

# 《数据挖掘：实用机器学习工具与技术（》

## 图书基本信息

书名：《数据挖掘：实用机器学习工具与技术（原书第3版）》

13位ISBN编号：9787111453816

出版时间：2014-5-1

作者：Ian H.Witten,Eibe Frank

页数：480

译者：李川,张永辉

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：[www.tushu000.com](http://www.tushu000.com)

# 《数据挖掘：实用机器学习工具与技术（》

## 内容概要

大数据时代应用机器学习方法解决数据挖掘问题的实用指南。

洞察隐匿于大数据中的结构模式，有效指导数据挖掘实践和商业应用。

weka系统的主要开发者将丰富的研发、商业应用和教学实践的经验和技术融会贯通。

广泛覆盖在数据挖掘实践中采用的算法和机器学习技术，着眼于解决实际问题

避免过分要求理论基础和数学知识，重点在于告诉读者“如何去做”，同时包括许多算法、代码以及具体实例的实现。

将所有的概念都建立在具体实例的基础之上，促使读者首先考虑使用简单的技术。如果简单的技术不足以解决问题，再考虑提升到更为复杂的高级技术。

新版增加了大量近年来最新涌现的数据挖掘算法和诸如Web数据挖掘等新领域的介绍，所介绍的weka系统增加了50%的算法及大量新内容。

本书是机器学习和数据挖掘领域的经典畅销教材，被众多国外名校选为教材。书中详细介绍用于数据挖掘领域的机器学习技术和工具以及实践方法，并且提供了一个公开的数据挖掘工作平台Weka。本书主要内容包括：数据输入/输出、知识表示、数据挖掘技术（决策树、关联规则、基于实例的学习、线性模型、聚类、多实例学习等）以及在实际中的运用。本版对上一版内容进行了全面更新，以反映自第2版出版以来数据挖掘领域的技术变革和新方法，包括数据转换、集成学习、大规模数据集、多实例学习等，以及新版的Weka机器学习软件。

# 《数据挖掘：实用机器学习工具与技术（》

## 作者简介

Ian H.Witten 新西兰怀卡托大学计算机科学系教授，ACM Fellow和新西兰皇家学会Fellow，曾荣获2004年国际信息处理研究协会（IFIP）颁发的Namur奖项。他的研究兴趣包括语言学习、信息检索和机器学习。

Eibe Frank 新西兰怀卡托大学计算机科学系副教授，《Machine Learning Journal》和《Journal of Artificial Intelligence Research》编委。

Mark A.Hall 新西兰怀卡托大学名誉副研究员，曾获得2005年ACM SIGKDD服务奖。

## 书籍目录

### 目 录

Data Mining : Practical Machine Learning Tools and Techniques , Third Edition

出版者的话

译者序

前言

致谢

### 第一部分 数据挖掘简介

#### 第1章 绪论2

##### 1.1 数据挖掘和机器学习2

###### 1.1.1 描述结构模式3

###### 1.1.2 机器学习5

###### 1.1.3 数据挖掘6

##### 1.2 简单的例子：天气问题和其他问题6

###### 1.2.1 天气问题7

###### 1.2.2 隐形眼镜：一个理想化的问题8

###### 1.2.3 鸢尾花：一个经典的数值型数据集10

###### 1.2.4 CPU性能：介绍数值预测11

###### 1.2.5 劳资协商：一个更真实的例子11

###### 1.2.6 大豆分类：一个经典的机器学习的成功例子13

##### 1.3 应用领域14

###### 1.3.1 Web挖掘15

###### 1.3.2 包含评判的决策15

###### 1.3.3 图像筛选16

###### 1.3.4 负载预测17

###### 1.3.5 诊断17

###### 1.3.6 市场和销售18

###### 1.3.7 其他应用19

##### 1.4 机器学习和统计学20

##### 1.5 将泛化看做搜索21

###### 1.5.1 枚举概念空间22

###### 1.5.2 偏差22

##### 1.6 数据挖掘和道德24

###### 1.6.1 再识别25

###### 1.6.2 使用个人信息25

###### 1.6.3 其他问题26

##### 1.7 补充读物27

### 第2章 输入：概念、实例和属性29

#### 2.1 概念29

#### 2.2 样本31

##### 2.2.1 关系32

##### 2.2.2 其他实例类型34

#### 2.3 属性35

#### 2.4 输入准备37

##### 2.4.1 数据收集37

##### 2.4.2 ARFF格式38

##### 2.4.3 稀疏数据40

##### 2.4.4 属性类型40

- 2.4.5 缺失值41
- 2.4.6 不正确的值42
- 2.4.7 了解数据43
- 2.5 补充读物43
- 第3章 输出：知识表达44
  - 3.1 表44
  - 3.2 线性模型44
  - 3.3 树45
  - 3.4 规则48
    - 3.4.1 分类规则49
    - 3.4.2 关联规则52
    - 3.4.3 包含例外的规则52
    - 3.4.4 表达能力更强的规则54
  - 3.5 基于实例的表达56
  - 3.6 聚类58
  - 3.7 补充读物60
- 第4章 算法：基本方法61
  - 4.1 推断基本规则61
    - 4.1.1 缺失值和数值属性62
    - 4.1.2 讨论64
  - 4.2 统计建模64
    - 4.2.1 缺失值和数值属性67
    - 4.2.2 用于文档分类的朴素贝叶斯68
    - 4.2.3 讨论70
  - 4.3 分治法：建立决策树70
    - 4.3.1 计算信息量73
    - 4.3.2 高度分支属性74
    - 4.3.3 讨论75
  - 4.4 覆盖算法：建立规则76
    - 4.4.1 规则与树77
    - 4.4.2 一个简单的覆盖算法77
    - 4.4.3 规则与决策列表80
  - 4.5 挖掘关联规则81
    - 4.5.1 项集81
    - 4.5.2 关联规则83
    - 4.5.3 有效地生成规则85
    - 4.5.4 讨论87
  - 4.6 线性模型87
    - 4.6.1 数值预测：线性回归87
    - 4.6.2 线性分类：Logistic回归88
    - 4.6.3 使用感知机的线性分类90
    - 4.6.4 使用Winnow的线性分类91
  - 4.7 基于实例的学习92
    - 4.7.1 距离函数93
    - 4.7.2 有效寻找最近邻93
    - 4.7.3 讨论97
  - 4.8 聚类97
    - 4.8.1 基于距离的迭代聚类98
    - 4.8.2 快速距离计算99

4.8.3	讨论	100
4.9	多实例学习	100
4.9.1	聚集输入	100
4.9.2	聚集输出	100
4.9.3	讨论	101
4.10	补充读物	101
4.11	Weka实现	103
第5章	可信度：评估学习结果	104
5.1	训练和测试	104
5.2	预测性能	106
5.3	交叉验证	108
5.4	其他评估方法	109
5.4.1	留一交叉验证	109
5.4.2	自助法	109
5.5	数据挖掘方法比较	110
5.6	预测概率	113
5.6.1	二次损失函数	114
5.6.2	信息损失函数	115
5.6.3	讨论	115
5.7	计算成本	116
5.7.1	成本敏感分类	117
5.7.2	成本敏感学习	118
5.7.3	提升图	119
5.7.4	ROC曲线	122
5.7.5	召回率-精确率曲线	124
5.7.6	讨论	124
5.7.7	成本曲线	125
5.8	评估数值预测	127
5.9	最小描述长度原理	129
5.10	在聚类方法中应用MDL原理	131
5.11	补充读物	132
第二部分	高级数据挖掘	
第6章	实现：真正的机器学习方案	134
6.1	决策树	135
6.1.1	数值属性	135
6.1.2	缺失值	136
6.1.3	剪枝	137
6.1.4	估计误差率	138
6.1.5	决策树归纳的复杂度	140
6.1.6	从决策树到规则	140
6.1.7	C4.5:选择和选项	141
6.1.8	成本-复杂度剪枝	141
6.1.9	讨论	142
6.2	分类规则	142
6.2.1	选择测试的标准	143
6.2.2	缺失值和数值属性	143
6.2.3	生成好的规则	144
6.2.4	使用全局优化	146
6.2.5	从局部决策树中获得规则	146

- 6.2.6 包含例外的规则149
- 6.2.7 讨论151
- 6.3 关联规则152
  - 6.3.1 建立频繁模式树152
  - 6.3.2 寻找大项集157
  - 6.3.3 讨论157
- 6.4 扩展线性模型158
  - 6.4.1 最大间隔超平面159
  - 6.4.2 非线性类边界160
  - 6.4.3 支持向量回归161
  - 6.4.4 核岭回归163
  - 6.4.5 核感知机164
  - 6.4.6 多层感知机165
  - 6.4.7 径向基函数网络171
  - 6.4.8 随机梯度下降172
  - 6.4.9 讨论173
- 6.5 基于实例的学习174
  - 6.5.1 减少样本集的数量174
  - 6.5.2 对噪声样本集剪枝174
  - 6.5.3 属性加权175
  - 6.5.4 泛化样本集176
  - 6.5.5 用于泛化样本集的距离函数176
  - 6.5.6 泛化的距离函数177
  - 6.5.7 讨论178
- 6.6 局部线性模型用于数值预测178
  - 6.6.1 模型树179
  - 6.6.2 构建树179
  - 6.6.3 对树剪枝180
  - 6.6.4 名目属性180
  - 6.6.5 缺失值181
  - 6.6.6 模型树归纳的伪代码181
  - 6.6.7 从模型树到规则184
  - 6.6.8 局部加权线性回归184
  - 6.6.9 讨论185
- 6.7 贝叶斯网络186
  - 6.7.1 预测186
  - 6.7.2 学习贝叶斯网络189
  - 6.7.3 算法细节190
  - 6.7.4 用于快速学习的数据结构192
  - 6.7.5 讨论194
- 6.8 聚类194
  - 6.8.1 选择聚类的个数195
  - 6.8.2 层次聚类195
  - 6.8.3 层次聚类的例子196
  - 6.8.4 增量聚类199
  - 6.8.5 分类效用203
  - 6.8.6 基于概率的聚类204
  - 6.8.7 EM算法205
  - 6.8.8 扩展混合模型206

- 6.8.9 贝叶斯聚类207
- 6.8.10 讨论209
- 6.9 半监督学习210
  - 6.9.1 用于分类的聚类210
  - 6.9.2 协同训练212
  - 6.9.3 EM和协同训练212
  - 6.9.4 讨论213
- 6.10 多实例学习213
  - 6.10.1 转换为单实例学习213
  - 6.10.2 升级学习算法215
  - 6.10.3 专用多实例方法215
  - 6.10.4 讨论216
- 6.11 Weka实现216
- 第7章 数据转换218
  - 7.1 属性选择219
    - 7.1.1 独立于方案的选择220
    - 7.1.2 搜索属性空间222
    - 7.1.3 具体方案相关的选择223
  - 7.2 离散化数值属性225
    - 7.2.1 无监督离散化226
    - 7.2.2 基于熵的离散化226
    - 7.2.3 其他离散化方法229
    - 7.2.4 基于熵的离散化与基于误差的离散化229
    - 7.2.5 离散属性转换成数值属性230
  - 7.3 投影230
    - 7.3.1 主成分分析231
    - 7.3.2 随机投影233
    - 7.3.3 偏最小二乘回归233
    - 7.3.4 从文本到属性向量235
    - 7.3.5 时间序列236
  - 7.4 抽样236
  - 7.5 数据清洗237
    - 7.5.1 改进决策树237
    - 7.5.2 稳健回归238
    - 7.5.3 检测异常239
    - 7.5.4 一分类学习239
  - 7.6 多分类问题转换成二分类问题242
    - 7.6.1 简单方法242
    - 7.6.2 误差校正输出编码243
    - 7.6.3 集成嵌套二分法244
  - 7.7 校准类概率246
  - 7.8 补充读物247
  - 7.9 Weka实现249
- 第8章 集成学习250
  - 8.1 组合多种模型250
  - 8.2 装袋251
    - 8.2.1 偏差-方差分解251
    - 8.2.2 考虑成本的装袋253
  - 8.3 随机化253



- 8.3.1 随机化与装袋254
- 8.3.2 旋转森林254
- 8.4 提升255
  - 8.4.1 AdaBoost算法255
  - 8.4.2 提升算法的威力257
- 8.5 累加回归258
  - 8.5.1 数值预测258
  - 8.5.2 累加Logistic回归259
- 8.6 可解释的集成器260
  - 8.6.1 选择树260
  - 8.6.2 Logistic模型树262
- 8.7 堆栈262
- 8.8 补充读物264
- 8.9 Weka实现265
- 第9章 继续：扩展和应用266
  - 9.1 应用数据挖掘266
  - 9.2 从大型的数据集里学习268
  - 9.3 数据流学习270
  - 9.4 融合领域知识272
  - 9.5 文本挖掘273
  - 9.6 Web挖掘276
  - 9.7 对抗情形278
  - 9.8 无处不在的数据挖掘280
  - 9.9 补充读物281
- 第三部分 Weka数据挖掘平台
- 第10章 Weka简介284
  - 10.1 Weka中包含了什么284
  - 10.2 如何使用Weka285
  - 10.3 Weka的其他应用286
  - 10.4 如何得到Weka286
- 第11章 Explorer界面287
  - 11.1 开始287
    - 11.1.1 准备数据287
    - 11.1.2 将数据载入Explorer288
    - 11.1.3 建立决策树289
    - 11.1.4 查看结果290
    - 11.1.5 重做一遍292
    - 11.1.6 运用模型292
    - 11.1.7 运行错误的处理294
  - 11.2 探索Explorer294
    - 11.2.1 载入及过滤文件294
    - 11.2.2 训练和测试学习方案299
    - 11.2.3 自己动手：用户分类器301
    - 11.2.4 使用元学习器304
    - 11.2.5 聚类和关联规则305
    - 11.2.6 属性选择306
    - 11.2.7 可视化306
  - 11.3 过滤算法307
    - 11.3.1 无监督属性过滤器307

- 11.3.2 无监督实例过滤器312
- 11.3.3 有监督过滤器314
- 11.4 学习算法316
  - 11.4.1 贝叶斯分类器317
  - 11.4.2 树320
  - 11.4.3 规则322
  - 11.4.4 函数325
  - 11.4.5 神经网络331
  - 11.4.6 懒惰分类器334
  - 11.4.7 多实例分类器335
  - 11.4.8 杂项分类器336
- 11.5 元学习算法336
  - 11.5.1 装袋和随机化337
  - 11.5.2 提升338
  - 11.5.3 组合分类器338
  - 11.5.4 成本敏感学习339
  - 11.5.5 优化性能339
  - 11.5.6 针对不同任务重新调整分类器340
- 11.6 聚类算法340
- 11.7 关联规则学习器345
- 11.8 属性选择346
  - 11.8.1 属性子集评估器347
  - 11.8.2 单一属性评估器347
  - 11.8.3 搜索方法348
- 第12章 Knowledge Flow界面351
  - 12.1 开始351
  - 12.2 Knowledge Flow组件353
  - 12.3 配置及连接组件354
  - 12.4 增量学习356
- 第13章 Experimenter界面358
  - 13.1 开始358
    - 13.1.1 运行一个实验358
    - 13.1.2 分析结果359
  - 13.2 简单设置362
  - 13.3 高级设置363
  - 13.4 分析面板365
  - 13.5 将运行负荷分布到多个机器上366
- 第14章 命令行界面368
  - 14.1 开始368
  - 14.2 Weka的结构368
    - 14.2.1 类、实例和包368
    - 14.2.2 weka.core包370
    - 14.2.3 weka.classifiers包371
    - 14.2.4 其他包372
    - 14.2.5 Javadoc索引373
  - 14.3 命令行选项373
    - 14.3.1 通用选项374
    - 14.3.2 与具体方案相关的选项375
- 第15章 嵌入式机器学习376

15.1	一个简单的数据挖掘应用	376
15.1.1	MessageClassifier ( )	380
15.1.2	updateData ( )	380
15.1.3	classifyMessage ( )	381
第16章	编写新的学习方案	382
16.1	一个分类器范例	382
16.1.1	buildClassifier ( )	389
16.1.2	makeTree ( )	389
16.1.3	computeInfoGain ( )	390
16.1.4	classifyInstance ( )	390
16.1.5	toSource ( )	391
16.1.6	main ( )	394
16.2	与实现分类器有关的惯例	395
第17章	Weka Explorer的辅导练习	397
17.1	Explorer界面简介	397
17.1.1	导入数据集	397
17.1.2	数据集编辑器	397
17.1.3	应用过滤器	398
17.1.4	可视化面板	399
17.1.5	分类器面板	399
17.2	最近邻学习和决策树	402
17.2.1	玻璃数据集	402
17.2.2	属性选择	403
17.2.3	类噪声以及最近邻学习	403
17.2.4	改变训练数据的数量	404
17.2.5	交互式建立决策树	405
17.3	分类边界	406
17.3.1	可视化1R	406
17.3.2	可视化最近邻学习	407
17.3.3	可视化朴素贝叶斯	407
17.3.4	可视化决策树和规则集	407
17.3.5	弄乱数据	408
17.4	预处理以及参数调整	408
17.4.1	离散化	408
17.4.2	离散化的更多方面	408
17.4.3	自动属性选择	409
17.4.4	自动属性选择的更多方面	410
17.4.5	自动参数调整	410
17.5	文档分类	411
17.5.1	包含字符串属性的数据	411
17.5.2	实际文档文类	412
17.5.3	探索StringToWordVector过滤器	413
17.6	挖掘关联规则	413
17.6.1	关联规则挖掘	413
17.6.2	挖掘一个真实的数据集	415
17.6.3	购物篮分析	415
	参考文献	416
	索引	431



# 《数据挖掘：实用机器学习工具与技术（》

## 版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:[www.tushu000.com](http://www.tushu000.com)