

本书是在国内外时间序列数据挖掘最新研究的基础上，以股票时间序列数据为实际应用背景，对时间序列挖掘中的模式表示、相似度量、聚类分析、相似查询、频繁模式发现、异常检测等方面进行了研究分析，提出了一些算法和解决方案，取得了一定成果，具有重要的参考价值。

D 多变量时间序列研究

DUOBIANLIANG SHIJIAN XULIE YANJIU

周大镛 著

河北人民出版社

本书出版得到河北经贸大学出版基金资助

多变量时间序列研究

Multivariate Time Series Research

周大镛 著

河北人民出版社

图书在版编目 (CIP) 数据

多变量时间序列研究 / 周大镛著. — 石家庄: 河北人民出版社, 2012. 12

ISBN 978-7-202-06818-2

I. ①多… II. ①周… III. ①多变量-时间序列分析
IV. ① 0211.61

中国版本图书馆 CIP 数据核字 (2012) 第 183652 号

书 名 多变量时间序列研究

著 者 周大镛

责任编辑 王 颖

美术编辑 于艳红

责任校对 付敬华

出版发行 河北人民出版社 (石家庄市友谊北大街 330 号)

印 刷 石家庄市桥西区新科技印刷厂

开 本 850 毫米×1168 毫米

印 张 6

字 数 15 0000

版 次 2012 年 12 月第 1 版 2012 年 12 月第 1 次印刷

印 数 1-1 000

书 号 ISBN 978-7-202-06818-2/C·241

定 价 20.00 元

版权所有 翻印必究

内 容 简 介

时间序列是按时间顺序排列的、随着时间变化且相互关联的数据序列，在经济、金融、科学观测和工程等各个领域广泛存在。时间序列的数据挖掘研究是一种从大量时间序列数据中发掘有价值的规律性信息的算法和实现技术。目前重点的研究内容包括时间序列的模式表示、相似查询、聚类 and 分类、异常检测等。本书在国内外时间序列数据挖掘最新研究的基础上，以股票时间序列数据为实际应用背景，对时间序列挖掘中的模式表示、相似度量、聚类分析、相似查询、频繁模式发现、异常检测等方面进行了研究分析，提出了一些算法和解决方案，取得了一定成果。

前 言

多变量时间序列数据广泛存在于金融、多媒体、医学等领域，随着时间推移，这类数据的存储规模呈现爆炸式地增长。因此，如何有效地管理和利用这些数据，挖掘其内在的各种变化模式，是一个具有挑战性和重要意义的理论与实际应用课题。该课题的研究对于我们正确认识事物变化，进行科学决策，识别各种异常行为等具有重要的指导意义。

本书在分析多变量时间序列数据特点和实际应用需求的基础上，针对多变量时间序列的方法和技术进行了研究，反映了当前多变量时间序列研究的最新成果。

全书共分 7 章。第 1 章是绪论，介绍多变量时间序列研究的现状和意义。第 2 章讨论多变量时间序列模式的表示，提出了一种基于主成分分析的多变量时间序列模式的表示方法。在实际问题的研究中，为了全面、系统地分析问题，往往要考虑众多有关变量，该方法用为数较少的、互不相关的新变量来反映原变量所提供的大

部分信息，通过对新变量的分析达到解决问题的目的。第 3 章研究多变量时间序列聚类分析，提出一种高效的多变量时间序列聚类算法 PCA-CLUSTER。已有的聚类算法大多采用 k 均值对低维数据进行聚类，不能对高维多变量时间序列数据进行有效聚类。首先利用主成分分析对多变量时间序列数据降维，然后通过改进 K-means 算法对选取的多变量时间序列的主成分进行有效聚类分析。第 4 章是多变量时间序列相似性查询，提出了一种基于距离的索引结构相似查询算法 Dbis。为了高效地执行多变量时间序列相似查询，在 Eros 相似距离度量的基础上，选择聚类质心作为参考点，依据参考点将每类变换到一维空间，利用 B^+ -tree 结构进行索引查询，实验表明查询效率和准确性都有较大提高。第 5 章是多变量时间序列关联模式挖掘，通过 PCA 分析方法把时间序列分割成状态序列，通过建立 FP-tree 来发现时间序列的关联模式，并做了实验结果比较分析。实验结果证明了该算法的可扩展性和实用性。第 6 章是多变量时间序列异常检测，提出了一种基于密度的多变量时间序列异常检测算法。在多变量时间序列中与其他序列显著不同的序列，称为异常序列。在 k -近邻局部异常检测算法的基础上，结合基于主成分分析的降维方法，可以高效检测多变量时间序列的异常。第 7 章是对本书的一个总结和今后作者研究的方向。

本书内容新颖，认真总结了作者的科研成果，取材国内外最新资料，反映了当前该领域的研究水平。论述概念清晰，表达准确，突出理论联系实际，富有启发性。本书对从事数据挖掘、时间序列分析、机器学习等科技人员具有重要参考价值。

目 录

第一章 绪 论	(1)
第一节 时间序列研究的实际需求	(1)
一、研究背景	(1)
二、研究的实际需求	(4)
三、时间序列数据挖掘的研究任务	(6)
第二节 时间序列数据挖掘的研究现状	(7)
一、时间序列数据挖掘概述	(7)
二、时间序列模式表示	(9)
三、时间序列相似度量	(13)
四、时间序列索引	(14)
五、时间序列模式发现	(15)
六、时间序列预测与数据可视化	(16)
第三节 本书工作	(18)
一、研究内容	(18)
二、创新点	(20)
三、组织结构	(21)
第二章 多变量时间序列模式表示和相似度量	(23)

第一节 相关研究	(24)
一、时间序列定义	(25)
二、复杂类型数据挖掘问题	(28)
第二节 时间序列模式表示	(29)
一、时间序列的模式表示定义	(30)
二、时间序列模式表示的好处	(30)
三、时间序列表示方法应遵循的原则	(31)
四、时间序列主要模式表示方法	(32)
第三节 相似度量	(40)
一、Minkowski 距离	(41)
二、动态时间弯曲距离	(42)
三、扩展的 Frobenius 范数距离	(47)
第四节 基于序列重要点的时间序列分割算法	(50)
一、序列重要点定义	(51)
二、基于序列重要点分割算法 PLR_SIP 描述	(52)
三、实验结果评价	(54)
第五节 本章小结	(57)
第三章 多变量时间序列聚类分析	(59)
第一节 相关研究	(60)
一、聚类分析的概念	(61)
二、聚类的特征	(62)
三、聚类间的距离	(63)

四、聚类有效性	(64)
五、聚类过程	(65)
六、数据挖掘中现有的聚类算法	(66)
第二节 多变量时间序列聚类算法	(74)
一、多变量时间序列主成分分析	(75)
二、PCA-CLUSTER 聚类算法	(79)
第三节 实验及结果分析	(81)
一、实验数据集	(81)
二、算法性能评价	(82)
第四节 本章小结	(86)
第四章 多变量时间序列相似查询	(87)
第一节 研究基础	(88)
一、时间序列相似定义	(90)
二、时间序列的相似度量	(91)
三、时间序列相似的查询方式	(92)
四、相似性查询的策略	(93)
五、相似查询的完备性	(96)
第二节 多变量时间序列的相似查询	(99)
一、B ⁺ -tree 索引技术	(100)
二、多变量时间序列相似查询算法	(102)
第三节 实验及结果分析	(105)
一、实验数据及环境	(105)
二、算法性能评价	(105)
第四节 本章小结	(110)

第五章 多变量时间序列关联模式挖掘	(111)
第一节 研究基础	(113)
一、关联规则的意义	(114)
二、经典关联规则挖掘算法研究	(115)
第二节 多变量时间序列的关联模式挖掘算法	(122)
一、多变量时间序列的关联模式挖掘算法描述	(123)
二、多变量时间序列关联模式算法的实现	(124)
第三节 实验及结果分析	(126)
一、实验数据及环境	(126)
二、算法性能评价	(127)
第四节 本章小结	(127)
第六章 多变量时间序列异常检测	(129)
第一节 研究基础	(130)
一、异常的描述	(131)
二、异常定义	(135)
三、异常检测的理论基础	(138)
四、经典异常检测算法	(140)
第二节 多变量时间序列的异常检测算法	(146)
一、多变量时间序列异常检测步骤	(146)
二、多变量时间序列异常检测算法描述	

.....	(146)
第三节 实验及结果分析	(148)
一、实验数据	(148)
二、实验结果	(149)
第四节 本章小结	(150)
第七章 总结与展望	(152)
第一节 总 结	(152)
第二节 进一步工作	(154)
参考文献	(157)

第一章 绪 论

本章阐述了全书研究工作的背景和意义，介绍了时间序列数据挖掘的研究任务和研究现状。同时，简要介绍了全文的研究内容、创新点以及组织结构。

第一节 时间序列研究的实际需求

一、研究背景

随着计算机应用技术和大容量存储技术的迅猛发展，以及条形码等数据获取技术的广泛应用，在经济、科学、工业、教育、交通以及军事等社会的各个领域，人们积累了大量的各种类型的数据，互联网的飞速发展更是加快了数据增长的步伐，世界上的数据正以惊人的速度增长。但是这些海量数据绝大部分都没有被提炼为知识，“丰富的数据与贫乏的知识”问题也日益突出。不同领域的人们期待着从这些大量数据中分析出其间所蕴涵的有价值的知识，这需要有一个强有力的工具来对其进行分析，数据挖掘 (Data Mining, DM) 技术应运而生。数据挖掘又叫做数据库中的知识发现 (Knowledge Discovery in Database)，就是从大量无序的数据中发现隐

含的、有效的、有价值的、可理解的模式，进而发现有用的知识，并得出时间的趋向和关联，为用户提供问题求解层次的决策支持能力^[1]。

时间是自然界无所不在的客观属性，大多数信息都具有相应的时间属性，因此在这些保存的数据中，绝大部分都是呈现时间序列类型的数据。时间序列（Time Series）是按照时间先后顺序排列的多次观测记录所组成的数据集。在社会生活的各个领域中都大量广泛存在时间类型序列，如金融证券市场中每天的股票价格变化；商业零售行业中，某项商品每天的销售额；气象预报研究中，某一地区的每天气温与气压的读数；在生物医学中，某一症状病人在每个时刻的心跳变化等。图 1-1 给出了两种典型的时间序列数据：股票数据、心电图数据。

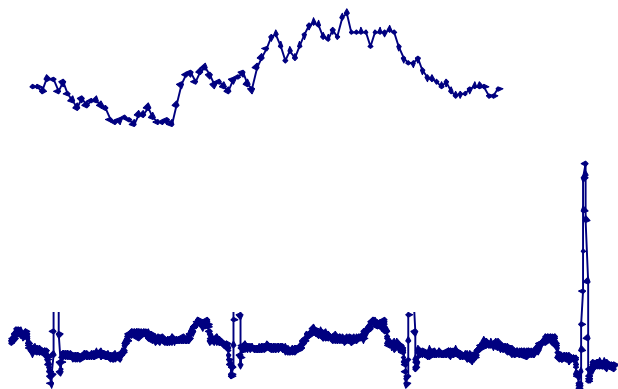


图 1-1 时间序列样例

时间序列是对历史事件的记录，展现事物变化的显示方式。随着时间推移和时间序列数据的大规模增长，人们需要更有效地管理海量时间序列数据库、更有效地发现海量时间序列数据中蕴含的规律（知识）以及更有效地理解复杂的时间序列数据，这对于我们揭示事物发展变化的内部规律，发现不同的事物之间的相互作用关系，为人们正确认识事物和科学决策提供依据等具有重要的实际意义。因此有关时间序列分析的研究一直以来就受到了许多研究人员的广泛重视，作为一个新兴的研究领域，在学术界受到广泛的关注，成为一个具有重要理论和实用价值的热点研究课题^[1-3]。

一般说来，时间序列分析可以分为传统方法和现在的数据挖掘方法。传统的时间序列分析方法作为概率统计学科的一个课题，主要关注的是序列的总体特征，通常是把序列匹配到某种整体模型上，然后通过分析模型来做描述、预测或是控制。经过几十年的发展，传统的时间序列分析取得了许多重要成果，在实际应用中发挥了重要作用。但是，面对新的数据库应用以及知识发现提出的新需求，传统的时间序列分析方法还是远远不能满足要求，存在着种种不足。

现在的数据挖掘方法关注的问题是针对序列的局部特征，分析不同数据源之间或者相同数据源在不同时间区间的相似性（或者差异性），从而从中提取关联规则，发现知识。传统的时间序列挖掘关注的都是数值型的时

序或者序列数据，而且多是连续值的。若按照数据来源或时间关系进行索引，这样的索引结构在基于相似查询场合，其效率是极其低下的，特别是对海量时间序列数据库，更是不可行的。另外，传统的时间序列分析方法缺乏对时间序列数据中蕴含的相似性分析，只能给用户提供数值结果，不能为用户提供个性化的关于时间序列形态的信息，人们通常更容易理解模糊化、形象化的描述。因此，新型的搜索系统需要为用户提供个性化的关于时间序列形态的信息。同时，对时间序列的数据挖掘也需要搜索系统能够提供关于时间序列形态的信息。

二、研究的实际需求

早期的时间序列分析技术对数据库应用领域 (如数据仓库以及知识发现等) 提出的新需求开始显得力不从心，因此从 1990 年代早期开始，时间序列数据挖掘 (Time Series Data Mining, TSDM) 作为一个新的研究领域应运而生，成为数据库知识发现 (KDD) 与数据挖掘领域的一个重要的分支。时间序列数据挖掘的定义：时间序列数据挖掘是时间序列数据库中知识挖掘的一个步骤，它发现序列数据中的时态模式或模型^[4]。时间序列数据通常属于高维、非结构化数据，并可能受到噪声污染。复杂的海量时间序列数据已经远远超出了人类的直观理解能力。因此，研究如何有效地从这些海量时间序列中自动抽取隐含的和潜在有用的重要信息，具有重要的理论价值和现实意义。在实际分析中，往往需要对时

间序列局部特征进行分析，比如发现频繁出现的变化模式；比较不同序列在某段时间内，其运动变化是否相似；监控时间序列是否在某一时刻产生异常等。这些发现型任务在许多应用领域中同样具有重要的意义。下面就是一些典型应用的例子：

◆电子商务网站每天都可能有上百万次的在线交易，生成大量有时间特性的记录文档，利用客户信息和海量交易数据，建立相应的数据挖掘算法和模型，从中发现了解客户的喜好和购买模式，设计出满足于不同客户群体需求的个性化网站，增强网站的竞争力。

◆在电子政务中，政府提供和保存着大量的信息，这些信息的有效处理显得特别重要。例如，政府发布的政府招标公告，规定了一个进行竞标的时间区间，超过时间的标书视为无效；政府制定的政策法规在不断的完善中，更新也比较快，这也要求电子政务系统可以自动地对失效的信息进行相应的处理。

◆在气象预报分析中，对于气压序列，需要找出在一段时间内那些频繁出现的变化模式，结合专家知识和经验，从中归纳气压变化的规律；或者给定一个气压序列和气温序列，要求找出这两个序列之间有那些是经常出现的相互关联的变化模式等。

◆在证券市场，每天的股票交易信息是一个典型的海量数据，众多投资者希望利用序列模式挖掘等相关挖掘技术从这些海量数据中发现某些规律，包括预测股票