

《数据挖掘技术与工程实践》

图书基本信息

书名：《数据挖掘技术与工程实践》

13位ISBN编号：9787111480767

出版时间：2014-10-1

作者：洪松林,庄映辉,李堃

页数：400

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.tushu111.com

《数据挖掘技术与工程实践》

内容概要

数据挖掘是当前最活跃的领域之一。本书作者根据自己20年数据挖掘方面的经验，总结了数据挖掘的理论和实践经验，提供了大量一线资料。本书首先介绍数据挖掘的概念和误区，然后介绍数据探索的方法，包括数据查探、数据描绘、数据变换、数据优化等，重点介绍了相关算法，包括：相关因子算法、聚类算法、分类算法、回归与测试算法等。不仅列举了详细示例，还介绍了算法在工程实践中的具体应用，特别是总结了自己独特的一些新算法，例如秩相关因子选择算法、矢量相关因子选择算法、密度分布聚类算法、概率特征模型算法等。还剖析了几个热门领域的实际应用，涉及医药学、信息安全、新闻分析、商品推荐、证券预测等领域的应用。最后归纳总结了数据挖掘应用系统的开发方案，并介绍一个数据挖掘工具的应用。本书可供数据挖掘、数据仓库、数据库等领域的技术人员参考，也可供想建立智能计算系统的企业信息系统管理人员参考。

《数据挖掘技术与工程实践》

作者简介

Hong Song Lin (洪松林) 福安易数据技术(天津)有限公司(F&E DATA TECHNOLOGY CORP.) 创始人, 外国专家局引智技术专家, 加拿大OCP认证专家, 有20年智能计算(数据仓库、商务智能及数据挖掘)方面的研究、设计、开发和培训经验。掌握北美先进的项目经验, 曾在加拿大安大略省卫生部(OMH)、蒙特利尔银行(BMO)、加拿大研科电讯公司(TELUS)、安省高教委(OCAS)等大型企业参与多个大型智能计算项目。近年来在国内主持多个智能计算产品的总体设计和研发工作, 将北美的智能计算技术及业务经验与中国的专业需求和数据环境有效地结合起来, 开发了以数据仓库、数据挖掘和数据统计为技术核心的智能数据分析产品, 国内首创, 并在北京、天津等地得到成功应用。

书籍目录

前 言

第1章 数据挖掘应用绪论1

1.1 认识数据挖掘1

1.1.1 数据挖掘概念2

1.1.2 数据挖掘与生活4

1.1.3 数据挖掘与知识6

1.2 数据挖掘应用基础6

1.2.1 事物与维度7

1.2.2 分布与关系9

1.2.3 描绘与预测11

1.2.4 现象和知识13

1.2.5 规律与因果13

1.3 数据挖掘应用系统工程14

1.3.1 数据层14

1.3.2 算法层18

1.3.3 应用层23

1.4 数据挖掘应用体会26

1.4.1 项目关键点26

1.4.2 技术与应用创新27

1.4.3 经验积累与应用28

1.5 无限三维嵌套空间假说28

1.5.1 一维空间29

1.5.2 二维空间29

1.5.3 三维空间29

1.5.4 突破三维空间30

1.5.5 五维空间31

1.5.6 六维空间31

1.6 本章小结32

第2章 数据探索与准备33

2.1 数据关系探索34

2.1.1 业务发现34

2.1.2 关系发现36

2.1.3 数据质量探索37

2.1.4 数据整合40

2.2 数据特征探索42

2.2.1 数据的统计学特征42

2.2.2 统计学特征应用48

2.3 数据选择52

2.3.1 适当的数据规模52

2.3.2 数据的代表性53

2.3.3 数据的选取54

2.4 数据处理56

2.4.1 数据标准化57

2.4.2 数据离散化58

2.5 统计学算法的数量条件60

2.5.1 样本量估计概念60

2.5.2 单样本总体均值比较的样本量估计 (T-Test) 61

- 2.5.3 两样本总体均值比较的样本量估计(T-Test)62
- 2.5.4 多样本总体均值比较的样本量估计(F-Test)63
- 2.5.5 区组设计多样本总体均值比较的样本量估计 (F-Test) 66
- 2.5.6 直线回归与相关的样本量估计66
- 2.5.7 对照分析的样本量估计67
- 2.6 数据探索应用68
 - 2.6.1 检验项的疾病分布69
 - 2.6.2 疾病中检验项的分布70
 - 2.6.3 成对检验项的相关分析71
 - 2.6.4 两种药物的应用分析71
- 2.7 本章小结73
- 第3章 数据挖掘应用算法74
 - 3.1 聚类分析74
 - 3.1.1 划分聚类算法 (K均值) 75
 - 3.1.2 层次聚类算法 (组平均) 79
 - 3.1.3 密度聚类算法84
 - 3.2 特性选择85
 - 3.2.1 特性选择概念85
 - 3.2.2 线性相关算法90
 - 3.2.3 相关因子SRCF算法91
 - 3.3 特征抽取100
 - 3.3.1 主成分分析算法101
 - 3.3.2 因子分析算法102
 - 3.3.3 非负矩阵因子分解NMF算法103
 - 3.4 关联规则104
 - 3.4.1 关联规则概念105
 - 3.4.2 Apriori算法105
 - 3.4.3 FP树频集算法106
 - 3.4.4 提升Lift107
 - 3.5 分类和预测107
 - 3.5.1 支持向量机107
 - 3.5.2 Logistic回归算法112
 - 3.5.3 朴素贝叶斯分类算法115
 - 3.5.4 决策树121
 - 3.5.5 人工神经网络125
 - 3.5.6 分类与聚类的关系129
 - 3.6 时间序列129
 - 3.6.1 灰色系统预测模型129
 - 3.6.2 ARIMA模型预测135
 - 3.7 本章小结136
- 第4章 数据挖掘应用案例137
 - 4.1 特性选择的应用137
 - 4.1.1 数据整合137
 - 4.1.2 数据描绘138
 - 4.1.3 数据标准化139
 - 4.1.4 特性选择探索139
 - 4.2 分类模型的应用——算法比较144
 - 4.2.1 数据整合144
 - 4.2.2 数据描绘145

- 4.2.3 数据标准化148
- 4.2.4 特性选择探索148
- 4.2.5 分类模型150
- 4.3 分类模型的应用——网络异常侦测151
 - 4.3.1 计算机网络异常行为152
 - 4.3.2 网络异常数据模型152
 - 4.3.3 分类模型算法应用156
- 4.4 算法的综合应用——肿瘤标志物的研究159
 - 4.4.1 样本选取160
 - 4.4.2 癌胚抗原临床特征主题分析164
 - 4.4.3 癌胚抗原临床特征规则分析167
 - 4.4.4 癌胚抗原临床特征规则的比较分析172
 - 4.4.5 癌胚抗原相关因子分析173
 - 4.4.6 不同等级癌胚抗原组差异分析176
- 4.5 数据挖掘在其他领域中的应用180
- 4.6 本章小结182
- 第5章 数据挖掘行业应用原理183
 - 5.1 传统医学科研方法的现状184
 - 5.1.1 传统医学科研的命题与假说184
 - 5.1.2 传统医学科研的数据应用185
 - 5.1.3 传统的医学科研的统计学应用186
 - 5.1.4 传统医学科研的流程186
 - 5.2 智能医学科研系统的需求187
 - 5.2.1 临床医学科研的问题187
 - 5.2.2 临床医学科研的解决思路188
 - 5.3 智能医学科研系统的设计思想190
 - 5.3.1 科研立项190
 - 5.3.2 科研设计与统计分析191
 - 5.3.3 样本数据收集与分析192
 - 5.4 智能医学科研系统的核心技术方法193
 - 5.5 智能医学科研系统的科研数据仓库建设194
 - 5.5.1 医学科研数据仓库建设的技術方法194
 - 5.5.2 医学科研数据仓库的建设过程196
 - 5.5.3 科研数据仓库的数据安全198
 - 5.6 智能医学科研系统的核心功能设计198
 - 5.7 智能医学科研系统的整体功能设计199
 - 5.7.1 智能医学科研系统主要功能200
 - 5.7.2 智能医学科研系统的模块设计和应用实现202
 - 5.7.3 智能医学科研系统的评估方法211
 - 5.8 智能医学科研系统的应用价值215
 - 5.9 本章小结218
- 第6章 数据挖掘应用系统的开发219
 - 6.1 数据挖掘应用系统的意义219
 - 6.2 IMRS系统设计221
 - 6.2.1 对数据源的分析221
 - 6.2.2 数据挖掘应用系统IMRS的总体设计224
 - 6.3 IMRS异常侦测模型的开发232
 - 6.3.1 异常侦测模型的功能展示232
 - 6.3.2 数据挖掘技术开发要点236

- 6.4 IMRS特征抽取模型的开发242
 - 6.4.1 特征抽取模型的功能展示242
 - 6.4.2 数据挖掘技术开发要点243
- 6.5 IMRS智能统计模型的开发255
 - 6.5.1 回归模型的开发实现255
 - 6.5.2 线性相关模型的开发实现267
- 6.6 IMRS的算法开发271
 - 6.6.1 相关因子算法SRCF的实现271
 - 6.6.2 朴素贝叶斯分类算法的实现275
- 6.7 本章小结280
- 第7章 数据挖掘应用系统的应用281
 - 7.1 分布探索282
 - 7.1.1 两维度聚类模型应用282
 - 7.1.2 高维度聚类模型应用287
 - 7.2 关系探索289
 - 7.2.1 关联规则的应用289
 - 7.2.2 特性选择的应用292
 - 7.3 特征探索297
 - 7.3.1 不稳定心绞痛的特征总结297
 - 7.3.2 动脉硬化心脏病的临床特征302
 - 7.4 异常探索305
 - 7.4.1 生理指标的异常侦测305
 - 7.4.2 异常侦测模型比较307
 - 7.5 推测探索308
 - 7.6 应用系统的高级应用310
 - 7.6.1 异常侦测的高级用法310
 - 7.6.2 关联规则的高级应用315
 - 7.7 本章小结320
- 第8章 数据挖掘工具的应用321
 - 8.1 应用Oracle Data Mining321
 - 8.1.1 ODM数据挖掘流程322
 - 8.1.2 ODM算法模型323
 - 8.1.3 ODM算法应用327
 - 8.2 应用IBM SPSS Modeler351
 - 8.2.1 IBM SPSS Modeler介绍351
 - 8.2.2 SPSS Modeler独立应用352
 - 8.2.3 SPSS Modeler与应用系统的联合应用359
 - 8.3 本章小结367
- 参考文献368

《数据挖掘技术与工程实践》

精彩短评

- 1、作者用大多数篇幅拿自己的一个医药方面的数据挖掘软件产品作例子，使用 Oracle 的数据挖掘API。列出的源代码多是 Oracle 的存储过程。文风是国产式的，也许直接去看 Oracle 的文档还舒服些。
- 2、深入浅出，非常好的数据挖掘方面的书，知识点在讲解之后，配有示例和相应的应用场景，易于理解。

《数据挖掘技术与工程实践》

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.tushu111.com