当代统计方法与应用丛书

李金昌 主编

基于统计视角的数据挖掘研究

JIYU TONGJI SHIJIAO DE SHUJU WAJUE YANJIU

徐雪琪 著

浙江工商大学出版社
Zhejiang Gongshang University Press

基于统计视角的数据挖掘研究

JIYU TONGJI SHIJIAO DE SHUJU WAJUE YANJIU

徐雪琪 著

浙江工商大学出版社

图书在版编目（CIP）数据

基于统计视角的数据挖掘研究 / 徐雪琪著. —杭州：浙江工商大学出版社，2010.12
ISBN 978-7-81140-268-1

Ⅰ. ①基… Ⅱ. ①徐… Ⅲ. ①数据采集－研究
Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2010）第 255669 号

基于统计视角的数据挖掘研究

徐雪琪 著

责任编辑 白小平
责任校对 周敏燕
封面设计 刘 韵
责任印刷 汪 俊
出版发行 浙江工商大学出版社
（杭州市教工路 198 号 邮政编码 310012）
（E-mail：zjgsupress@163.com）
（网址：http://www.zjgsupress.com）
电话：0571－88904980，88831806（传真）
排 版 杭州中大图文设计有限公司
印 刷 杭州杭新印务有限公司
开 本 880mm×1230mm 1/32
印 张 6.25
字 数 170 千
版 印 次 2010 年 12 月第 1 版 2010 年 12 月第 1 次印刷
书 号 ISBN 978-7-81140-268-1
定 价 25.00 元

“当代统计方法与应用”丛书编委会

总　序

统计学的突出特征是通过数据分析得到客观而真实的结论，帮助人们提高对研究对象的认识，因而统计学是探求真理不可缺少的工具。作为定量认识的有力工具，统计方法适用于对自然、社会、经济等广泛现象的探索，具有普适性。正如著名统计学家 C. R. 劳所说：当今，在人类活动努力的一切范围内，统计学已经成为一种万能的、强有力的和不可缺少的研究工具。

随着科学技术的发展，统计学也在不断吸收、借鉴和融合其他学科知识的过程中，深化、丰富和完善其自身理论，不仅应用领域进一步拓展，而且为推出新技术和新方法提供了很好的契机和平台。与此同时，社会经济发展过程中出现的一些新情况、新问题，也对统计学发展提出了新的、更高的要求。当前，就统计学的研究对象而言，所涉及的范围如此之广，所面临的“未知”如此之多，迫使我们需要不断夯实和拓展统计学科基础，进一步加强统计方法及其应用研究，并把相应的研究成果加以总结和整理。这正是我们依托浙江省高校人文社科重点研究基地（浙江工商大学统计学）编撰“当代统计方法与应用”丛书的目的所在。

浙江工商大学统计学专业是学校传统优势学科之一，在国内相同领域具有较大的影响力。学院于 1980 年开始招收本科生，1990 年获准设立硕士学位点，1996 年被评为浙江省重点扶植学科，2002 年被评为浙江省重点学科和重点专业，2003 年获准设立博士学位点，2004 年被列为浙江省首批省属高校人文社科重点研究基地，2007 年入选国家特色建设专业。迄今为止，浙江工商大学已经为社

会培养了约1500名统计学本科生、200多名统计学硕士生和20余名统计学博士生。

学科、专业的建设与发展有赖于团队建设和科学研究。在大家的共同努力下，浙江工商大学统计学科团队成为了省级教学团队和省级科研创新团队，在教学与科学研究方面取得了大量的成果。近五年来，团队在《统计研究》、《经济研究》、《管理世界》、《财贸经济》、《数量经济技术经济研究》等一流杂志及核心期刊发表学术论文150多篇，先后承担了国家自然科学基金1项、国家社会科学基金7项，省自然科学基金3项、省哲学社会科学规划项目8项，其他省部级项目30余项，累计项目总经费达800多万元，获省部级二、三等奖项20多项。目前，浙江工商大学统计学已基本形成具有较强影响力的“统计基本理论与方法”、“经济统计方法与应用”、“管理统计方法与应用”、“概率统计方法与应用”等研究方向。

“当代统计方法与应用”丛书涵盖浙江工商大学统计学科团队成员的科研著作、博士论文，以及基地立项项目的最终研究成果。希望该丛书的出版，有利于推动统计方法的创新，有利于拓宽统计应用的领域，有利于提升统计分析的高度，为促进我国统计学科的建设与发展起到应有的积极作用。

李金昌

二〇一〇年春

前　言

数据挖掘是一个年轻、活跃的研究领域。从数据挖掘概念的提出至今，这一研究领域吸引了越来越多不同专业背景的研究者。从目前国内外的研究现状看，从事数据挖掘研究的人员主要是计算机领域的学者和专家，统计界的学者和专家对数据挖掘的研究相对较少。而随着数据存储技术的不断发展，可用于数据分析的数据量越来越大，数据形式也越来越丰富，数据源也不再单纯是静态的，通过各种途径、各种渠道产生的、处于不断更新中的数据源越来越多，所有这些，对传统的统计分析技术提出了极大的挑战。笔者认为数据挖掘的出现，正是统计学适应这一变化的新的发展方向，数据挖掘并不是为了替代传统的统计分析技术，而是统计分析方法的延伸和扩展。本书从统计学视角研究数据挖掘，以期从统计学角度对数据挖掘理论的研究有所突破和创新，同时对统计学理论在数据挖掘方向的发展做出探索。

全书主要包括以下八部分内容。

第一部分阐明选题的理论与现实意义，回顾相关内容的研究动态，开展文献综述，提出研究内容结构与方法，给出研究的难点和创新。

第二部分从统计视角对数据挖掘理论体系进行研究。通过对数据挖掘与统计学理论基础、方法等方面的比较，提出了基于统计视角的数据挖掘理论体系，有利于改变目前研究中人们对统计学与数据挖掘之间理解的模糊状态。

第三部分对现有数据挖掘中的统计方法进行综述。首先，对数据挖掘数据源、属性类型和功能三个基本问题作了界定；其次，对实现其中的关联、聚类、分类和回归四项功能的统计方法进行综述，并对其中一些统计方法从数据挖掘应用角度作了一些完善和改进。

第四部分对数据挖掘中的统计方法作了进一步研究。主要进行了特征描述统计方法研究和聚类挖掘距离函数和相似系数研究。在特征描述统计方法研究中，提出了在本书设计的可视化数据挖掘系统 LavaMiner 中将把特征描述作为一个独立的挖掘功能模块来实现的思路，提出了特征描述过程模型，进而根据建立的过程模型系统地研究了各个步骤可采用的统计方法。在聚类挖掘距离函数和相似系数研究中，系统地研究了各类属性的距离函数和相似系数，并分析了各个距离函数和相似系数的优缺点或适用性。

第五部分进行数据挖掘质量问题研究。从数据挖掘整个过程考虑把数据挖掘质量问题分为以下三类：源数据的质量问题、数据集成时的质量问题和数据分析时的质量问题，并针对不同问题从统计学的角度分别提出了相应有效的解决方法。

第六部分是可视化数据挖掘原型的实现。首先从数据挖掘原型的应用场景、使用对象、过程模型和模型表示四个方面研究了数据挖掘原型设计基础，然后设计并部分实现了可视化数据挖掘原型系统 LavaMiner，该系统具有灵活的操作过程、便捷的扩展方式和全面的接口封装三大特点。这是本书的另一项重要研究内容。

第七部分是数据挖掘技术在实际数据库上的应用示例。完成了数据挖掘在浙江省联通炫铃用户实际数据库中的应用示例，首先，介绍了实验数据库；其次，分不同时间段来实现最有价值客户的挖掘；最后，作了铃音关联挖掘。

第八部分对全书研究内容进行总结，并对下一步的研究作出展望。

在本书的撰写过程中，参考了国内外众多学者的相关研究成果，

书中对于他人成果的引用、参考做了标注和说明，但仍有可能存在一些遗漏的思想借用或观点转述，若未能注明，在此深表歉意！

由于本人的水平和学识有限，不妥之处在所难免，恳请读者和学界同人批评和指正！

作 者

2010 年 12 月

目　录

第一章
绪　论

第一节　选题意义

一、理论意义

数据挖掘是一个年轻、活跃的研究领域。从数据挖掘概念的提出至今，这一研究领域吸引了越来越多不同专业背景的研究者。从技术层面讲，数据挖掘集人工智能、统计学、数据库管理、数据仓库、可视化、并行计算、决策支持为一体，利用数据库、数据仓库技术存储和管理数据，利用统计学方法和人工智能分析数据。从目前国内外的研究现状看，从事数据挖掘研究的人员主要是计算机领域的学者和专家，统计界的学者和专家对数据挖掘的研究较少。而随着数据存储技术的不断发展，可用于数据分析的数据量越来越大，数据形式也越来越丰富，数据源也不再单纯是静态的，通过各种途径、各种渠道产生的、处于不断更新中的数据源越来越多，所有这些，对传统的统计分析技术提出了极大的挑战。笔者认为数据挖掘的出现，正是统计学适应这一变化的新的发展方向，数据挖掘并不是为了替代传统的统计分析技术，而是统计分析方法的延伸和扩展。所以笔者在多年思考和研究的基础上，确立以“基于统计视角的数据挖掘研究”为博士论文选题，以期从统计学角度对数据挖掘理论的研究有所突破和创新，同时对统计学理论在数据挖掘方向的发展做出探索。

二、现实意义

随着数据库、数据仓库技术的飞速发展以及数据库管理系统的广泛应用，人们积累的数据正以前所未有的速度急剧增加，最近几十年产生了很多超大型数据库，遍及超级销售市场、银行、天文学研究、粒子物理研究、化学研究、医学研究以及政府统计等各个领域。例如，全球最大零售商 Wal-Mart 的数据仓库容量已达到 101TB，美孚石油公司计划存贮的有关石油开采数据将达 10^{14} 字节，人类基因组计划也已收集了几千兆个相关数据，等等。在这个充满数据的数字化、信息化、全球化的时代，如此规模甚至更大的数据库将是人们不得不面对的一个突出问题。如果我们对数据库内的数据量作一个形象的说明，假定一个数据库包含有 15TB 的数据，那么这些数据以书面形式发布，则需要一个 724.05km 长的书架来装载。但是我们知道，数据库作为一种资源，本身并没有什么直接的价值，有价值的是从中所能获得的知识和信息。数据挖掘正是基于这种需要而发展起来的。从数据挖掘演变的过程来看，早期的数据挖掘研究较多地注重于与人工智能的结合来分析数据，而最近人们逐渐发现数据挖掘中有许多工作可以由统计方法来完成，并且认为最好的策略是将统计方法与数据挖掘有机地结合起来。所以，在此背景下开展基于统计视角的数据挖掘研究有如下的现实意义：

第一，本书对数据挖掘与统计方法的结合进行研究，使统计方法适应数据量的变化，继续发挥其处理数据、分析数据的重要作用。

第二，本书从数据挖掘整个过程对数据质量问题进行研究，有利于指导实际数据挖掘工作，提高数据挖掘质量。

第三，本书设计并部分实现的可视化数据挖掘原型系统，具有很好的实用价值。

第二节 研究动态与文献综述

一、研究动态

1989年8月在美国底特律召开的第11届国际联合人工智能学术会议(IJCAI—89)上,Gregory Piatetsky-Shapiro组织了"数据库中的知识发现"(KDD:Knowledge Discovery in Database)专题讨论会,该讨论会的重点是强调发现(Discovery)的方法以及发现的知识(Knowledge)两个方面,这是基于数据挖掘概念的首次国际学术会议。

Gregory在为该会命名时,曾考虑过数据挖掘(data mining)、知识挖掘(knowledge mining)、知识抽取(knowledge extraction)、数据库挖掘(database mining)这四个术语。但是,因为数据挖掘已经在数据库领域被使用过,所以觉得不具备吸引力,而且统计学界对"数据挖掘"这个术语抱有偏见,认为挖掘的语义太单调,并且没有指明挖掘的内容。"知识挖掘"和"知识抽取"并不比"数据挖掘"更理想,而"数据库挖掘"是HNC公司的注册商标(database mining TM)。因此,最终选择了数据库中的知识发现,简称KDD,作为专题讨论会的名称。随着该学术会议的召开,KDD开始在人工智能(AI)和机器学习(machine learning)领域变得流行起来了。

学术界有部分观点将KDD作为知识发现的整个过程的统称,而将数据挖掘作为KDD过程中的一个步骤。然而数据库领域研究人员经常与商业伙伴以及出版社接触,非学术人员(商业伙伴和出版商)更容易接受"数据挖掘"这一术语,因此,数据挖掘在商业出版社比KDD或其他名词更为流行。目前数据挖掘的研究已经不仅仅局限于数据库中数据的挖掘,挖掘的对象包括文本、图像、声音等多种非结构化数据格式,从这个意义上讲,数据挖掘比KDD更贴近这一术语本身代表的含义。到2010年12月1日,在www.google.com.hk上搜索"data mining",搜索结果有1680万条,对"KDD"搜索结果只有

158 万条，而“Knowledge Discovery in Database”的搜索结果也只有636 万条。这也从一个侧面反映出数据挖掘这一术语的流行程度。因此，到目前为止，数据挖掘和 KDD 这两个术语几乎是同义的，本书所指的数据挖掘也是取其广义的概念，即表示知识发现的整个过程，而不仅仅是应用算法这个步骤。

随后在 1991 年、1993 年和 1994 年都举行了 KDD 专题讨论会，来自各个领域的研究人员和应用开发者集中讨论了数据统计、海量数据分析算法、知识表示和知识运用等问题。随着参与科研和开发人员的不断增加，国际 KDD 组委会于 1995 年把专题讨论会发展成为国际年会。在加拿大的蒙特利尔市召开了第一届 KDD 国际学术会议。在这次会议上“数据挖掘”(Data Mining)概念第一次由 Usama Fayyad 提出。以后每年召开一次，其会议名称全称为“ACM SIGKDD(Special Interested Group on Knowledge Discovery in Databases) International Conference on Knowledge Discovery and Data Mining”。参加人数由几十人发展到上千人，研究重点也逐渐从发现方法转向系统应用，并且注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。其中 1997 年第三届 KDD 国际学术大会上进行的数据挖掘工具的竞赛评奖活动，就是一个生动的证明。1998 年，在美国纽约举行的第四届知识发现与数据挖掘国际学术会议上，与会者不仅进行了学术讨论，而且领略了 30 多家软件公司展示的数据挖掘软件产品。最近一次的 KDD 国际学术会于 2010 年 7 月 25 日－28 日在美国华盛顿举行。

除了美国人工智能协会主办的 KDD 年会外，还有许多的数据挖掘年会，包括 PAKDD、PKDD、SIAM－Data Mining 等。PAKDD (Pacific-Asia conference on knowledge discovery and data mining) 是亚太平洋地区数据挖掘年会，从 1997 年开始，每年召开一次，至今已召开了 14 届，其中 1999 年的 PAKDD 在我国北京召开，2007 年的 PAKDD 也在我国南京举行，最近一届 PAKDD 于 2010 年 6 月 21－24 日在印度的 Hyderabad 召开。PKDD(European symposium on principles of data mining and knowledge discovery)是欧洲数据

挖掘会议，也是从1997年开始，每年召开一次，至今也已召开了14届，最近一届于2010年9月20－24日在巴塞罗那召开。SIAM-Data Mining(Society for Industrial and Applied Mathematics)是SIAM组织召开的数据挖掘讨论会，2001年4月召开第1届讨论会，专注于科学数据的数据挖掘，以后每年召开一次，至今已召开了10届，第十届SIAM数据挖掘国际会议于2010年4月29日－5月1日在美国Columbus召开。

国外已有许多专门的工作组，从事数据挖掘领域的研究。比较著名的有R. Agrawal领导下的IBM Almaden实验室的数据挖掘工作组；J. Han带领下的SFU工作组；Stanford大学的Ullman领导的关联规则研究小组；Minnesota大学的Kumar领导的并行数据挖掘研究小组；新西兰怀卡托大学Ian H. Witten教授领导下的Weka工作组，等等。他们提出了许多好的数据挖掘算法，并实现了数据挖掘工具，为该领域的发展奠定了一定的基础。其中Ian H. Witten教授在2004年荣获了国际信息处理研究协会(IFIP)颁发的Namur奖项，这是一个两年一度、用于奖励那些在信息和通信技术的社会应用方面做出杰出贡献及具有国际影响的荣誉奖项。2005年8月，在第11届ACM SIGKDD国际学术会议上，Weka工作组荣获了数据挖掘和知识探索的最高服务奖，Weka被誉为数据挖掘和机器学习历史上的里程碑。

此外，数据库、人工智能、信息处理、知识工程等领域的国际学术刊物也纷纷开辟了KDD专题或专刊。IEEE的*Knowledge and Data Engineering*会刊率先在1993年出版了KDD技术专刊，所发表的5篇论文代表了当时KDD研究的最新成果和动态，较全面地论述了KDD系统方法论、发现结果的评价、KDD系统设计的逻辑方法，集中讨论了鉴于数据库的动态性冗余、高噪声和不确定性、空值等问题，KDD系统与其他传统的机器学习、专家系统、人工神经网络、数理统计分析系统的联系和区别，以及相应的基本对策。目前较有影响的学术期刊是*Data Mining and Knowledge Discovery*，1997年3月创刊，由M. Fayyad主办，Kluwers出版社出版。

不仅如此，在 Internet 上还有不少 KDD 电子出版物，其中以半月刊 *Knowledge Discovery Nuggets* 最为权威。另一份在线周刊为 DS(DS 代表决策支持)，1997 年 10 月 7 日开始出版，可向 dstrial@tgc.com 提出免费订阅申请。ACM-SIGMOD 还在 SIGKDD 成员中，出版了一种季刊电子通信 SIGKDD Explorations。在网上，还有一个自由论坛 DM Email Club，人们通过电子邮件相互讨论 DMKD 的热点问题。

与国外相比，国内对 DM 的研究稍晚，但最近几年有较大的发展。1993 年国家自然科学基金首次支持我们对该领域的研究项目。目前，国内的许多科研单位和高等院校竞相开展数据挖掘的基础理论及应用研究。例如，复旦大学施伯乐教授领导开发了数据挖掘工具集 AMINER，北京大学智能科学系的唐世渭和杨冬青教授领导开发了基于空间数据挖掘的客户分析系统模型 CASDM。此外，清华大学周立柱教授领导的数据挖掘研究小组，四川大学唐常杰教授领导的针对时间序列方面的数据挖掘研究小组，中国科技大学蔡庆生教授领导的针对关联规则的研究小组，复旦大学朱扬勇教授领导的数据挖掘工作组等，都取得了许多重要的研究成果。在数据挖掘算法研究方面，中科院计算所史忠值研究员、清华大学石纯一、陆玉昌教授、武汉大学李德仁院士、北京科技大学杨炳儒教授、复旦大学周傲英教授等都取得了许多重要的研究成果。国内统计学界，中国人民大学统计学院开辟了“统计学与数据挖掘”研究专栏，并于 2001 年春季成立了数据挖掘研究中心。该中心是中国人民大学统计学院下的二级非营利性学术组织，现有专兼职研究人员 10 人，现任中心主任为吴喜之教授。它是国内较早开展数据挖掘应用和理论探索的团队，也是在经济学科下较早研究数据挖掘应用的组织，现在正承担着国家级重大项目 2 个。厦门大学计划统计系也于 2007 年底成立了数据挖掘中心，它致力于商业智能和数据挖掘的学术研究和实务应用。目前拥有成员逾百人，团队领军人物朱建平教授也有许多研究成果，其专著《数据挖掘的统计方法与实践》于 2005 年 12 月由中国统计出版社出版。台湾辅仁大学谢邦昌教授是我国目前统计领域从