4 ROC 图..... 151	
12.5 ROC 曲线..... 153	
12.6 寻找最佳分类器..... 153	
12.7 本章小结..... 155	
12.8 自我评估练习..... 155	
<b>第 13 章 处理大量数据</b> ..... 157	
13.1 简介..... 157	
13.2 将数据分发到多个处理器..... 159	
13.3 案例研究: PMCRI..... 161	
13.4 评估分布式系统 PMCRI 的有效性..... 163	
13.5 逐步修改分类器..... 167	
13.6 本章小结..... 171	
13.7 自我评估练习..... 171	
<b>第 14 章 集成分类</b> ..... 173	
14.1 简介..... 173	
14.2 估计分类器的性能..... 175	
14.3 为每个分类器选择不同的训练集..... 176	
14.4 为每个分类器选择一组不同的属性..... 177	

14.5	组合分类: 替代投票系统	177	17.2	事务和项目集	209
14.6	并行集成分类器	180	17.3	对项目集的支持	211
14.7	本章小结	181	17.4	关联规则	211
14.8	自我评估练习	181	17.5	生成关联规则	213
<b>第 15 章</b>	<b>比较分类器</b>	<b>183</b>	17.6	Apriori	214
15.1	简介	183	17.7	生成支持项目集: 一个示例	217
15.2	配对 $t$ 检验	184	17.8	为支持项目集生成规则	219
15.3	为比较评估选择数据集	189	17.9	规则兴趣度度量: 提升度和杠杆率	220
15.4	抽样	191	17.10	本章小结	222
15.5	“无显著差异”的结果 有多糟糕?	193	17.11	自我评估练习	222
15.6	本章小结	194	<b>第 18 章</b>	<b>关联规则挖掘 III:</b> 频繁模式树	225
15.7	自我评估练习	194	18.1	简介: FP-growth	225
<b>第 16 章</b>	<b>关联规则挖掘 I</b>	<b>195</b>	18.2	构造 FP-tree	227
16.1	简介	195	18.2.1	预处理事务数据库	227
16.2	规则兴趣度的衡量标准	196	18.2.2	初始化	229
16.2.1	Piatetsky-Shapiro 标准 和 RI 度量	198	18.2.3	处理事务 1: $f, c, a,$ $m, p$	230
16.2.2	规则兴趣度度量应用 于 chess 数据集	200	18.2.4	处理事务 2: $f, c, a,$ $b, m$	231
16.2.3	使用规则兴趣度度量 来解决冲突	201	18.2.5	处理事务 3: $f, b$	235
16.3	关联规则挖掘任务	202	18.2.6	处理事务 4: $c, b, p$	236
16.4	找到最佳 $N$ 条规则	202	18.2.7	处理事务 5: $f, c, a,$ $m, p$	236
16.4.1	$J$ -Measure: 度量 规则的信息内容	203	18.3	从 FP-tree 中查找频繁 项目集	238
16.4.2	搜索策略	204	18.3.1	以项目 $p$ 结尾的 项目集	240
16.5	本章小结	207	18.3.2	以项目 $m$ 结尾的 项目集	248
16.6	自我评估练习	207	18.4	本章小结	254
<b>第 17 章</b>	<b>关联规则挖掘 II</b>	<b>209</b>			
17.1	简介	209			

18.5	自我评估练习	254	21.2	构建H-Tree: 更新数组	283
<b>第 19 章</b>	<b>聚类</b>	<b>255</b>	21.2.1	currentAtts 数组	284
19.1	简介	255	21.2.2	splitAtt 数组	284
19.2	<i>k</i> -means 聚类	257	21.2.3	将记录排序到适当的 叶节点	284
19.2.1	示例	258	21.2.4	hitcount 数组	285
19.2.2	找到最佳簇集	262	21.2.5	classtotals 数组	285
19.3	凝聚式层次聚类	263	21.2.6	acvCounts 阵列	285
19.3.1	记录簇间距离	265	21.2.7	branch 数组	286
19.3.2	终止聚类过程	268	<b>21.3</b>	<b>构建H-Tree: 详细示例</b>	<b>287</b>
19.4	本章小结	268	21.3.1	步骤 1: 初始化 根节点 0	287
19.5	自我评估练习	268	21.3.2	步骤 2: 开始读取 记录	287
<b>第 20 章</b>	<b>文本挖掘</b>	<b>269</b>	21.3.3	步骤 3: 考虑在节 点 0 处分裂	288
20.1	多重分类	269	21.3.4	步骤 4: 在根节点上 拆分并初始化新的 叶节点	289
20.2	表示数据挖掘的文本 文档	270	21.3.5	步骤 5: 处理下一组 记录	290
20.3	停用词和词干	271	21.3.6	步骤 6: 考虑在节点 2 处分裂	292
20.4	使用信息增益来减少 特征	272	21.3.7	步骤 7: 处理下一组 记录	292
20.5	表示文本文档: 构建 向量空间模型	272	21.3.8	H-Tree 算法概述	293
20.6	规范权重	273	<b>21.4</b>	<b>分裂属性: 使用信息 增益</b>	<b>295</b>
20.7	测量两个向量之间的 距离	274	<b>21.5</b>	<b>分裂属性: 使用 Hoeffding 边界</b>	<b>297</b>
20.8	度量文本分类器的性能	275	<b>21.6</b>	<b>H-Tree 算法: 最终版本</b>	<b>300</b>
20.9	超文本分类	275	<b>21.7</b>	<b>使用不断进化的 H-Tree 进行预测</b>	<b>302</b>
20.9.1	对网页进行分类	276			
20.9.2	超文本分类与文本 分类	277			
20.10	本章小结	279			
20.11	自我评估练习	280			
<b>第 21 章</b>	<b>分类流数据</b>	<b>281</b>			
21.1	简介	281			

21.8	实验: H-Tree与TDIDT	304	22.8	创建备用节点	322
21.8.1	lens24 数据集	304	22.9	成长/遗忘备用节点及其后代	325
21.8.2	vote 数据集	306	22.10	用备用节点替换一个内部节点	327
21.9	本章小结	307	22.11	实验: 跟踪概念漂移	333
21.10	自我评估练习	307	22.11.1	lens24 数据: 替代模式	335
<b>第 22 章</b>	<b>分类流数据 II: 时间相关数据</b>	<b>309</b>	22.11.2	引入概念漂移	335
22.1	平稳数据与时间相关数据	309	22.11.3	使用交替 lens24 数据的实验	336
22.2	H-Tree 算法总结	311	22.11.4	关于实验的评论	343
22.2.1	currentAtts 数组	312	22.12	本章小结	343
22.2.2	splitAtt 数组	312	22.13	自我评估练习	343
22.2.3	hitcount 数组	312	附录 A	基本数学知识	345
22.2.4	classtotals 数组	312	附录 B	数据集	357
22.2.5	acvCounts 数组	313	附录 C	更多信息来源	371
22.2.6	branch 数组	313	附录 D	词汇表和符号	373
22.2.7	H-Tree 算法的伪代码	313	附录 E	自我评估练习题答案	391
22.3	从 H-Tree 到 CDH-Tree: 概述	315	参考文献	419	
22.4	从 H-Tree 转换到 CDH-Tree: 递增计数	315			
22.5	滑动窗口法	316			
22.6	在节点处重新分裂	320			
22.7	识别可疑节点	320			

# 第 1 章

## 数据挖掘简介

### 1.1 数据爆炸

现代计算机系统以令人难以置信的速度从各种来源收集数据。从街头的 POS 机到用于支票结算、现金提取和信用卡交易的机器，再到太空中的地球观测卫星，都在不断收集大量信息。

下面列举一些数据量：

- 目前美国宇航局的地球观测卫星每天生成 1TB 数据。
- 人类基因组计划针对数十亿个遗传碱基中的每一个存储数千个字节。
- 许多公司都维护着大型客户交易数据仓库。一个较小数据仓库可能包含超过一亿个事务。
- 每天在自动记录设备上记录大量数据(如信用卡交易文件和网络日志，以及监控系统记录的非符号数据)。
- 估计有超过 6.5 亿个网站，其中一些网站非常庞大。
- Facebook 拥有超过 9 亿用户，每天估计产生 30 亿个帖子。
- 据估计，Twitter 用户约有 1.5 亿，每天发送 3.5 亿条推文。

随着存储技术的不断发展，无论是商业数据仓库、科研实验室还是其他地方，都越来越多地以较低成本存储大量数据，同时人们逐渐认识到这些数据中可能包含以下知识：对公司的兴衰至关重要的知识，可导致重大科学发现的知识，可更准确地预测

天气和自然灾害的知识,可找出致命疾病原因及治愈方法的知识,以及可能关乎生死的知识。然而,这些数据大部分仅被存储——人们很少对这些数据进行更深入的检查。所以准确地说,世界正变得“数据丰富但知识贫乏”。

机器学习技术有可能解决一直困扰公司、政府和个人的数据爆炸问题。

## 1.2 知识发现

“知识发现”(Knowledge Discovery)被定义为“从数据中提取隐含的、先前未知的和潜在可用的信息”。该过程虽然只是数据挖掘的一部分,却是核心部分。

图 1.1 显示了完整的知识发现过程的略微理想化的版本。

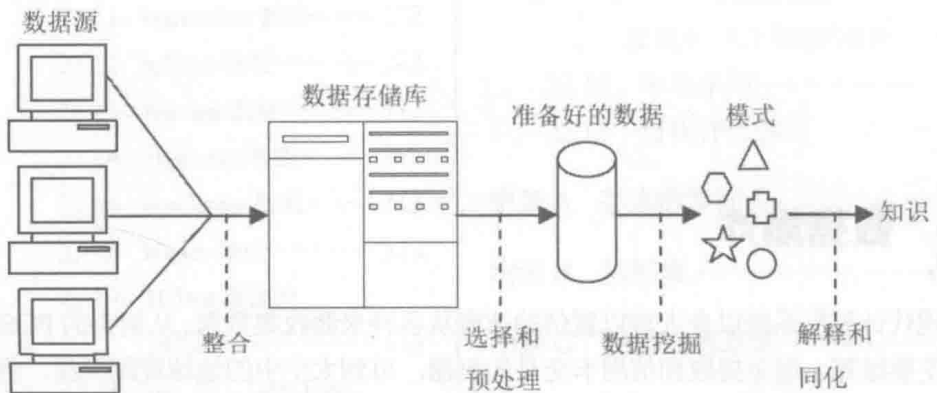


图 1.1 知识发现过程

数据可能有许多来源,被集成并放在一些通用数据存储中。然后将其中一部分数据预处理成标准格式,再将这些“准备好的数据”传递给数据挖掘算法,该算法根据规则或某种其他类型的“模式”生产输出。最后对这些输出进行解释并给出潜在有用的新知识——这就是知识发现的“圣杯”。

通过上面的简述可清楚地看到,数据挖掘算法是知识发现的核心,但并非全部。数据的预处理和结果的解释也非常重要。与其说完成这些任务是一门精确的科学,还不如说是一门艺术。虽然本书也会介绍数据的预处理并解释结果,但重点讨论的是知识发现的数据挖掘阶段涉及的算法。