厦门大学统计学系列丛书

The Statistical Method and Practice of Data Mining

数据挖掘的统计方法及实践

朱建平 著

中国统计出版社
China Statistics Press

厦门大学统计学系列丛书

The Statistical Method and Practice of Data Mining

数据挖掘的统计方法及实践

朱建平 著

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

数据挖掘的统计方法及实践/朱建平著.
—北京:中国统计出版社,2005.6
ISBN 7—5037—4764—1

Ⅰ.数…
Ⅱ.朱…
Ⅲ.数据采集—统计方法—高等学校—教材
Ⅳ.TP274—32

中国版本图书馆 CIP 数据核字(2005)第 067179

数据挖掘的统计方法及实践

作　　者/朱建平
责任编辑/申明九
封面设计/艺编广告·刘　璐　周　颖
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 75 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
电　　话/(010)63459084、63266600—225000(发行部)
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/787×1092mm　1/18
字　　数/120 千字
印　　张/13.5
版　　别/2005 年 10 月第 1 版
版　　次/2005 年 10 月第 1 次印刷
书　　号/ISBN 7—5037—4764—1/TP·14
定　　价/25.00 元

厦门大学统计学系列丛书编审委员会

出版说明

厦门大学统计学专业由王亚南先生创办于1950年，迄今已有50余年历史；厦门大学国民经济与核算研究所由钱伯海先生创办于1983年，是国内最早建立的国民核算专门研究机构。厦大统计学人秉承严谨治学的优良传统，通过长期探索和反复实践，逐步形成了自己的学术风格和办学特色，积累了丰硕的学科建设成果，尤其在统计理论方法和国民经济核算两大领域内的建树令人瞩目。鉴于以上各方面的建设成就，厦门大学统计学科现已成为国内统计学术研究和人才培养的重要基地之一，并分别于1987年和2001年连续两次被国家教委和教育部评为国家级重点学科。

学科建设之根本在于学术创新和成果积累。我们深信，没有一流的学术成果，就没有一流的学科建设。过去数十年间，在厦门大学统计学科的旗帜之下，涌现了诸如钱伯海、黄良文、罗季荣、翁礼馨、吴玑端等众多名宿和大师，他(她)们正是以其精湛、睿智、广博的著述为学界所熟知。而今，又一批有抱负、有作为的中青年统计学者集聚在这一旗帜下，希冀以自己的智慧和努力为现代统计学的理论大厦添砖加瓦。

为适应新时期统计学国家重点学科和“211工程”建设的需要，在中国统计出版社的大力支持下，我们特筹划出版“厦门大学统计学系列丛书”。该套丛书拟由厦大统计学人的最新科研和教学成果中，遴选部分优秀的专著和教材，经系统整理和精心加工后汇集出版。其宗旨是注重学术研究的继承性、创新性、前沿性和代表性：不是既有成果的简单重复，而应与时俱进，不断发展，推陈出新；不是传统观念的老调重弹，而应倡导创新思维，拓展研究领域，跟踪学术前沿；不是文献素材的拼凑或课题报告的汇编，而应力求博采众长，融会贯通，提炼升华，自成一家之言。当然，这些目标能否达到，还有赖于各位撰述者的鼎力襄助，以及各位编审委员的敏锐判断和敬业精神。同时，也竭诚欢迎海内外统计界同仁给予建设性的批评和建议，帮助我们把这套丛书出好！

厦门大学统计学系列丛书编审委员会

2005年1月

前　言

随着计算机应用和 Internet 的日益普及，各行各业都开始采用计算机及相应的信息技术进行管理和决策，这使得各企事业单位生成、收集、存储和处理数据的能力大大提高，数据量与日俱增，海量数据层出不穷。在信息爆炸的今天，人们已经意识到数据最值钱的时代已经到来。显然，大量信息在给人们带来方便的同时也带来一系列问题，比如：信息量过大，超过了人们掌握、消化的能力；一些信息真伪难辨，从而给信息的正确应用带来困难；信息组织形式的不一致性导致难以对信息进行有效统一处理等等，这种变化使传统的数据库技术和数据处理手段已经不能满足要求。Internet 的迅猛发展也使得网络上的各种资源信息异常丰富，在其中进行信息的查找真如大海捞针。

“丰富的数据与贫乏的知识”问题越来越突出，人们也希望能够在对大量数据分析的基础上进行科学研究、商业决策或企业管理，但是目前所有的数据分析工具很难对数据进行深层次的处理，使得人们只能望“数”兴叹。面对海量数据库和大量复杂信息，如何才能从信息海洋中提取有价值的知识，进一步提高信息的利用率，这就引发了一个新的研究方向：知识发现（Knowledge Discovery in Database）以及相应的数据挖掘（Data Mining）理论和技术的研究。数据挖掘正是为了解决传统分析方法的不足，并针对大规模数据的分析处理而出现的。数据挖掘从大量数据中提取出隐藏在数据之后的有用的信息，随着其技术的发展，人们逐渐地开始认识到数据挖掘中有许多

工作可以由统计方法来完成，并取得了较好的效果，为人们的正确决策提供了很大的帮助。因此，本书的内容选定在数据挖掘的统计方法及实践上。

2000年11月在导师张润楚教授的主持下，在南开大学数学科学学院开设了Data Mining讨论班，我开始了数据挖掘领域的研究，并从大量的资料中寻找着数据挖掘研究之“道”，为我的研究内容寻找角度、理顺思路。在这样的条件下，我被该领域蕴涵的统计学问题、独特的思想方法以及广泛的应用前景所吸引，因而阅读了不少相关的文献，经过张润楚教授的精心指导，通过多年的反复思考及思想的提炼，理出了一条从统计角度研究数据挖掘技术和方法的思路。在该领域的研究过程中，导师张润楚教授和师母肖芸茹教授时时鼓励我克服所遇到的困难，我向他们对我的培养和指导致以崇高的敬意！。

近年来，在与厦门大学经济学院计划统计系硕士和博士研究生研讨后，我完成了关于数据挖掘的统计方法研究和实践方面的讲义，在此基础上提炼并整理出了这本著作。虽然本书只是从统计角度探讨数据挖掘方面的内容，但涉及面广，参考了国内外大量资料，谨向原作者致以诚挚的谢意！

在该领域的研究中，美国宾州州立大学管理科学系Dennis Lin教授、台湾辅仁大学统计资讯学系谢邦昌教授提供了部分信息和资料；在本书的完成过程中我国著名统计学家厦门大学黄良文教授多次给予鼓励；厦门大学经济学院计划统计系曾五一教授、杨灿教授在研究上给予了大力的支持，确保了本书的顺利完成；在本书的部分数据预处理和资料整理过程中，博士生刘云霞、硕士生来升强和郑晶晶做了不少的工作，在此一并表示衷心的感谢！

这里特别需要提及的是我的父母及家人，他们的祝福和鼓励一直伴随着我。我的妻子成健女士默默地为我的学习和研究做了许多工作，并承担起了繁重的家务和教育儿子的重任；我们的儿子朱森在学习上积极进取的精神和自强不息的努力，增加了我专心致志地投入到学习和研究之中的信心，此书献给他们，以表达我对他们无法言喻的感激之情。

通过本书的撰写，我深深地体会到：统计学应该随时地关注数据分析，哪里有数据，哪里就应该有统计分析。统计学方法在数据挖掘科学中发挥着重要作用，是数据挖掘研究的重要途径之一。数据挖掘需要统计方法，统计方法的恰当应用将会为数据挖掘和知识发现的结果起到意想不到的效果。

本书的完成，可以说是我在数据挖掘领域研究的一个阶段性的总结。为了进一步完善现有的研究成果，还有大量的研究工作需要去做，由于我水平有限，对数据挖掘的方法及应用的研究还很粗浅，有的内容甚至是阶段性的研究成果，不妥之处恳请专家批评指正。

朱建平

于厦门大学海滨

2005 年 1 月 5 日

摘　要

随着计算机和电子数据技术的不断发展以及 Internet 和各种局域网的广泛普及，人们获得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，遍及超级市场销售、银行存款、天文学、粒子物理、化学、医学以及政府统计等领域. 那么，如何从这些大型数据库中发现有用的信息、模式和知识？如何开发有效的挖掘方法？已成为众多科技工作者共同关注的焦点. 在过去几年，一个被称为“数据挖掘”(Data Mining) 的新领域得到了快速发展，这是一个介于统计学、模式识别、人工智能、机器学习、数据库技术以及高性能并行计算等领域的交叉新学科，已在经济、商业、金融、天文等行业得到了成功的应用，在国际上掀起了一股空前的研究热潮.

从总体上，国外在数据挖掘领域中的研究内容十分广泛，已经取得了明显的成果，如 Han, J. and Fu, Y. (1995) 等人对于定量关联规则以及其他种类的关联规则的发现研究， Mehta, M. (1996) 等人针对大型数据库快速分类算法的研究， Owen, A. B. (1999) 对分类与回归的管状邻域研究， Friedman, J. H. (1997) 对最近邻分类方法的改进，以及我们所列文献中对聚类规则的研究、数据泛化、简约和特征提取研究等. 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其应用研究，例如：模糊方法在知识发现中的应用研究；对数据立方体代数的研究；对关联规则开采算法的优化和改造；非结构化数据的知识发现以及 Web 数据挖掘. 然而，对数据挖掘技术的统计方法专门研究的不多，为了发展数据挖

掘的技术与理论，更好地应用于实际，我们将就数据挖掘的统计方法及其应用展开研究，其中包括事务性数据库的压缩、数据的排序及有向聚类分析、关联规则的进一步研究、针对时序稠密数据的基函数拟合与预测等.

我们在第一章中从技术和商业的角度介绍了数据挖掘的定义，数据挖掘与知识发现的关系，以及数据挖掘的研究对象.

第二章我们从国内外在数据挖掘领域中的发展状况，概括出了数据挖掘领域的研究成果，探讨了数据挖掘与统计学的关系. 以 Friedman, J. H. 和 Tukey, J. W. (1974) 的投影寻踪，及 Zhang, R. C. (1996) 在 Stiefel 流形上的均匀抽样等方法的思想，说明了统计学者在数据挖掘领域中所做的贡献.

第三章我们主要介绍数据的准备. 数据准备是模型开发过程中最重要的步骤之一. 从最简单的分析到最复杂的模型，所使用的数据质量是项目成功的关键. 好的数据和有效的技术一样，决定着一个模型产生有力结果的能力. 这里我们探讨了数据的选取、数据预处理、数据的变换和数据的衍生. 为我们后面的研究和分析奠定基础.

第四章我们研究了事务性数据库的压缩问题. 首先构造了事务性数据库信息系统 $S=\{U,Q,V,f\}$，其中 U 是对象 (事务) 的一个有限集 $\{x_1,x_2,...,x_N\}$，N 为所研究的事务总量，称 U 为论域；Q 是描述对象 (事务) 属性的一个有限集合，称为属性集；$V=\bigcup_{q\in Q}V_q$ 是属性项 $q\in Q$ 的值域；f 是 $U\times Q$ 到 V 的一个映射. 根据不可识别性的定义，以减少事务数增加可识别性，将原信息系统 S 压缩为 $S^*=\{U^*,Q,V,f\}$，这里 U^* 为等价关系 R_Q 将论域 U 分成等价类

族. 在此进而，我们相对于具有某种属性特征的子集 $A \subseteq Q$，将原信息系统 S 压缩为 $S_A^* = \{U^*, A, V_A, f\}$. 同时，我们利用 Shannon 的信息熵，对压缩后的信息系统或数据库的信息损失进行了统计分析.

事务数据库的列联描述也是数据库压缩的一个方面，我们定义

$$u_{ij} \triangleq \{x : f(x, q_{\alpha_1}) = a_{\alpha_1}^{(i)} \wedge f(x, q_{\alpha_2}) = a_{\alpha_2}^{(j)},$$
$$a_{\alpha_1}^{(i)} \in V_{q_{\alpha_1}}, \quad a_{\alpha_2}^{(j)} \in V_{q_{\alpha_2}}, \quad \forall x \in U\},$$

来构造二维条件属性项之间的列联关系；我们定义

$$u_{ij}^t \triangleq \{x : f(x, q_t) = a_i \wedge f(x, d) = b_j,$$
$$a_i \in V_{q_t}, \quad b_j \in V_d, \quad \forall x \in U\},$$

构造条件属性项与决策属性项之间的列联关系；定义

$$u_{ij} \triangleq \{x : f(x, q_i) = a_j, \quad q_i \in Q, \quad a_j \in V_{q_i}, \quad \forall x \in U\},$$

构造属性项与属性值之间的列联关系.

另外，我们基于 Rough 集理论、相依性、广义线性分析、多重相关性等理论和方法对事务数据库的属性项压缩进行了探讨. 其研究结果表明，Rough 集理论与现代统计方法的结合对数据库的压缩有明显的效果.

在数据挖掘的应用是检验方法好坏的重要途径. 利用我们所构造的数据库压缩方法，对某地区中国移动手机用户从 2001 年 9 月到 2002 年 3 月的缴费情况的共计 149632 条信息的原始数据库进行了压缩分析. 不仅验证了方法的可行性，而且还得到了许多有实际意义的结论.

第五章研究了数据的排序及有向聚类问题. 就排序而言, 我们以信息源和综合属性为标准研究了事务项 (样本) 的排序, 以 PCA 的思想针对事务数据库, 提出了加权构造综合属性函数方法, 利用该方法对某地区中国移动手机用户的消费情况进行综合评价, 通过可视化和统计分析验证了方法的可靠性. 同时, 针对信息系统 $S = \{U, Q, V, f\}$ 压缩成的 $r \times c$ 的双因素单向有序列联资料, 我们定义了平均秩效应为

$$\overline{R}_j = \frac{(\sum_{t=1}^{j-1} n_{.t} + 1) + (\sum_{t=1}^{j} n_{.t})}{2} = \sum_{t=1}^{j-1} n_{.t} + \frac{n_{.j} + 1}{2}, \quad j = 1, 2, ..., c,$$

秩效应为

$$R_i = q_i' \overline{R}, \qquad i = 1, 2, ..., r,$$

其中 $\overline{R} = (\overline{R}_1, \overline{R}_2, ..., \overline{R}_c)'$. 我们提出了对属性项的秩效应排序方法, 并在实际中得到了应用.

最近人们对于最近邻聚类分析及其应用研究较多 (如 Owne, B. A. (1999)), 对有序样本的聚类分析也有研究 (如 Fisher 的最优求解法), 然而对于有序资料的近邻聚类分析研究较少, 对此我们构造了有序近邻聚类分析方法, 并与 Fisher 的最优求解法比较, 不仅结果基本一致, 而且提高了运算的速度. 同时, 我们还构造了有序平均秩效应聚类分析方法, 通过计算机的可视化分析及 Smirnov 检验, 验证了该方法的可行性和可靠性.

第六章我们对数据挖掘中建立数据之间关联规则的方法进一步研究. 关联规则挖掘的研究是近几年研究较多的数据挖掘方法, 关联规则的概念首先是由 Agrawal, R., Imieliski T. and Swami, A. (1993)

提出，关联规则挖掘的主要对象是事务数据库. 我们以相应分析理论对关联规则的方法进行深入的研究，对于相应分析的研究已经引起了人们的重视 (如胡国定，张润楚 (1989) , Ven de Velden and Nedecker (2000)). 在此，我们从总信息变差的角度，以二维列联表的独立性检验入手，在探讨相应分析某些性质的同时，研究了相应分析与独立性检验的内在关系，并得出了一些有意义的结论，即

1) 在 χ^2 距离意义下，以重心距离反映 $\mathbf{F}$ 的总信息变差与以原点距离反映的总信息变差之间相差单位 1.

2) 设二维列联表的频率矩阵为 $\mathbf{F}=(f_{ij})_{r\times c}$，样本容量为 k. 检验两因素独立性的 χ^2 统计量为 W_0，以重心和原点计算因素 A 分布轮廓的度量协差阵分别为 $\mathbf{S}_r\mathbf{D}_c^{-1}$ 和 $\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}$，则

$$k\mathtt{tr}(\mathbf{S}_r\mathbf{D}_c^{-1})=W_0 \quad \text{或者} \quad k(\mathtt{tr}(\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1})-1)=W_0.$$

3) 独立性检验的 χ^2 统计量 W_0 是卡方标准化频率矩阵在正交于矩阵 $\mathbf{S}$ 或 $\mathbf{Q}$ 的最大特征值为 1 时对应的平凡子空间的空间的 k 倍变差.

在实际中，针对所研究的对象进行相应分析是否有意义，或者说对于所给的数据是否值得做这种相应分析，这就是所谓的相应分析的适应性问题. 对此我们提出了相应分析适应性检验的基本思想，利用奇异值分解理论，论证了适应性检验方法的合理性，并构造出了第 l 步适应性检验的统计量为

$$W_l=k\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-f_{i\cdot}f_{\cdot j}-\sum_{m=1}^{l}\sqrt{\beta}_m v_{mi}u_{mj})^2}{f_{i\cdot}f_{\cdot j}}.$$

在 H_0 成立时，统计量 W_l 渐近服从自由度为 $(r-l-1)(c-l-1)$ 的 χ^2 分布. 根据我们所构造的方法，概括出了相应分析适应性依次

检验的程序. 同时, 我们通过计算机模拟对多度相应分析及多维相应分析的结果进行了可视化研究, 验证了该方法的可靠性.

对于两因素之间关联程度的量度有不少学者进行过讨论 (如 Pearson, K. (1904), Cramer, H. (1946), Kendall, M. and Stuart, A. (1979)), 但是对于相应分析多度关联程度的量度很少讨论, 我们对此构造了相应分析 l 度关联系数, 即

$$C_l = \frac{\Phi_l^2}{\min\{(r-l-1),(c-l-1)\}} = \frac{\sum_{i=l+1}^{l_0}\beta_i}{\min\{(r-l-1),(c-l-1)\}}.$$

如果 C_l 值越大, 说明选取 l 个特征值 $\beta_1, \beta_2, ..., \beta_l$ 进行相应分析是有意义的.

我们在第七章研究了时序稠密数据集的拟合和建模问题. 从算法及数据处理的角度, 对多项式基函数、拉格朗日 (Lagrange) 基函数和伯恩斯坦 (Bernstein) 基函数进行了比较, 确定了 Bernstein 基函数建模是分析时序稠密数据集较为理想的方法, 并对 Bernstein 基函数的性质进行了刻划.

设稠密时间序列数据集为 $\mathbf{X}_i$, $i = 0, 1, \ldots, n$, 我们定义了以 Bernstein 基函数建立的模型为

$$\mathbf{X}(t) = \sum_{j=0}^{m} \mathbf{b}_j B_{j,m}(t) + \varepsilon(t),$$

其中 $\mathbf{b}_j$ 为待定的控制点, $B_{j,m}(t) = C_m^j t^j (1-t)^{m-j}$, $j = 0, 1, \ldots, m$ 为 Bernstein 基函数. 利用模型曲线的凸包性质, 构造了未来某一现象的发展方向, 并对此进行了预测.

我们利用所建立的上述模型对上证指数 (1A0001) 从 1996 年 1 月 2 日到 2001 年 9 月 21 日收盘价 (1362 天) 这一稠密时序资料

拟合建模，并利用分阶段资料进行移动预测模拟．对拟合和预测的结果，利用可视化方法进行了验证，取得了良好的效果．

这里需要提及的是，我们利用所研究的方法，对某地区中国移动通讯用户消费数据库、某大学大学生隐形教育调查资料和上证指数收盘价信息进行了剖析，不仅检验了方法的实用性和可靠性，而且还得到了许多有意义的结论．

Abstract

Today, people acquire data at a fantastic speed that they could not imagine before, with the development of computer and electronic data technique and a widespread use of Internet and various LANs. In recent decades, many overlarge databases have appeared in various fields, such as supermarket sales, banking savings, astronomy, particle physics, chemistry, medicine and governmental statistics and so on. How to discover useful information, pattern and knowledge from those overlarge databases and how to develop effective ways of mining data have been a focus for many scientific researchers. In the past few years, a new field called " Data Mining " has gained rapid development, which is a new crossing-subject relating to statistics, pattern identification, AI, mechanic-learning, database-techniques and high-property parallel calculation, etc. Meanwhile, it has successfully been applied in economy, commerce, finance and astronomy and brought an unprecedented tide in the world.

Generally speaking, the content of research in the field of data mining is rich. Many obvious achievements have been acquired in foreign countries, which include Han, J. and Fu, Y.'s (1993) discovery and research on the quantitative association rules and other kinds of association rule, Mehta, M.'s (1996) research on the fast classification algorithm for large databases, Owen, A. B.'s (1999) research on the Tubular Neighbors for Regression and Classification,

Friedman, J. H.'s (1997) improvement on the nearest neighbor classifiers, and the research on cluster regular, and data generalization and reduction and character extracting displayed in the references by us. Currently, the research on the basic theory and application of data mining has been carried out in many institutions and universities in our country. For example, the applied research on fuzzy method in the knowledge discovery, the research on data cube algebra, the optimum and reformation of association rule mining algorithm, the knowledge discovery of non-structured data and Web data mining. However, there is just little specific research on the statistical methods in the data mining. In order to develope the technologies and theories of data mining and apply them to practice better, this dissertation carries out a research on the statistical methods and applications of data mining, which includes the compression of transactional databases, data sorting, orderly classification analysis, and the more research of association rule, the fitting of the basic function and forecasting for the dense time series data, and so on.

In Chapter one, we introduce the definition of data mining from the angle of technology and commerce, the relationship between data mining and knowledge discovery, and the object of data mining .

In Chapter two, we generalize the research achievements in the field of data mining and explore the relationship between data mining and statistics, according to the development of the research on data mining around the world. We also showe the statisticians' con-