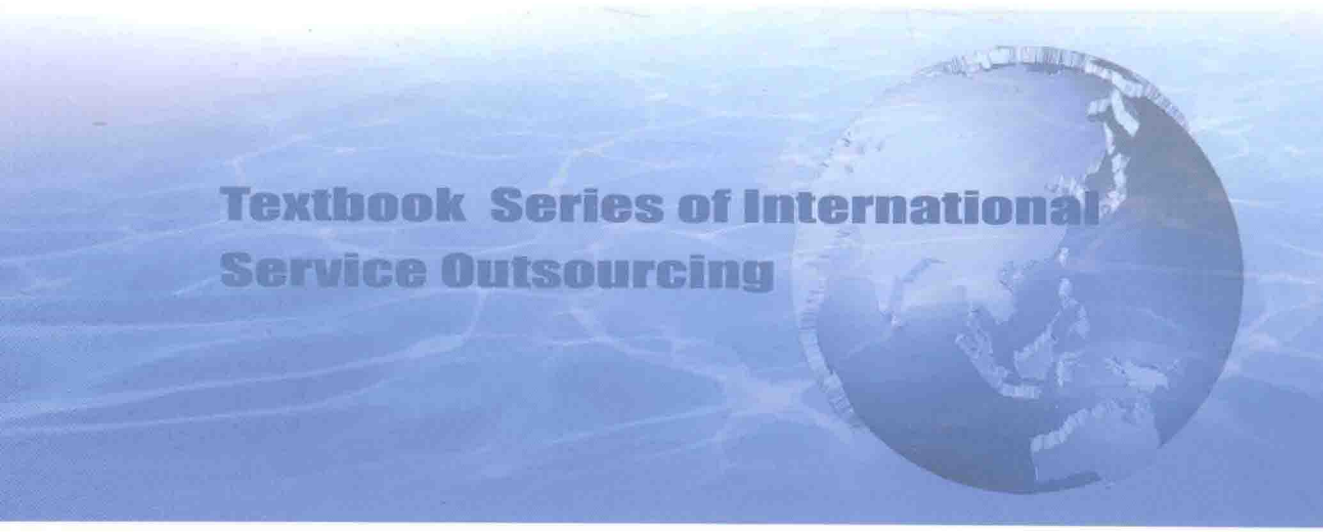
国际服务外包系列教材

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



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
电子邮件：esp@esp.com.cn
天猫网店：经济科学出版社旗舰店
网址：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

编委会名单

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训基地，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训的专门机构。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训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训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训基地，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、经济和管理类等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础概念和理论，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科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

目　录