

国际服务外包系列教材

数据挖掘基础与 应用实例

Foundation and
Application of Data Mining

蒋盛益
张钰莎 编 著
王连喜

Textbook Series of International
Service Outsourcing



经济科学出版社
Economic Science Press



国际服务外包系列教材

数据挖掘基础与 应用实例

Foundation and
Application of Data Mining

蒋盛益
张钰莎 编 著
王连喜

Textbook Series of International
Service Outsourcing



经济科学出版社
Economic Science Press

图书在版编目 (CIP) 数据

数据挖掘基础与应用实例 / 蒋盛益, 张钰莎, 王连喜编著.

—北京: 经济科学出版社, 2014. 12

国际服务外包系列教材

ISBN 978 - 7 - 5141 - 5240 - 1

I. ①数… II. ①蒋…②张…③王… III. ①数据处理 - 教材 IV. ①TP274

中国版本图书馆 CIP 数据核字 (2014) 第 281922 号

责任编辑: 王冬玲

责任校对: 刘 昕

责任印制: 邱 天

数据挖掘基础与应用实例

蒋盛益 张钰莎 王连喜 编著

经济科学出版社出版、发行 新华书店经销

社址: 北京市海淀区阜成路甲 28 号 邮编: 100142

总编部电话: 010 - 88191217 发行部电话: 010 - 88191522

网址: [www. esp. com. cn](http://www.esp.com.cn)

电子邮件: [esp@ esp. com. cn](mailto:esp@esp.com.cn)

天猫网店: 经济科学出版社旗舰店

网址: [http: //jjkxcbs. tmall. com](http://jjkxcbs.tmall.com)

北京万友印刷有限公司印装

787 × 1092 16 开 22.5 印张 550000 字

2015 年 5 月第 1 版 2015 年 5 月第 1 次印刷

ISBN 978 - 7 - 5141 - 5240 - 1 定价: 45.00 元

(图书出现印装问题, 本社负责调换。电话: 010 - 88191502)

(版权所有 侵权必究 举报电话: 010 - 88191586)

电子邮箱: [dbts@ esp. com. cn](mailto:dbts@esp.com.cn)

编委会名单

主任委员：隋广军

副主任委员：顾也力 郑建荣

委 员：黄永智 毕惠阳 李铁立

姜灵敏 林吉双 黄立军

熊海涛 蒋吉频 魏 青

曾 增

总 序

自 21 世纪以来,我国承接美欧日等国家和地区的国际服务外包呈加速发展之势;2013 年,我国承接国际服务外包执行金额为 454.1 亿美元,现已成为全球第二服务外包接包国。伴随服务外包产业的迅速发展,我国能熟练从事国际服务外包业务中高端人才的短缺问题日益凸显出来。因此,尽快培养国际服务外包产业所需的中高端人才,已成为促进我国服务外包产业持续、快速和健康发展的当务之急。

广东外语外贸大学国际服务外包研究院和国际服务外包人才培养基地,是全国普通高等院校最早成立的国际服务外包研究和人才培养的专门机构。2009 年 10 月以来,国际服务外包研究院承接国际服务外包的理论研究和政府咨询等课题 60 余项,发表论文 300 余篇;目前,广东外语外贸大学国际服务外包研究院已成为华南地区国际服务外包理论研究中心、政府决策咨询智库。五年来,广东外语外贸大学国际服务外包人才培养基地共培训软件架构师、软件测试工程师和网络工程师等 IT 类高校“双师型”教师 200 余人;培养和培训 ITO、BPO、KPO 等适用型大学毕业生 2 000 余人;为 IBM、西艾、从兴等服务外包企业定制培训服务外包商务英语和相关业务流程专业人才 500 余人;培训服务外包企业和政府中高层管理人员近万人。经过几年来对服务外包人才培养模式与实践的有益探索,广东外语外贸大学国际服务人才培养基地已成为广东省服务外包“双师型”教师资源库、大学毕业生适用型人才交付中心、企业和政府管理人员短期培训中心。

广东外语外贸大学作为广东省国际服务外包中高端人才培养基地,为更好地发挥学校在国际化人才培养的优势,进一步提高国际服务外包和国际服务经济人才培养的质量,特组织专家学者编写了本套教材。本次编写和出版的教材包括《服务外包概论》、《国际服务外包实务》、《广东国际服务外包案例》、《国际服务外包营销》、《印度国际服务外包经典案例》、《服务外包园区发展的理论与实践》、《国际服务经济概论》、《国际服务贸易战略与实务》、《国际金融服务实务》、《国际服务经济组织与管理概论》、《Java 软件工程师培训教程》、《云计算基础、应用与产业发展》、《数据挖掘基础与应用实例》、《物联网与产业发展》、《中外艺术创意经典 100 例》和《服务外包:理论·实践·创新》共

16 部。

培训国际服务外包和国际服务经济产业所需的中高端人才是一项系统工程，其中，编写出能够既反映国际服务外包和国际服务经济发展理论又符合国际服务外包和国际服务经济发展实践的教材就显得尤其重要，我们希望本套教材的出版能够为国际服务外包和国际服务经济人才的培养尽一份力量，同时，我们也真诚地欢迎各位读者对本套教材的不足之处提出修改的意见和建议，以期进一步提高我们教材编写的质量。

广东外语外贸大学国际服务外包系列教材编写委员会

2014 年 10 月

前言

随着信息技术的普及、大数据时代的到来,越来越多的人意识到数据对于一个单位巨大的价值,数据深度分析的需求也越来越大,但数据的激增加深了数据处理的难度。如何实现从信息到知识的转变?如何从海量数据中挖掘出有价值的信息?数据挖掘技术成为一种有效的工具。为了适应社会对数据挖掘技术人才的需求,许多高等院校在计算机、统计、经济和管理类专业逐步开设了数据挖掘课程。

数据挖掘是一个多学科交叉的综合研究领域。它融合了数据库技术、人工智能、机器学习、统计分析等多个学科领域的技术。本书在介绍了数据挖掘的基本原理和方法之后,讲述了多个应用领域案例,旨在使读者了解和掌握数据挖掘技术的理念和算法,熟悉数据挖掘技术应用的流程和分析方法,引导读者理解和利用数据挖掘技术解决实际领域中的现实问题,从而为今后的数据分析工作夯实基础。全书分为三大部分,包括上篇——入门篇、中篇——基础篇和下篇——提高篇,共10章。

入门篇从全局视角对数据挖掘的基本概念、任务、建模过程、应用前景以及数据挖掘工具 Clementine 软件进行介绍。具体地,第1章阐述数据挖掘的基本理论以及数据挖掘的应用前景,第2章介绍 Clementine 软件的基本使用,并围绕“跨行业数据挖掘过程标准”CRISP-DM 介绍数据挖掘过程的6个阶段。

基础篇对数据挖掘的主流分析技术进行介绍,并对一些经典算法进行了详细的描述和示例讲解,同时对部分算法进行了对比。本部分包括第3章至第7章:第3章介绍数据预处理技术,第4章讲述分类与回归方法,第5章阐述聚类分析的方法,第6章介绍关联分析方法,第7章对离群点检测方法进行了分析。

提高篇是入门篇与基础篇内容的延伸与拓展,是数据挖掘技术在不同行业领域的具体应用,包括第8章至第10章。第8章介绍 RFM 方法,第9章主要介绍文本挖掘中的基本概念和理论,包括文本预处理中的分词、文本表示和文本特征选择,文本挖掘中的文本分类、聚类、自动摘要和情感分析等方面的内容;第10章为在线社会关系挖掘,介绍了社会网络分析的基本概念,并重点介绍了社团发现方法,通过通信行业和微博领域的实际案例展示了社会网络分析的价

值。

本书除了介绍数据挖掘的经典方法之外，也参考了很多国内外的研究成果，同时也融入了作者们的部分研究成果。

本书的出版融汇了许多人的辛勤劳动。全书由蒋盛益策划框架和统稿，蒋盛益负责第1章和第7章的编写，张钰莎负责第2章、第4章、第6章的编写，王连喜负责第3章的编写，蒋盛益、张钰莎负责第5章的编写，蒋盛益、杨博泓负责第8章的编写，陈东沂、庞观松负责第9章的编写，吴美玲、杨博泓负责第10章的编写。本书的出版得到了广东外语外贸大学国际战略研究院和经济学出版社的大力支持，书中参考了许多学者的研究成果，在此一并表示衷心感谢。

由于作者学识水平有限，再加上时间仓促，书中难免会存在不足和疏漏，敬请广大读者不吝批评指正。

作者

2015年1月

目 录

上篇 数据挖掘入门篇

第1章 数据挖掘概述	3
1.1 数据挖掘引例	3
1.1.1 Target 和怀孕预测指数	3
1.1.2 Amazon 和个性化推荐	3
1.1.3 Google 用搜索关键词监测流感	4
1.1.4 智能搜索	4
1.2 数据挖掘简介	4
1.2.1 数据挖掘产生的背景	4
1.2.2 数据挖掘的定义	5
1.2.3 数据挖掘任务	6
1.2.4 数据挖掘过程	8
1.2.5 数据挖掘十大经典算法	8
1.3 数据挖掘应用	10
1.3.1 商业领域的应用	10
1.3.2 互联网技术领域的应用	12
1.3.3 其他应用领域	14
1.4 数据挖掘工具及软件	16
1.4.1 数据挖掘工具分类	16
1.4.2 数据挖掘工具选择需要考虑的问题	16
1.4.3 数据挖掘工具介绍	17
1.5 数据挖掘技术的前景	19
1.6 数据挖掘与隐私保护	20
1.7 本章小结	21
习题1	21

第2章 Clementine 概述	22
2.1 Clementine 简介	22
2.2 Clementine 数据流操作	23
2.2.1 生成数据流的基本过程	23
2.2.2 节点操作	24
2.2.3 超节点	26
2.3 输入、输出节点介绍	27
2.3.1 数据源节点	27
2.3.2 类型节点	32
2.3.3 表节点	33
2.3.4 数据导出节点	34
2.4 数据可视化节点介绍	35
2.4.1 数据审核节点	35
2.4.2 网络节点	37
2.5 数据挖掘建模过程	39
2.5.1 业务理解	40
2.5.2 数据理解	41
2.5.3 数据准备	41
2.5.4 建模	42
2.5.5 评估	43
2.5.6 部署	43
2.6 辛普森悖论	44
2.7 本章小结	45
习题2	45

中篇 数据挖掘基础篇

第3章 数据预处理	49
3.1 数据预处理概述	49
3.2 数据清理	50
3.2.1 缺失值的处理	50
3.2.2 噪声数据的处理	51
3.2.3 不一致数据的处理	52
3.3 数据集成	52
3.4 数据变换	53

3.4.1 数据泛化	53
3.4.2 规范化	54
3.4.3 特征构造	55
3.4.4 数值属性离散化	56
3.5 数据归约	58
3.5.1 数据立方体聚集	58
3.5.2 特征选择	60
3.5.3 抽样	60
3.6 Clementine 中相关节点介绍	61
3.6.1 导出节点	61
3.6.2 特征选择节点	65
3.6.3 抽样节点	66
3.6.4 选择节点	66
3.6.5 分区节点	66
3.6.6 分箱节点	68
3.6.7 平衡节点	70
3.6.8 排序节点	71
3.7 本章小结	71
习题3	72
第4章 分类与回归	73
4.1 分类与回归技术概述	73
4.2 决策树分类方法	74
4.2.1 决策树的基本概念	74
4.2.2 构建决策树的要素	75
4.2.3 Hunt 算法	80
4.2.4 C4.5 算法	81
4.2.5 CART 算法	88
4.2.6 C4.5 与 CART 算法对比	92
4.3 贝叶斯分类方法	93
4.3.1 贝叶斯定理	94
4.3.2 朴素贝叶斯分类算法	95
4.3.3 贝叶斯信念网络	98
4.4 K-最近邻分类方法	100
4.4.1 最近邻分类的基本概念	101
4.4.2 KNN 算法优缺点	102
4.5 Logistic 回归	102

4.5.1	二元 Logistic 回归模型	102
4.5.2	Logistic 回归模型的系数估计	103
4.5.3	显著性检验	104
4.5.4	回归方程的拟合优度检验	105
4.6	分类模型的评价	108
4.7	回归分析	110
4.7.1	线性回归模型的表示	110
4.7.2	线性回归模型的检验	111
4.7.3	非线性回归	113
4.8	集成分类	115
4.8.1	集成学习的过程描述	115
4.8.2	构建集成分类器的方法	116
4.8.3	集成分类方法的优缺点	116
4.9	Clementine 中相关节点介绍	116
4.9.1	C5.0 节点	116
4.9.2	C&R Tree 节点	119
4.9.3	BayesNet 节点	120
4.9.4	线性回归节点	122
4.9.5	逻辑回归节点	124
4.9.6	Ensemble 节点	126
4.9.7	分析节点	127
4.9.8	评估节点	128
4.10	案例 4-1: 分类技术在信用风险贷款分析中的应用	133
4.10.1	商业理解	133
4.10.2	数据理解	133
4.10.3	数据准备	135
4.10.4	数据建模	138
4.10.5	模型评估	142
4.10.6	模型部署	144
4.11	案例 4-2: Logistic 回归在旅游公司目录销售中的应用	144
4.11.1	商业理解	144
4.11.2	数据理解与数据准备	145
4.11.3	数据建模	146
4.11.4	部署	148
4.12	本章小结	149
	习题 4	149

第5章 聚类分析	153
5.1 聚类分析概述	153
5.2 相似性度量	154
5.2.1 数据及数据类型	154
5.2.2 属性之间的相似性度量	155
5.2.3 对象之间的相似性度量	157
5.3 K-means 算法及其改进	161
5.3.1 基本 K-means 算法	161
5.3.2 二分 K-means 算法	163
5.3.3 K-means 算法的拓展	163
5.4 一趟聚类算法	166
5.4.1 算法描述	166
5.4.2 一趟聚类阈值的选择策略	166
5.5 两步聚类算法	168
5.5.1 构建 CF 树	168
5.5.2 两步聚类的“亲疏程度”度量	170
5.5.3 簇数目的确定	170
5.6 聚类算法评价	171
5.6.1 确定簇数	171
5.6.2 测定聚类质量	172
5.7 Clementine 中相关节点介绍	173
5.7.1 K-means 聚类节点	174
5.7.2 Two-step 聚类节点	176
5.7.3 Khonen 聚类节点	176
5.8 案例 5-1: 电信客户细分与流失分析	179
5.8.1 商业理解	179
5.8.2 数据理解	179
5.8.3 数据准备	180
5.8.4 数据建模	181
5.8.5 结果评估	185
5.9 案例 5-2: 聚类城镇及在市场营销中的应用	186
5.9.1 创造城镇特征	186
5.9.2 创建簇	187
5.9.3 利用主题簇调整区域边界	189
5.10 本章小结	190
习题 5	190

第6章 关联规则	192
6.1 关联规则挖掘概述	192
6.2 关联规则挖掘的基本概念	193
6.3 Apriori 算法	194
6.3.1 Apriori 性质	195
6.3.2 频繁项集的产生	195
6.3.3 规则的产生	199
6.3.4 关联规则的评价	201
6.4 关联规则扩展	204
6.4.1 关联规则分类	204
6.4.2 多层次关联规则	204
6.4.3 多维度关联规则	205
6.4.4 定量关联规则	205
6.4.5 基于约束的关联规则	206
6.4.6 序列模式挖掘	206
6.5 Clementine 中 Apriori 节点介绍	207
6.6 案例 6-1: 移动业务关联分析	209
6.6.1 商业理解	209
6.6.2 数据理解阶段	209
6.6.3 数据准备阶段	211
6.6.4 建模阶段	213
6.6.5 模型评估	216
6.6.6 部署阶段	218
6.7 案例 6-2: 超市购物篮分析	219
6.7.1 商业理解	219
6.7.2 数据理解	219
6.7.3 数据准备	220
6.7.4 建立模型	221
6.7.5 模型评估和应用	224
6.8 本章小结	225
习题 6	225
第7章 离群点检测	228
7.1 离群点检测概念	228
7.2 基于统计的方法	229
7.3 基于相对密度的离群点检测方法	230

目 录

7.4 基于聚类的离群点检测方法	236
7.4.1 基于对象的离群因子检测方法	237
7.4.2 基于簇的离群因子检测方法	239
7.4.3 基于聚类的动态数据离群点检测方法	241
7.5 离群点检测方法的评估	242
7.6 Clementine 中的 Anomaly 节点介绍	242
7.7 案例 7-1: 离群点检测在癌症诊断中的应用	244
7.7.1 商业理解	244
7.7.2 数据理解	244
7.7.3 数据准备	245
7.7.4 数据建模与评估	245
7.8 案例 7-2: 离群点检测在网络入侵检测中的应用	246
7.8.1 商业理解	246
7.8.2 数据理解	247
7.8.3 数据准备	248
7.8.4 数据建模与评估	249
7.9 本章小结	251
习题 7	251

下篇 数据挖掘提高篇

第 8 章 RFM 分析	255
8.1 RFM 分析的基本原理	255
8.2 RFM 模型的应用场景	256
8.3 Clementine 中相关节点介绍	257
8.3.1 RFM 汇总节点	258
8.3.2 RFM 分析节点	259
8.4 案例 8-1: 识别促销的目标客户	261
8.4.1 数据理解	261
8.4.2 识别消费额度高的客户	262
8.4.3 预测促销目标客户的响应	264
8.5 案例 8-2: RFM 模型在销售数据分析中的应用	267
8.5.1 数据理解	267
8.5.2 数据准备	268
8.5.3 数据建模	269
8.5.4 结果评估	271

8.6 本章小结	272
第9章 文本挖掘	273
9.1 分词技术	273
9.1.1 分词挑战	273
9.1.2 分词方法	274
9.1.3 常见分词工具	276
9.2 文本向量化	277
9.2.1 向量空间模型	277
9.2.2 文本特征选择	278
9.3 文本聚类	279
9.3.1 文本相似度计算	279
9.3.2 文本聚类过程	280
9.4 文本分类	281
9.4.1 文本分类的概念	281
9.4.2 常用文本分类算法	281
9.4.3 常用基准语料与模型评估	285
9.5 文档自动摘要	286
9.5.1 文档自动摘要的类型	287
9.5.2 相关技术	287
9.5.3 自动文摘的关键问题	289
9.5.4 性能评估	290
9.6 文本情感分析	291
9.6.1 文本情感分析概念	291
9.6.2 文本情感分析技术	292
9.6.3 文本情感分析的应用	293
9.7 案例9-1: 跨语言智能学术搜索系统	294
9.7.1 混合语种文本分词	295
9.7.2 基于机器翻译的跨语言信息检索	295
9.7.3 不同语种文本的搜索结果聚类	296
9.7.4 基于聚类的个性化信息检索	296
9.7.5 基于聚类的查询扩展	297
9.7.6 其他检索便利工具	298
9.7.7 系统性能评估	298
9.8 案例9-2: 基于文本分类的微博平台潜在客户识别	304
9.8.1 商业理解	304
9.8.2 数据理解	305

目 录

9.8.3 数据准备	305
9.8.4 数据建模	306
9.8.5 模型评估及应用	307
9.9 本章小结	312
第10章 社会网络分析	314
10.1 社会网络分析概述	314
10.1.1 社会网络分析相关概念	314
10.1.2 中心性	315
10.1.3 权威性	316
10.2 社区检测	316
10.2.1 基于分割的 GN 算法	317
10.2.2 基于模块度优化的 CNM 算法	319
10.2.3 面向加权网络的随机漫步模型算法	320
10.2.4 BGLL 算法与层次性	321
10.2.5 CPM 算法与重叠性	322
10.2.6 动态网络的社区检测算法	323
10.2.7 社区检测质量评价方法	326
10.2.8 社会网络分析软件	327
10.3 案例 10-1: 基于社区检测的通信业客户细分	328
10.3.1 数据理解	328
10.3.2 数据预处理	328
10.3.3 社团检测	329
10.3.4 社团的通话特征分析	330
10.3.5 社团的客户属性分析	330
10.3.6 社团的中心客户发现	331
10.3.7 基于社团检测的电信客户细分的应用	333
10.4 案例 10-2: 微博用户圈识别	333
10.4.1 数据理解	333
10.4.2 数据预处理	333
10.4.3 社团检测	335
10.4.4 结果分析	335
10.5 本章小结	337
附录 数据挖掘常用资源列表	338
参考文献	340