四川大学哲学社会科学学术著作出版基金丛书
SICHUAN DAXUE ZHEXUE SHEHUI KEXUE XUESHU ZHUZUO CHUBAN JIJIN CONGSHU

宏观经济数据挖掘理论与方法

HONGGUAN JINGJI SHUJU WAJUE LILUN YU FANGFA

何跃／著

四川大学出版社

SICHUAN DAXUE ZHEXUE SHEHUI KEXUE XUESHU ZHUZUO CHUBAN JIJIN CONGSHU
四川大学哲学社会科学学术著作出版基金丛书

宏观经济数据挖掘理论与方法

何跃／著

HONGGUAN JINGJI SHUJU WAJUE LILUN YU FANGFA

四川大学出版社

责任编辑：陈　纯
责任校对：段悟吾
封面设计：墨创文化
责任印制：王　炜

图书在版编目(CIP)数据

宏观经济学数据挖掘理论与方法 / 何跃著. —成都：四川大学出版社，2013.11
ISBN 978-7-5614-7356-6

Ⅰ.①宏… Ⅱ.①何… Ⅲ.①宏观经济分析－数据采掘－研究 Ⅳ.①F015

中国版本图书馆 CIP 数据核字（2013）第 282910 号

书名　宏观经济学数据挖掘理论与方法

著　　者　何　跃
出　　版　四川大学出版社
地　　址　成都市一环路南一段 24 号 (610065)
发　　行　四川大学出版社
书　　号　ISBN 978-7-5614-7356-6
印　　刷　郫县犀浦印刷厂
成品尺寸　170 mm×240 mm
印　　张　14.75
字　　数　275 千字
版　　次　2013 年 12 月第 1 版
印　　次　2013 年 12 月第 1 次印刷
定　　价　28.00 元

◆读者邮购本书，请与本社发行科联系。
电话：(028)85408408/(028)85401670/
(028)85408023　邮政编码：610065
◆本社图书如有印装质量问题，请
寄回出版社调换。
◆网址：http://www.scup.cn

四川大学哲学社会科学学术著作出版基金丛书
编委会

丛书序

四川大学（以下简称川大）是中国近代创办的最早一批高等教育机构中的一所。近十余年来，又经两次“强强合并”，成为学科覆盖面较广、综合实力较强的一所综合性大学。一百多年来，川大的人文社会科学在学校日益壮大的过程中，从国学研究起步，接受现代科学的洗礼，与不同的学术流派融合互动、共同成长，形成了今日既立足于中国传统，又积极面向世界的学术特征。

作为近代教育机构，四川大学的历史要从1896年设立的四川中西学堂算起。但具体到人文社会科学研究，则可以追溯到清同治十三年（1874年）由张之洞等人创办的四川尊经书院。在短短二十几年的办学历史中，书院先后培养出经学家廖平、思想家吴虞等一大批在近代中国学术思想史上影响巨大的学者，也因此使四川成为国内研究经、史、文章等中国传统之学的重镇。此后，在20世纪相当长的一段时间里，以国学为主要研究对象的近代“蜀学”成为川大人文社会科学研究的主流，拥有张森楷、龚道耕、林思进、向楚、向宗鲁、庞俊、蒙文通、刘咸炘、李植、李培甫、伍非百等一大批国内知名的学者。

近代蜀学在研究内容上以传统学术为主，在观念与方法上则立意求新。廖平的经学思想曾经作为19世纪晚期变法维新的基本理论依据之一，其知识背景上也不乏西学色彩。20世纪20年代成长起来的一批学者如庞俊、刘咸炘等人，更是亲自参与了中国传统学术向现代学术的转变。其中，蒙文通由经向史，同时又广涉四部之学，在晚年更是力图从唯物史观的角度探索中国社会与思想的演进，最能代表这一学术传统的是包容、开放并具有前瞻性的眼光。

自20世纪20年代开始，现代社会科学的深入研究也逐渐在川大开展。1922年至1924年，吴玉章在此教授经济学课程，鼓励学生通过社会科学的研究，思考“中国将来前途怎样走”的问题。1924年，学校设立了10个系，在人文社会科学6个系中，除了延续着蜀学风格的中文系外，教育、英文、历史、政治、经济5个系均着力于新的社会科学研究。这一科系的设置格局一直持续到30年代初的国立四川大学时期。

川大的另一源头是私立华西协合大学（以下简称华大）。作为教会学校，华大文科自始即以“沟通中西文化与发扬中西学术”为宗旨，而尤擅长于西式学问。其中，边疆研究最放异彩。1922 年创办的华西边疆研究学会（West China Border Research Society）及其会刊《华西边疆研究学会杂志》（*Journal of the West China Border Research Society*）在国际学术界享有盛誉。华大博物馆以“搜集中国西部出土古物、各种美术品，以及西南边疆民族文物，以供学生课余之参考，并做学术研究之材料”为目标，在美籍学者葛维汉（David Crockett Graham）的主持下，成为国内社会科学研究的另一基地。

华大社会科学研究的特点：一是具有较强的国际色彩，二是提倡跨学科的合作，三是注重实地踏勘；对边疆文化、底层文化和现实问题更为关注，与国立川大校内更注重“大传统”和经典研习的学术风格形成了鲜明对比。双方各有所长，其融合互补也成为 20 世纪三四十年代两校人文社会科学发展的趋向。从20 世纪 30 年代中期开始，华大一方面延请了庞俊、李植等蜀学传人主持中文系，加强了其国学研究的力量；另一方面致力于学术研究的中国化。一批既有现代社会科学的训练，又熟悉中国古典文化的中国学者如李安宅、郑德坤等成为新的学术领袖。

1935 年，任鸿隽就任国立四川大学校长后，积极推动现代科学的发展。1936 年 5 月，川大组建了西南社会科学调查研究处，在文科中首倡实地调研的风气，也代表了川大对西南区域跨学科综合性研究的发端。此后，经济学、社会学、民族学、考古学等领域的学者组织开展了大量的实地考察工作，掌握了西南地区社会文化的第一手资料。在历史学方面，较传统史学而言更注重问题导向和新材料之扩充的“新史学”也得到了蓬勃发展，并迅速成为国内史学界的重镇。20 世纪 30 年代后期开始，川大校内名师云集。张颐（哲学）、朱光潜（美学）、萧公权（政治学）、赵人儁（经济学）、徐中舒（历史学）、蒙文通（历史学）、赵少咸（语言学）、冯汉骥（考古学、人类学）、闻宥（民族学、语言学）、任乃强（民族学）、胡鉴民（民族学）、彭迪先（经济学）、缪钺（历史学）、叶麟（文艺心理学）、杨明照（古典文学）等一批大师级学者均在此设帐，有的更任教终身，为川大文科赢得了巨大声誉。

在不同学术流派的融合中，川大人文社会科学形成了自己的特点：一方面具有传统学术通观明变之长；另一方面又具有鲜明的现代学术意识。1952 年，在院系调整中，随着华大文科的并入，更使川大人文社会科学进入了飞速发展的新时期。半个多世纪以来，在继续保持传统优势学科如古典文学、语言学、历史学、考古学、民族学发展的基础上，新的学科如宗教学、理论经济学、敦

煌学、比较文学、城市史等也成长起来，涌现出了一大批在国内外学术界受到极高赞誉的学者，为川大文科未来的进一步发展奠定了良好的基础。

2006年是川大建校110周年，为了继续发扬深厚的学术传统，推动人文社会科学研究的新繁荣，学校决定设立“四川大学哲学社会科学学术著作出版基金”，资助川大学者尤其是中青年学者原创性学术精品的出版。我们希望通过这套丛书的出版，有助于川大学术大师的不断涌现和学术流派的逐渐形成，为建设具有中国特色、中国风格、中国气派的哲学社会科学作出贡献。

前　言

数据资源的分析、整合在宏观经济研究中起着越来越重要的作用。国民经济在其经济活动和运行过程中，积累了大量的数据，包括总供给与总需求、国民经济的总值及其增长速度、国民经济中的主要比例关系等。要想准确判断宏观经济的发展变化趋势，必然需要充分的信息支持才能得到正确的结论。数据挖掘作为一种系统地提取和分析大量数据的工具，能有效地从宏观经济不断积累与更新的数据库中提取隐含在其中的、人们事先不知道的、但是又有用的信息和知识。因此，数据挖掘被引入宏观经济研究领域，对大量的经济数据进行探索和分析，揭示其中的规律，它对宏观经济的评估、预测和预警具有重要意义。

全书共分 8 章。第 1 章介绍了数据挖掘和宏观经济的基础知识，同时简要介绍了数据挖掘方法在宏观经济中的应用，还探讨了宏观经济评估、预测与预警分析过程；第 2 章着重解决宏观经济中对数据的预处理问题，简单介绍了原始数据中可能存在的问题以及数据预处理的功能，并对常见的数据预处理技术和预测 GDP 数据预处理过程进行了详细的探讨；第 3 章讨论了分类算法，对分类算法的基本概念作了介绍，讲述了分类算法的应用，同时介绍了分类算法中的支持向量机分类和朴素贝叶斯分类方法的原理及其应用；第 4 章系统地介绍了数据挖掘的又一重要方法——聚类分析，阐述了聚类分析方法的概念及其应用，并详细介绍了聚类分析方法中的 Ward 聚类分析以及对聚类结果的评估；第 5 章首先对常用的宏观经济预测方法，包括 ARMA、ARIMA、ARCH、AC、GMDH 和线性组合预测模型以及对几种预测方法的应用情况进行了简单的介绍，其次阐述了如何使用软件实现各种预测方法的详细操作，并以中国宏观经济指标预测的案例说明预测方法；第 6 章论述了数据挖掘方法中的关联规则方法，简要介绍了关联规则的基本概念，特别介绍了经典算法——Apriori 算法的原理，还讨论了关联规则以及在宏观经济预警分析中的应用；第 7 章则从方法的概念、基本原理及其应用和具体的软件实现方面着重介绍了 DEA——数据包络分析法，并举例说明；第 8 章在介绍主成分分析方法概念

及基本原理的基础上，进一步对主成分分析方法的运用进行了简要介绍，结合实例详细地演示了用软件实现应用的方法。

本书凝聚了作者多年来的研究成果，深入浅出、通俗易懂、图文并茂，把相对复杂的数据挖掘技术及其在宏观经济中的应用简明扼要地呈现在读者面前。

在本书的编写过程中，作者参考了大量的专业书籍和相关研究文献，得到了杨小朋、田盼、许沛沛、吕一清、尹静、冯韵、叶学芳、张丽丽、郭秋艳、王迪、朱虹明、帅马恋、雷挺、刘玉婷、熊涛的帮助，以及四川大学商学院的大力支持。在此一并向他们表示衷心的感谢。

本书受到四川大学中央高校基本科研业务费研究专项项目（skcb201206）的资助，同时也受到教育部人文社会科学研究规划基金项目（11YJA630029）的资助，在此表示衷心感谢。

由于编者水平有限，编写时间仓促，书中可能存在错误，敬请读者批评指正，以便再版或重印时纠正。

编　者

2013 年 7 月

目　录

第 1 章　宏观经济数据挖掘基础知识

随着经济的发展，人们越来越认识到应用数据挖掘方法研究宏观经济问题的重要意义。本章介绍数据挖掘和宏观经济的基础知识，同时介绍数据挖掘方法在宏观经济中的应用，还将探讨宏观经济评估、预测与预警分析过程。

1.1　数据挖掘概述

近年来，数据挖掘引起了信息产业界的极大关注，其主要原因是迫切需要将大量存在的数据转换成有用的信息和知识，并进行广泛且适当的应用。本小节探讨了数据挖掘的定义、功能及研究方向，并重点介绍了数据挖掘的过程。

1.1.1　数据挖掘的定义

数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用信息和知识的过程[1]。数据挖掘就是从数据中挖掘知识。很多术语和数据挖掘意思相近，如数据分析、知识发现、数据融合以及决策支持等。在人工智能领域，习惯称其为知识发现，而在数据库领域习惯叫作数据挖掘。

数据挖掘的源数据可以是像关系数据库中的结构化数据；也可以是文本、图形、图像等半结构化数据。数据挖掘可以理解为一个利用各种分析方法和分析工具在海量数据中建立模型和发现数据间关系的过程，运用这些关系和模型可以支持决策和做出预测。

1.1.2　数据挖掘的功能

数据挖掘目标是从数据中发现隐含的、有意义的知识，还可以通过预测未来的趋势和行为，做出有前瞻性的、基于知识的决策。数据挖掘的具体功能有

以下几个方面[2]。

（1）概念描述

概念描述就是描述某类对象的内涵，同时概括这类对象的相关特征。概念描述分为两类：特征描述和区别性描述。特征描述用于描述某类对象的共同特征，而区别性描述用于不同类对象的区别。比如，高层用户可能不关心每口井每个月的钻井进尺，而更愿意观察高层数据，如按地区、井型对钻井分组，观察每组钻井的进尺。

（2）关联分析

关联数据是数据中存在能够被发现的重要指示。关联就是变量与变量间存在的某种联系。关联分析的目的就是找出数据中隐藏着的关联规则。关联分析能够发现隐藏的关联规则，这些关联规则有可能有效地支持用户进行决策。比如，经过关联分析，能够发现当钻井的设计井深在2165～2303米时，同时工期预计在38天以内，最后总费用有83.87%的概率在261.15万元以内的规则，这样的井占所有钻井的10%以上。发现了这些联系后，用户就能够及时调整资金保证钻探工作的顺利进行。

（3）分类与预测

分类，就是依照所要分析的对象的属性来分类、定义、建组。如客户关系管理中，时常将客户按照交易量分为忠诚度高、中、低三个类别，可以根据各个类别的特点制订营销方案。分类的关键是确定分类的标准或者规则。因此，首先需要根据属性特征，为每一种类别找到一个合理的描述或模型，再根据前面确定好的规则或者模型对新的数据进行分类。

预测，就是利用历史数据建立模型，以获得未来变化的趋势或者评估给定样本可能具有的属性值或者值的范围。如可以根据历史数据，预测公司在下一年可能产生的总进尺和总费用；也可以根据历史数据，预测公司在下一年可能用去的生产总时间与非生产时间，这对公司制订年度计划有着非常重要的指导意义。

（4）聚类分析

聚类分析又被称为无指导学习，其目的是客观地按照被处理对象的特征分类，将有相同特征的对象分为一类。

分类与聚类的区别是：分类需要预定义类别和预先训练样本，而聚类是直接从数据源获取数据，并没有预先定义好的类别和训练样本存在，所有的对象都根据彼此相似的程度进行归类。

聚类分析按照对象本身的相似性将其聚集在一起，然后对聚类结果进行分析解释。可以把钻井队按其所钻的进尺分为不同类别，这样就可以把有难度的工程分配给有经验的钻井队以提高施工的效率。

（5）趋势分析

趋势分析又被称为时间序列分析，是时序数据挖掘最基本的内容。运用趋势分析可以从相当长的时间发展中发现规律和趋势。趋势分析和关联分析相似，其目的也是为了找出数据之间的联系，但趋势分析的侧重点在于分析数据间的前因后果关系。

（6）孤立点分析

数据库中包含的一些与数据一般行为或者模型不一致的数据就是孤立点。孤立点分析又被称为孤立点挖掘。大部分的数据挖掘都将孤立点视为噪声或异常丢弃，但这些孤立点可能具有分析意义，如欺骗检测。

（7）偏差分析

偏差分析是对差异和极端特例的描述，又被称为比较分析，用于揭示事物偏离常规的异常现象，如标准外的特例、数据聚类外的离群值等。

寻找出偏差的数据并对其进行分析是很有意义的。偏差分析的基本方法是：寻找参照值与观测结果之间有意义的差别。偏差包括潜在的知识、分类结果中错误的实例、不满足某一规律的特例、实际观测结果与预测模型的预测值偏差等。

1.1.3　数据挖掘的过程

数据挖掘是一个完整的过程，从大量数据中挖掘先前未知的、有效的、可使用的信息，并使用这些信息做出决策或转换成有用的的知识。

数据挖掘的一般步骤[3]如图 1－1 所示。