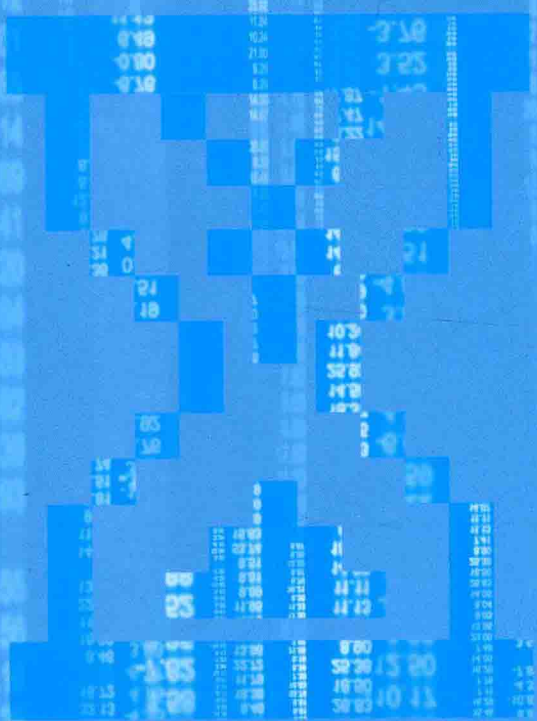
JINRONG SHIJIAN XULIE DE FENXI YU WAJUE

金融时间序列的分析与挖掘

吴学雁◎著



SPM 南方出版传媒
广东科技出版社 | 全国优秀出版社

JINRONG SHIJIAN XULIE DE
FENXI YU WAJUE

金融时间序列的分析与挖掘

SPM 南方出版传媒
广东科技出版社 | 全国优秀出版社
·广 州·

图书在版编目（CIP）数据

金融时间序列的分析与挖掘/吴学雁著. —广州：广东科技出版社，2018.7

ISBN 978-7-5359-7002-2

Ⅰ.①金… Ⅱ.①吴… Ⅲ.①金融—时间序列分析 Ⅳ.①F830

中国版本图书馆 CIP 数据核字（2018）第 165900 号

金融时间序列的分析与挖掘

Jinrong Shijian Xulie De Fenxi Yu Wajue

责任编辑：夏　丰　黄子丰

封面设计：友间文化

责任校对：罗美玲

责任印制：彭海波

出版发行：广东科技出版社

（广州市环市东路水荫路 11 号　邮政编码：510075）

http：//www.gdstp.com.cn

E-mail：gdkjyxb@gdstp.com.cn（营销）

E-mail：gdkjzbb@gdstp.com.cn（编务室）

经　　销：广东新华发行集团股份有限公司

印　　刷：佛山市浩文彩色印刷有限公司

（南海区狮山科技工业园 A 区　邮政编码：528225）

规　　格：787 mm×1 092 mm　1/16　印张 9.25　字数 185 千

版　　次：2018 年 7 月第 1 版

2018 年 7 月第 1 次印刷

定　　价：40.00 元

前　言

在金融领域中，时间序列是非常重要的一种数据类型，例如证券市场中的股票价格和交易量、外汇市场上的汇率、期货和黄金的交易价格等，这些数据都形成了持续不断的时间序列。金融市场中的时间序列主要使用基础分析和技术分析方法进行分析，这两种方法使用简单，但是无法对数据中隐含的更深层次的规律和特征进行挖掘。数理统计分析方法是目前金融时间序列分析中比较常用的方法，随着数据量的不断增加，这种方法的分析能力存在一定的缺陷，各种数理统计分析方法都无法有效地处理较大规模的数据集，也不适合从大量数据中主动地发现各种潜在的规则。因此面对金融行业不断涌现的海量数据，需要寻找新的数据分析和挖掘的方法。本书将数据挖掘技术运用到金融时间序列研究中，使用关联规则、聚类分析等数据挖掘方法对金融时间序列中的隐含模式进行挖掘，本文的创新点主要基于以下几个方面：

（1）针对金融时间序列需要保留形态特征与趋势特征的特点，提出了适合金融时间序列的多层次极值点分段表示法（MEPS），此方法能在多个层次上最大限度地保留关键特征点信息，从而能更好地捕捉和表示时间序列的形态和走势。

（2）针对金融时间序列需要保留形态特征与趋势特征的特点，在 MEPS 算法的基础上提出了分层的动态时间弯曲相似性度量方法（HDTW）及其改进方法 IHDTW，将时间序列在不同层次上进行分段，然后计算对应分段层次中子序列间的相似性，最后汇总得到序列间的相似度，在算法中对动态时间弯曲算法（DTW）进行了改进，并且考虑到了分层的均匀因素及趋势因素，实验结果证明能大大提高相似性度量的效果和效率。

（3）金融市场的运行非常复杂，其中人的因素也非常重要，为了在金融时间序列挖掘的过程中更好地体现用户的实际需求，提出了基于事件的时间序列相似性度量方法（SMBE），此算法通过对事件的定义引入用户在相似性度量时的偏好与需求，并设计了基于 SMBE 的层次聚类算法，完全以事件的相似性为中心进行聚类，定义了类间相似度和类间一般距离两个参数，并以它们之间的比较作为判断类间距离的依据，使得时间序列相似性度量及其聚类的结果更加符合实际金融市场的状态与需求。

（4）多元时间序列的跨事务关联规则挖掘可以发现不同时间序列在不同时间段发生的事件之间的联系，对于准确预测金融时间序列的走势具有非常重要的意义。本书提出 O－Apriori 算法实现了多元时间序列的跨事务关联规则挖掘，设计并定义了频繁状态矩阵来存放项集的频繁状态，构建事务布尔矩阵来存放事务与项集的关系，只需在初始化阶段扫描一次数据库产生初始的频繁状态矩阵和事务布尔矩阵，并根据多元时间序列的跨事务关联规则的定义对递推寻找频繁 k 项集的过程进行约简，大大提高了关联规则挖掘的效率。然后提出了基于可变支持度的 O－Apriori 算法（VSO－Apriori），设置变化的最小支持度阈值来对应不同级别的频繁项集，从而能够挖掘出更多有效的关联规则。

（5）随着金融市场的快速发展，对于实时金融时间序列数据流的处理与分析也成为非常迫切的需求。针对实时金融时间序列数据流挖掘的需求，本书提出了基于滑动挖掘区间的动态关联规则挖掘方法（SI－DARM）和基于形态特征的数据流聚类算法。SI－DARM 算法能在线对不同的挖掘区间进行频繁项集的挖掘，并能跟踪及表现频繁项集的模式变化趋势。基于形态特征的数据流聚类算法通过在线保留子序列的特征信息来实现任意时刻的聚类需求，可以随着新的数据流的到来追踪聚类结果的演化过程。实验证明，这两种算法都能非常有效地进行金融时间序列数据流的动态挖掘，对于挖掘与发现不断变化的金融市场的隐含规律具有非常重要的意义。

本书的基础资料和主要内容来源于作者攻读博士学位期间的研究，同时也与作者主持和参与的科研项目相关，其中主要包括：国家自然科学基金项目“经理级 CIO 拼创行为对 IS 战略匹配的作用机理研究：以复杂适应系统理论为视角”（编号 71472051）和教育部人文社会科学研究项目“社会化媒体环境下顾客契合行为研究”（编号 13YJCZH200）。

目　　录

第一章　引　　言

第一节　金融市场信息化的发展

现代金融业已经成为一国社会经济发展的重要推动力和国家竞争力的重要组成部分。在全球经济一体化的今天，金融业面临着巨大的机遇与挑战，对金融市场本质规律的认识和把握更直接关系到金融市场的稳定、高效与安全，探求金融市场的变化规律，提高金融管理与投资的效率也成为各国政府与投资机构孜孜以求的目标之一。

当前金融创新已成为金融企业的核心竞争力，而95%的金融创新的实现都极度依赖于信息技术，由信息技术构建的金融工具，能实时对金融风险进行识别、度量和控制，提高金融管理的决策效率。金融管理决策正在向智能化方向发展，金融信息化正成为当前金融业面临的重要课题。金融信息化对金融市场乃至整个经济社会具有非常重要的影响作用。

金融信息化[1]是指构建在通信网络、计算机、信息资源与人力资源等要素基础上的国家信息结构框架，由具有统一技术标准，能以不同速率传送数据、语言和视频影像的综合信息网络构成，将具备智能交换和增值服务的多种金融信息网络系统联网结体，创造金融经营、管理、服务新模式，以人为本，全方位地服务社会，并极大提升金融企业核心竞争力的现代化工程。目前，我国金融信息化事业已经取得了显著成绩，基于服务、经营、管理和监管理念的金融信息化技术体系框架基本形成，高效、安全的金融信息化安全保障体系初步建成，为推动金融改革与创新，提升中国金融业的核心竞争力做出了突出贡献。金融信息化已成为中国金融平稳安全运营最基本的生存支撑环境，信息化已经提升到金融业的战略高度。

金融业的信息化进程可以概括为：以数据大集中为前提，以完善的综合业务系统为基础平台，以数据仓库为工具，以信息安全为技术保障，打造出现代化、网络化的金融企业。在金融业信息化的进程中产生了大量的数据，尤其是网络金融业的发展更是产生了海量的数据。金融机构的许多业务活动（客户分析、投资决策、风险管理、价格预测等）越来越依赖于对大量历史数据的分析，我国的投资者与金融机构也越来越清楚地认识到分析金融数据、从中挖掘出有价值的信息是实现科学化管理决策的必要手段与核心工作。

第二节　金融市场的传统分析方法

金融市场是一个庞大的系统，受到各方面因素的多重影响，具有非常复杂的运动规律，而时间序列数据则是其综合外在的表现形式。由于金融市场中的数据以时间序列为主，因此金融市场分析常常被称为金融时间序列分析。

一、基础分析与技术分析

基础分析与技术分析主要用于证券市场中的时间序列分析。基础分析主要用于股票长期走势的分析，常常用来判断是否对某只股票进行投资。技术分析主要用于股票短期走势的分析，常常用来对股票的买卖时机进行判断。

基础分析是指通过分析影响证券市场供求关系的相关因素来确定股票的真正价值，判断股票市场的走势，从而为投资者提供选择股票的依据。基础分析基于因果关系论的观点，具有很强的逻辑性，但是在实践中却很难操作。影响市场变化的因素千变万化，比如宏观经济形势、公司的运营状况、国家的金融政策等都可能影响股票市场的行为，因此基础分析者需要具有获取完备且及时信息的能力，显然这是非常困难的。另外，准确度量各种市场因素影响的程度也是非常困难的，因此基础分析主要是基于定性的层面来进行描述。

技术分析是指完全根据市场行情的变化来进行分析，主要是通过图表和各种技术指标来判断股票未来的价格变化趋势，常用的指标包括 OBV 指标、BOLL 指标和 KDJ 指标等，这些指标都是通过对历史时间序列数据进行简单统计而得到的，比如收盘价格、开盘价格、交易量和涨跌指数等。

二、数理统计分析

数理统计方法是目前金融时间序列分析的一类主要方法，运用各种数理统计模型来进一步分析数据中更加复杂的规律及各种统计特征。

（1）统计特征的检验分析。

检验分析方法主要用于分析序列中是否存在某种统计特征，这些统计特征被用来对数据进行进一步的分析。

（2）相关分析。

通过计算序列间的相关系数来分析变量间的相关程度，例如可以通过计算相关系数来判断股价的变化幅度对股票交易活跃程度的影响。

（3）回归分析。

通过对变量进行拟合来分析变量间的依赖关系，例如对股票价格和交易量进行曲线拟合可以分析它们之间的依赖关系。回归分析通常用于分析序列的趋势及趋势的变动。

(4) 基于自回归移动平均模型的分析。

自回归移动平均模型 (ARMA) 是一类重要的时间序列分析模型，对时间序列建立合适的 ARMA 模型有两个主要的用途：一是分析序列的某些重要特征；二是进行短期预测。

(5) 金融资产的定价分析与投资组合分析。

使用已有的数学工具，例如定价模型等，确定各种金融资产的可能价格与风险，并结合优化算法构建合理的投资组合，降低投资风险，提高期望收益率。

第三节 数据挖掘技术的兴起与发展

基础分析和技术分析方法主要对证券市场中的时间序列进行分析，这两种方法使用简单，但是无法对数据中隐含的更深层次的规律和特征进行挖掘。数理统计分析方法是目前金融时间序列分析中比较常用的方法，但是，面对数据量的不断增加的状况，这种方法的分析能力存在一定的缺陷。各种数理统计分析方法都无法有效地处理较大规模的数据集，也不适合从大量数据中主动地发现各种潜在的规则。因此面对金融行业不断涌现的海量数据，需要寻找新的数据分析和挖掘的方法。

随着数据库技术、人工智能技术和并行计算等技术的发展与融合，数据挖掘技术应运而生。数据挖掘技术是一门新兴的交叉学科，是现代科学技术相互渗透的结果，其基本目标就是从大量的数据中提取隐含的、潜在的和有用的知识和信息，为金融时间序列的隐含模式挖掘提供了有效的途径。

1989 年 8 月在底特律召开了第一届知识发现 (Knowledge Discovery in Database, KDD) 国际学术会议，正式提出了数据库中的 KDD 的概念。随后在 1991 年、1993 年和 1994 年都举行了 KDD 的专题讨论会，来自各个领域的研究人员在一起讨论海量数据的分析算法、知识表达和知识运用等问题。KDD 这一概念不断被大家所认可和熟知，KDD 国际会议逐渐发展成为年会。

数据挖掘 (Data Mining, DM) 是知识发现的核心，正如 Han[2] 所说，在产业界、媒体和数据库研究界，DM 比 KDD 更为流行。从 1989 年至今，数据挖掘的定义也在不断地改进与完善。本文采用文献 [3] 中的广义的数据挖掘定义：数据挖掘是从数据集中识别有效的、新颖的、潜在有用的，以及最终可理解的模式的高级处理过程，它包括数据清理、数据集成、数据选择与变换、数据挖掘、模式评估与知识表达等过程，应用各种方法从数据序列中发现隐含的规律和模式。图 1 - 1 展示了广义的数据挖掘的过程，它包括四个主要的步骤：

(1) 问题/任务定义。

数据挖掘的目的是为了从大量的数据中发现令人感兴趣的、有用的信息。在数据挖掘的第一阶段，首先要定义数据挖掘要解决的商业问题，即通过数据挖掘过程发现何种知识。在问题定义的过程中，数据挖掘人员需要和领域专家及最终用户紧密协作，明确实际商业问题对数据挖掘的要求，也需要对各种学习算法进行对比后确定可行的学习算

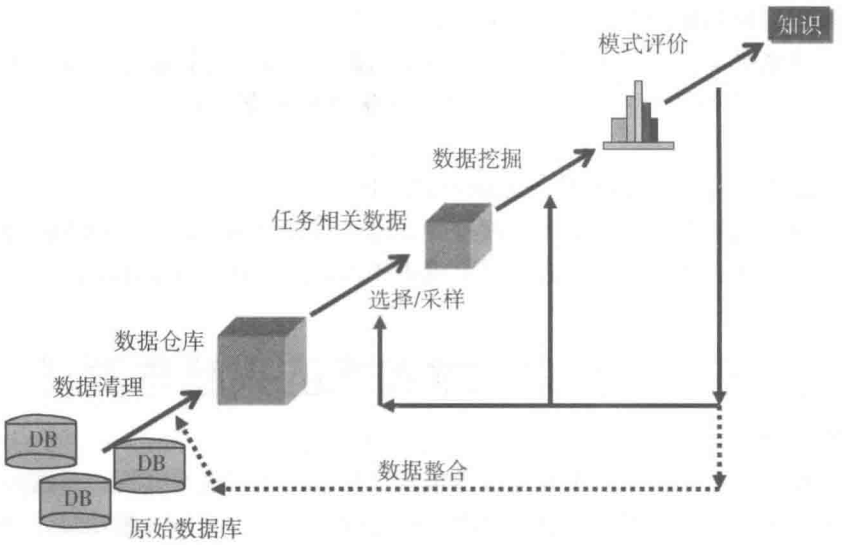


图 1－1　数据挖掘的过程

法，在后续的数据选取过程中也要根据定义的商业问题来选择相关数据。

（2）数据收集与数据预处理。

从原始数据库中选取数据进行集中，经过数据清理过程后形成数据仓库，从数据仓库中根据所定义的商业问题选取相关数据，一般可分为数据选取、数据预处理和数据变换三个步骤。数据选取是根据用户需要从原始数据库中抽取数据的过程。数据预处理过程一般包括消除噪声、消除重复记录、填补缺失值等，一般在数据仓库建立时完成。数据变换主要是对数据进行降维操作，即找出数据最本质的特征以减少数据挖掘过程中考虑的特征或变量数目。

（3）数据挖掘。

在数据挖掘阶段，首先根据定义的商业问题明确数据挖掘的任务，例如数据分类、数据聚类、关联规则挖掘或者异常检测等，然后根据数据挖掘任务选择相应的数据挖掘算法来完成数据挖掘的过程。数据挖掘算法的选择主要考虑两点：一是进行数据挖掘的数据的特点，例如金融数据需要选择可以保留数据点趋势和形态特征的算法；二是根据用户对挖掘结果的需求，有的用户希望获取容易理解的描述型知识，而有的用户希望获得预测型知识。

（4）结果的解释与评价。

数据挖掘得到的模式不一定是有效的，可能存在冗余或者无关的模式，因此需要对无效的模式进行删除。如果挖掘出的模式不符合用户的需求，还需要返回到上一个阶段重新进行数据挖掘的过程，直到找到符合要求的模式为止。另外，对于数据挖掘的结果要进行解释，转换为用户容易理解的方法。

数据挖掘的潜在应用领域非常广泛，从政府管理决策到商业营销策划，从科学问题研究到工业企业的生产流程优化等都可以找到数据挖掘技术的用武之地。如今数据挖掘技术的应用已经在许多领域取得成功，例如客户关系管理、产品质量分析、Internet 站

点访问模式和基因工程研究等。许多金融机构也认识到充分利用数据挖掘技术进行科学化管理决策和创新业务的重要性。美洲银行采用数据挖掘工具分析信用卡交易数据，全面了解客户的消费习惯和偏好，为其提供个性化的服务。HNC 公司的标志产品 Falcon 是信用卡侦测的专用软件，它利用嵌入式神经网络模型来预测信用和借贷欺诈，获得了巨大的商业成功。Customer Analytics 公司将预测模型融入金融服务业的客户关系管理软件中来计算顾客购买信托基金的倾向度。另外，根据软件供应商提供的报告，美国银行、FCC 国家银行、联邦住房贷款抵押公司、Mellon 银行和美国联邦储蓄银行等等一大批著名的金融机构都在进行某种形式的数据挖掘，但往往出于商业竞争的需要而隐瞒。

与国外相比，我国金融行业的发展相对缓慢，20 世纪 90 年代初才建立了上海与深圳证券交易市场，基金、期货等业务也是最近才发展起来，但是金融行业信息化的发展却非常迅速。目前，各个金融机构都基本建立了集中式的数据中心，金融业已经成为一个数据高度密集的行业。各大银行、交易所、证券公司、投资公司以及基金公司都保存了大量的历史交易数据，这些数据为数据挖掘的应用提供了很好的数据基础。然而虽然国内的金融行业已经进入了数据密集时期，但是与国外相比，在对数据进行深层次的挖掘和利用方面还处于较低的水平。因此利用数据挖掘技术对金融市场中的海量数据进行深入的分析与挖掘已成为当前国内金融行业的迫切需要。

第四节 本书的研究目的与内容

数据挖掘技术可以运用到金融领域的很多方面。在金融领域，时间序列是非常重要的一种数据类型，例如证券市场中的股票价格和交易量、外汇市场上的汇率、期货和黄金的交易价格等，这些数据都形成了持续不断的时间序列。本书所述的工作主要针对金融时间序列进行研究，使用数据挖掘方法对金融时间序列中的隐含信息进行挖掘，研究目的主要基于以下几个方面：

（1）满足对金融数据更深层次挖掘和主动挖掘的需求。传统的金融时间序列分析方法不能对金融数据进行较深层次的挖掘，尤其是不能主动地寻找隐含的信息与规律，通过数据挖掘技术能够较好地解决这一问题。

（2）结合金融时间序列的特征进行分析和挖掘，提高挖掘的准确度和效果。已有的时间序列挖掘算法的适用性是有限的，不能兼顾所有的时间序列类型和应用。文献[4]指出，对同一种挖掘方法而言，随着测试数据集的不同，方法的有效性、准确性和效率会出现完全不同的结果，有些挖掘方法对某些数据集甚至无法使用。因此，针对金融时间序列的特征和特定需求来进行数据分析与挖掘是非常重要的。

（3）针对金融时间序列实时性的需求，研究在线的动态数据挖掘算法。随着金融行业的数据量和业务量不断增加，对金融数据分析与处理的实时性需求也越来越高，因此研究动态的数据挖掘方法，对金融时间序列进行实时的分析与预测具有非常重要的意义。

一、本书的研究对象

本书的研究对象是金融时间序列。时间序列数据是一类特殊的序列数据，它是由一系列随时间变化的值组成的，测量这些值的时间间隔可以是等间距或者不等间距的。时间序列数据可以是连续的或者离散的，本书主要研究离散的时间序列数据。

定义1-1（离散时间序列数据）：对过程中的某个变量 s 进行观察测量，在一系列时刻 t_1，t_2，…，t_n（t_i 为自变量，且 $t_1 < t_2 < \cdots < t_n$）得到的离散有序数集合为 $\{(s_{t_1}, t_1), (s_{t_2}, t_2), \cdots, (s_{t_n}, t_n)\}$，称为离散时间序列数据，简称为时间序列。在时间序列分析领域，一般都使用 t_i 表示数据出现的先后顺序，通常令 $t_1 = 1$，$t_{i+1} = t_i + 1$，时间序列可记为 $\{s_1, s_2, \cdots, s_n\}$。

定义1-2（金融资产的增幅）：在金融时间序列分析过程中，金融资产的增幅是投资者重要的关注点之一，在本文中会多次使用这个概念。若某项金融资产在时刻 t_1 和 t_2 的价格分别为 p_1 和 p_2，则在 t_1 与 t_2 期间该金融资产的增幅 r 定义见（1-1）。

$$r = \frac{(p_2 - p_1)}{p_1} \tag{1-1}$$

二、本书的研究内容

本书结合金融时间序列的特征与金融分析过程的需求深入研究了金融时间序列模式挖掘的相关方法，研究内容主要包括以下几个方面：

（1）金融市场是一个由社会、经济、心理等诸多因素综合作用的复杂系统，作为其外在表现的金融时间序列具有不同于其他一般时间序列的特点。在很多情况下，金融时间序列的形态和趋势都蕴含了大量的信息，而这些信息可以通过若干关键点来表示。然而很多时间序列分段和表示的方法都破坏了序列的形态，或者平滑掉了关键点。因此，针对金融时间序列的这一特点，本书提出了多层次极值点分段表示法（MEPS），在多个分段层次上保留对应的关键点（特征点）信息，从而能充分捕捉和表示时间序列的形态和趋势。

（2）在金融时间序列中不可能频繁出现形态完全一样的两条序列（子序列），但是具有相似形态的序列（子序列）却是非常常见的。针对金融时间序列的这一特点，本书首先提出了基于分层的动态时间弯曲相似性度量方法（HDTW），在不同层次上采用动态时间弯曲方法（DTW）对序列间的相似性进行度量。然后在HDTW算法的基础上进一步提出了改进方法（IHDTW），一方面对DTW方法进行了改进，另一方面在相似性度量的过程中考虑了趋势因素和均匀分布因素，进一步提高了相似性度量的准确性和效果。另外，用户的实际需求也是相似性度量过程中一个非常重要的因素，本书通过对事件的定义来抽象用户的实际需求，并提出了基于事件的时间序列相似性度量方法（SMBE），使得时间序列相似性度量的结果更加符合用户的实际应用需求。

（3）多元时间序列的跨事务关联规则挖掘可以发现不同时间序列在不同时间段发

生的事件之间的联系，对于准确预测金融时间序列的走势具有非常重要的意义。本书提出的 O - Apriori 算法（VSO - Apriori），构建频繁状态矩阵来存放项集的频繁状态，构建事务布尔矩阵来存放事务与项集的关系，只需在初始化阶段扫描一次数据库产生初始的频繁状态矩阵和事务布尔矩阵，并根据多元时间序列的跨事务关联规则的定义对递推寻找频繁 k 项集的过程进行约简，大大提高了关联规则挖掘的效率。然后提出了基于可变支持度的 O - Apriori 算法，设置变化的最小支持度阈值来对应不同级别的频繁项集，从而能够挖掘出更多有效的关联规则。最后提出了一种多元时间序列跨事务动态关联规则挖掘的方法（SI - DARM），采用滑动时间窗口的思想对关联规则的挖掘区间进行划分，对每个挖掘区间进行频繁项集的寻找和关联规则的挖掘，该算法能对多条实时的时间序列数据流进行动态关联规则挖掘，并能跟踪及表现频繁项集在不同挖掘区间的变化趋势。

（4）聚类分析也是金融时间序列挖掘中非常重要的一项内容，通常用来为其他的数据挖掘任务提供先期的分类结果。结合（2）中提出的时间序列相似性度量方法，本书首先提出了一种基于 IHDTW 的聚类算法，采用共享最近邻相似度（SNN）的思想构建序列间的相似性计数矩阵，利用相似性计数矩阵来寻找聚类中心序列，获得了较好的聚类效果。然后，提出了一种基于 SMBE 的层次聚类算法，该方法专门针对满足用户需求的事件相似性进行聚类计算，采用类间相似度和类间一般距离两个参数的比较作为判断类间距离的依据，大大提升了聚类的效果。最后提出了一种基于形态特征的多时间序列数据流的实时聚类算法，针对金融时间序列的特点，在数据概要设计中通过保留重要特征点的方式进行子序列特征信息的提取，在新的数据到来时采用动态滑动窗口设计保证了多条数据流之间的数据同步，该算法可以实现任意时刻的数据流聚类，并且能够实时追踪聚类结果的演化过程。

（5）综合运用（3）中提出的金融时间序列的关联规则挖掘算法和（4）中提出的金融时间序列的聚类算法，提出了金融时间序列的综合预测方法，并以实际的沪深 A 股交易市场的数据为例对预测方法进行了验证，该方法可以对 3 个交易日内股票价格的变化区间和 60 个交易日内股票价格变化的趋势进行比较准确的分析与预测。

第二章　时间序列数据挖掘研究及其应用

在时间序列挖掘过程中，首先需要解决两个关键的问题：一是时间序列的表示；二是如何对时间序列与时间序列之间以及时间序列的不同子序列之间的相似性进行度量。在这两个基础问题之上，可以进行时间序列数据挖掘的各种任务，包括：时间序列的分类与聚类、时间序列的关联规则挖掘与预测等。本章主要针对与本书研究内容相关的几个领域，结合现有的时间序列数据挖掘的研究进行综述，具体包括：时间序列的分段与表示方法、时间序列的相似性度量方法、时间序列的关联规则挖掘算法以及时间序列的聚类算法。

第一节　时间序列的分段与表示

把时间序列通过其他的方式进行表示通常是为了对原始数据进行维度约简。进行维度约简最直接的方式就是采样，设置采样率$\frac{m}{n}$，m 是时间序列的长度，n 是约简后的维数。直接使用采样点来表示时间序列的方法非常简单直观，但是，如果采样率太低，被采样的时间序列的形状容易被歪曲。因此有大量的学者针对“如何更好地表示时间序列”这一命题展开了研究。从整体上来看，时间序列的表示方法可以分为两大类：一是基于时域的表示方法，二是基于变换域的表示方法。

一、基于时域的分段与表示

1．使用分段的特征值来表示序列

Keogh 等提出了分段累积近似方法（Piecewise Aggregate Approximation，PAA），该方法将时间序列分成等长的 k 段，并且使用每一段的平均值作为该段的 k 维特征矢量。这种方法的结果可以用任意的 L_p 范数来表示，算法非常快速并易于完成，但是该算法是等长分段的，并且平滑掉了时间序列的局部特征，造成了信息的遗漏和错误。为了修正等长分段带来的缺陷，Keogh 进一步提出了适应性分段常数近似法（Adaptive Piecewise Constant Approximation，APCA），该算法使用不同长度的分段来适应序列的形态，比 PAA 有更好的处理能力。除了使用分段的均值来表示序列的分段之外，也有一些其他的算法使用其他的分段特征值来表示序列，例如 Bagnall 等[5] 使用分段的平均值作为比较参数，将每个序列数据点转换为位的形式。

2．使用分段的近似曲线来表示序列

使用分段的近似曲线来表示序列主要有两种方法，一种是线性插值，一种是线性回

归。Keogh 提出了一系列线性插值的方法。首先在 1997 年提出了分段线性表示法（Piecewise Linear Representation，PLR）。1998 年，Keogh 和 Pazzani 通过考虑分段的权重加强了 PLR 算法。1999 年，他们又在 PLR 算法中加入了用户的相关反馈。分段线性表示法具有很好的形态表示与分割能力，并且分段在形态上相对独立，可以对同一时间序列在不同分辨率下进行观察，具有较高的噪声过滤和数据抽象能力。文献［6］～［7］中均使用了该方法。

3．使用重要特征点来表示序列

通过序列中的重要点来表示序列，约简序列的维度也是一种非常有效的方法。Chung 等首先提出了重要感知点（Perceptually Important Points，PIP）的概念并且在股票数据的技术模式匹配中进行应用。Perng 等使用界标模型来识别时间序列中的重要点进行相似性度量。Man 和 Wong 提出使用栅结构来表示时间序列中的高峰和低谷，这些点被称为控制点。Pratt 和 Fink 定义了极值点，并且选择某些极值点来表示时间序列而删除掉其他的点、这种方法过滤了较小的波动数据，保留了主要的极大值点与极小值点。

4．使用符号化来表示序列

分段符号化的方法首先将时间序列分成若干段，然后对每一分段使用特征符号来表示。

Lin 等[8]提出了符号累积近似（Symbolic Aggregate Approximation，SAX）的方法，该算法在 PAA 算法的结果上进行转换，将序列表示为一个符号串。由于 SAX 是基于 PAA 的符号化表示方法，SAX 也就继承了 PAA 的缺陷，即 SAX 只能对时间序列段的均值进行符号表示，容易忽略时间序列形态变化和关键点的重要信息。针对这种情况，钟清流等[9]同时考虑均值和方差并且将它们转换为符号，实现二维空间下的符号化表示。Megalooikonomous 等提出使用码字来表示时间序列，这些码字来源于关键序列形成的码字表，这一工作在 2005 年被进一步扩展为基于多层次的方法。Morchen 和 Ultsch 提出一种无监督的离散化过程，该算法使用质量评分和持续状态作为离散化的依据，并在算法中考虑了时间因素。特别地，Wang 等使用矢量量化（Vector Quantization，VQ）来对时间序列进行符号化表示。它先将训练集中时间序列进行等长度分段，再利用传统的 VQ 算法对它们进行矢量量化表示，形成用符号表示的编码表。

另外，文献［10］中采用了符号化的方法来表示序列。

二、基于变换域的分段与表示

1．离散傅立叶变换

离散傅立叶变换（Discrete Fourier Transform，DFT）也是一种时间序列表示与描述的常用方法。该方法最早由 Agrawal 等提出，Rafiei 和 Mendelzon 不断扩展了使用 DFT 进行相似性查询的方法。Janacek 等也是在离散傅立叶变换的频域内来进行时间序列的相似性度量。

该算法计算简便，能把时间序列信号的大部分能量集中到很少的几个系数中，时间复杂度也比较低。但是该算法在数据截取过程中舍弃了信号的高频成分，平滑了信号的