解兵博士文丛
Hubing Doctor Thesis Collection

基于分形技术的金融数据分析方法研究

RESEARCH ON FINANCIAL DATA ANALYTICAL METHOD BASED ON FRACTAL TECHNOLOGY

倪丽萍/著　倪志伟/导师

合肥工业大学出版社
HEFEI UNIVERSITY OF TECHNOLOGY PRESS

基于分形技术的金融数据分析方法研究

倪丽萍　著　　导师　倪志伟

合肥工业大学出版社

图书在版编目(CIP)数据

基于分形技术的金融数据分析方法研究/倪丽萍著．—合肥：合肥工业大学出版社，2015.12

ISBN 978-7-5650-2570-9

Ⅰ.①基… Ⅱ.①倪… Ⅲ.①金融—数据处理—研究 Ⅳ.①F830.41

中国版本图书馆 CIP 数据核字(2015)第 293019 号

基于分形技术的金融数据分析方法研究

倪丽萍 著　　倪志伟 导师　　责任编辑 权 怡

出 版	合肥工业大学出版社	版 次	2015 年 12 月第 1 版
地 址	合肥市屯溪路 193 号	印 次	2016 年 4 月第 1 次印刷
邮 编	230009	开 本	710 毫米×1010 毫米 1/16
电 话	编校中心：0551-62903210	印 张	7.75
	市场营销部：0551-62903198	字 数	102 千字
网 址	www.hfutpress.com.cn	印 刷	安徽联众印刷有限公司
E-mail	hfutpress@163.com	发 行	全国新华书店

ISBN 978-7-5650-2570-9　　定价：26.00 元

摘　要

由于金融业信息化建设的快速发展，金融数据量不断增多，如何对这些数据进行有效的分析是研究的热点问题。近年来，人们根据金融数据动态、复杂、非线性的特点，引入了非线性理论，以期更加准确地根据这些数据揭示金融市场的运作规律。

本书围绕金融数据分析领域中的热点和难点问题，对基于分形技术的金融数据分析方法进行研究。根据金融数据的特点，研究了金融一元、多元时间序列分形维数的定义、计算方法和意义。在此基础上，将分形维数与数据挖掘算法相结合用于分析金融数据，解决关键问题——相似性分析、维数约简以及预测等。

本书的主要内容如下：

(1)论述了金融数据的研究背景和意义，介绍了分形理论的发展概况，总结了运用分形技术分析金融数据的原理和方法。

(2)介绍了金融时间序列中常用的分形维数计算方法，并探讨了维数求解的后期过程中数据的拟合方法。本书分别运用最小二乘法和最小二乘分段方法对数据进行拟合，相关的实验结果表明最小二乘分段拟合方法能够使数据的拟合性能得到提高，进而提高维数计算的准确率。

(3)提出了一种趋势分形维数的定义和计算方法，以更好地表征金融时间序列的波动特征。该维数分为阴线维和阳线维两种。通过实验对股票数据、汇率数据和期货数据进行研究，发现阳线维或阴线维相对于传统的分形维数能够更好地反映金融市场的情况。

(4)研究了金融时间序列中的相似性分析方法。提出了将趋势分形维数和 K-means 聚类算法相结合的相似性分析方法，并对股指序列进行了相似性聚类研究。运用该方法时，首先利用趋势分形维数表示时间序列，进而

利用K-means聚类算法对表示后的序列进行聚类。通过与利用传统分形维数表示的聚类结果相比较，利用趋势分形维数表示的聚类结果更加准确，说明趋势分形维数比传统分形维数更能准确地和更细致地描述聚类结果。这也进一步表明了趋势分形维数的重要意义和作用。

(5)分析和比较了多元时间序列分形维数计算方法的异同点，进而提出一种多元时间序列维数计算方法。相关实验结果表明，该方法简便、可行，取得的效果较好。

(6)提出一种基于蚁群算法和分形维数的属性选择方法，以解决多元金融时间序列降维问题。在属性选择的基础上对多元时间序列的预测问题进行了研究。研究表明，该方法的效果较好，有利于识别出关键属性，提高预测的准确率。

关键词：分形维数　趋势分形维数　一元时间序列　多元时间序列　相似性分析　属性选择　时间序列预测

Abstract

With the rapid development of information construction of financial industry, more and more data has been created. How to make effective analysis of these data becomes a key issue. In recent years according to the dynamic property, complex and nonlinearity of financial data, researchers introduce nonlinear theory to reveal the market operational rules better. Fractal technology is a branch of nonlinear theory and related researches show that fractal is a common phenomenon in financial market.

The purpose of this dissertation is to do research on financial data analytical method based on fractal technology around hot and difficult questions of financial data analysis area. According to the character of financial data, fractal dimension definition, meaning and estimation methods of univariate and multivariate time series are researched; On these bases fractal dimension and data mining algorithm are combined to solve some financial data analysis problems which are similarity analysis, dimension reduction and prediction.

The primary work of this dissertation includes:

(1) The background and significance of this dissertation are discussed; the development of the fractal theory is introduced; The principles and methods of fractal technology in analyzing financial data are summarized.

(2) Some commonly used fractal dimension estimation methods of financial time series are introduced; Data fitting method in late solving process of estimating fractal dimension is discussed. Least square fitting method and least square sectional fitting method are used to fitting the data respectively. The related experimental results show least square sectional fitting method can improve fitting performance and the accuracy of fractal

dimension.

(3)In order to represent the fluctuation of financial time series better a tendency fractal dimension and its estimation method are proposed. This fractal dimension includes positive fractal dimension and negative fractal dimension. The experiments on stock data, exchange rate data and futures data demonstrate that both positive fractal dimension and negative fractal dimension indicate the uptrend or downtrend of financial market better than traditional fractal dimension indicates.

(4) The similarity analysis methods of financial time series are researched. A method combined with the tendency fractal dimension and K-means algorithm is proposed, and then stock index series similarity clustering is researched with this proposed method. In this method tendency fractal dimension is firstly used to represent the time series and then K-means algorithm is used to cluster the different index series. By comparison with the method that using traditional fractal dimension and K-means algorithm to cluster the stock index series, the results show tendency fractal dimension has the advantages of accurate and delicate description capability. Experimental results further demonstrate the meaning and function of tendency fractal dimension.

(5) Fractal dimension estimation methods of multivariate time series are analyzed and compared. An extended method is proposed to estimating the fractal dimension of multivariate time series. The fractal estimation method is simple, convenient and feasible and can get proper result.

(6) The feature selection method based on Ant Colony algorithm and fractal dimension is proposed for the problem of feature reduction of multivariate time series. Forecasting problem of multivariate financial time series is researched on the basis of this feature selection method. Experimental results show this improved feature selection algorithm can recognize determinant attributes and improve the accuracy of forecasting.

Keywords: Fractal Dimension; Tendency Fractal Dimension; Univariate Time Series; Multivariate Time Series; Similarity Analysis; Feature Selection; Time Series Prediction

目　录

第1章 绪 论

1.1 研究背景及意义

近年来，金融数据分析一直是人们研究的一个热点问题，原因有以下两方面：一方面，金融数据具有自身的特点，金融数据一般具有随机性、受外界因素影响较大的特点，且大部分数据具有高维、复杂、动态特性，因而需要准确地从这些数据中挖掘有价值的信息；另一方面，对金融数据进行分析可以使金融企业或金融投资者获得很多有价值的信息，不仅有利于金融企业或机构了解自己目前的运营情况，防范金融风险，而且有利于金融投资者了解金融市场的本质，更好地进行金融投资。

金融领域大量的数据是以时间序列的形式存在的，如股票交易数据、汇率价格、期货价格等。对这些金融时间序列数据进行分析通常有两大类方法，一是传统的时间序列分析方法；二是数据挖掘方法。传统的时间序列分析方法以模型分析方法为主，而模型分析方法建立在假设理论和数学基础之上[1]，每一种模型都有其适用条件。因而实际运用这种分析方法时往往具有一定的难度和局限性。数据挖掘方法弥补了传统的时间序列分析方法的不足。运用该方法时，可以从大量的、复杂的甚至含有噪声和模糊的实际数据中，提取出一定的有用规则，而且不需要事先知道数据的分布情况，得到的知识和信息也往往是单纯利用模型分析方法所不及的。

目前，运用于金融数据分析的数据挖掘算法有属性选择算法、分类、聚类、时间序列挖掘方法等。随着人们获取数据能力的提高，获取金融数据的手段不断多样化，相对于过去，可获取的数据量较大，且数据结构的复杂性

也在不断提高，因而不断发展数据挖掘技术，引入一些新的理论，可以更好地对数据进行分析。例如：模糊理论、云模型的引入提高了数据挖掘算法解决不确定性问题的能力，优化理论的引入提高了数据挖掘算法的处理速度和处理准确性，非线性科学理论的引入为数据挖掘算法解决复杂、非线性系统中的问题提供了可行的技术方案。近年来的研究表明，利用传统的线性理论分析金融数据往往不能得到很理想的结果，因而越来越多的研究者引入非线性理论用于分析金融数据，具体的有神经网络、混沌、分形、孤立波、经验模态分解等诸多方法[2~5]。其中，分形是一个比较活跃的研究领域。由于一个事物的形状、结构或分布具有自相似性，则可以利用分形理论对其进行分析。现实中有很多事物或现象都具有自相似性特点，在金融领域中这种分形现象更普遍，但需要说明的是这种相似性通常是从统计意义上说的。

Mandelbrot 通过研究股票价格序列，发现其具有分形特点，可以利用分形分布对其进行描述，此后，大量的研究者纷纷利用各种方法对金融市场进行了实证研究[6~11]。研究表明金融市场具有明显的分形特征，同时发现，分形理论中重要的定量指标——分形维数，在金融领域中有着特殊的作用。人们开始利用分形维数对一些金融现象如金融市场的大幅波动、股价未来趋势走向等进行了相关解释。引入分形理论分析金融数据虽然在一定程度上拓展了原有的市场分析理论，然而大多数研究仍然停留在实证研究和定性描述上。本书将利用分形维数及分形数据挖掘算法对一元及多元金融数据进行挖掘分析，利用定性与定量相结合的方式进一步挖掘金融市场中有价值的信息。其中主要的研究内容包括一元及多元金融时间序列分形维数的定义及计算方法，基于分形维数与数据挖掘算法的金融时间序列数据分析方法，包括金融时间序列数据的降维、相似性计算和趋势预测。

1.2 金融时间序列的分类

金融时间序列作为时间序列的一个特殊应用，具有非线性、非平稳性等特点，因而它属于非线性时间序列，需要利用非线性方法对其进行分析。如果从时间序列中所包含的变量个数的角度对其进行分类，金融时间序列可

以分为单变量金融时间序列(一元金融时间序列)和多变量金融时间序列(多元金融时间序列)两种。

时间序列是由一组按时间点所观测到的数据值组成的,可以把它表示成 $y=x_i(t);[i=1,\cdots,n;t=1,\cdots,m]$。当 $n=1$ 时,则为单变量时间序列,即随时间所观测到的数据值为一维数据。相应地,若 $n\geqslant 2$ 时,则为多变量时间序列,即随时间所观测到的数据值为多维数据。

多变量时间序列与一般多维数据集既有相似处也有不同的地方。相似的地方在于各变量之间并不是孤立的,具有一定的相关性,可能是非线性相关的。不同的地方在于多变量时间序列具有更大的复杂性。其中,影响数据变动的各变量都是随时间波动的,因而对其分析具有一定的难度。

在金融时间序列中,某一个时刻所获取的数据往往是有多个维度的。例如股票时间序列数据,其包含股票的开盘价、收盘价、最高价、最低价、成交量等多个变量。

1.3 金融数据分析方法

1.1 节中提到,金融数据分析的两种主要方法为传统的时间序列分析方法和数据挖掘方法。下面将分别介绍这两种方法的原理和应用情况。

1.3.1 传统的时间序列分析方法

传统的时间序列分析方法以数理统计模型为基础,通过假设、参数估计、检验等方式得到描述时间序列规律的模型[12]。简单地说,就是利用一些统计计量模型对金融市场进行建模,其分析方法有图表法、指标法和模型法[13]。其中模型法是目前最为常用的一种方法。

分析金融时间序列时,常用的模型有平稳时间序列模型、非平稳时间序列模型以及波动率模型。具体内容如下:

描述平稳时间序列的一个通用模型为自回归移动平均(ARMA)模型,ARMA 模型包含作为特例的自回归模型(AR)和移动平均(MA)模型[14]。该模型可以用 ARMA(p,q)表示,具体形式如公式(1-1)[13]所示。

$$y_t=\varphi_1 y_{t-1}+\varphi_2 y_{t-2}+\cdots+\varphi_p y_{t-p}+\mu_t-\theta_1 u_{t-1}-\theta_2 u_{t-2}-\cdots-\theta_q u_{t-q} \tag{1-1}$$

其中，$\{y_t\}$为一给定的时间序列；p 为自回归项；q 为移动平均项。模型中，时间序列的当前值不仅与误差值 μ_t 有关而且还与历史数据 $y_{t-1},\cdots,y_{t-p}$ 以及历史误差值 $u_{t-1},\cdots,u_{t-q}$ 相关[15]。

非平稳时间序列模型，是指方差、均值或两者都可能是变化的，典型的模型为差分自回归移动平均(ARIMA)模型。ARIMA 模型是 ARMA 模型的扩展，该模型可以用 ARIMA(p,d,q)表示。其中，参数 AR、p、MA、q 与 ARMA 中的含义相同；d 为时间序列转变成平稳序列所做的差分次数(阶数)。对于一个非平稳时间序列，先通过差分方法或其他技术方法将其转化成平稳时间序列，再利用 ARMA 模型进行建模。

目前在金融领域，人们往往利用 ARMA 模型或 ARIMA 模型对金融时间序列进行预测。有研究者将这两种模型用于预测汇价、股票收盘价以及期货价格等[16,17]。然而，更多情况下是将其与其他模型相结合，共同对金融数据进行分析。

金融数据具有波动性的特点，对其进行分析时，最常使用的模型为波动模型。其中分析时间序列波动性的一个模型为自回归条件异方差(ARCH)模型。ARCH 模型的描述如下：

p 阶自回归条件异方差 ARCH(p)模型，其定义由均值方程(1-2)和条件方程(1-3)给出：

$$y_t=\beta x_t+\varepsilon_t \tag{1-2}$$

$$h_t=var(\varepsilon_t|\Omega_{t-1})=a_0+a_1\varepsilon_{t-1}^2+a_2\varepsilon_{t-2}^2+\cdots+a_p\varepsilon_{t-p}^2 \tag{1-3}$$

其中，Ω_{t-1}表示 $t-1$ 时刻所有可得信息的集合；h_t 为条件方差。方程(1-3)中误差项 ε_t 的条件方差 h_t 分为两部分：一个常数项 a_0 和前 p 个时刻关于变化量的信息(用前 p 个时刻的残差平方表示)[18]。

在上述模型基础上，大量研究者根据研究中出现的实际问题得出了其衍生模型，如广义的 ARCH 模型(GARCH 模型)、NARCH、log-ARCH、GJR、TARCH 等。

近年来，人们通过研究发现，金融市场具有分形结构，导致传统的计量

模型仅仅能刻画金融市场短期相关性，而不能反映长期记忆特点。因而有学者将分形理论与上述统计分析模型相结合，形成新的描述时间序列的模型，如分数阶差分自回归滑动平均(ARFIMA)模型、部分整合自回归条件变异数(FIGARCH)模型，将分形差分参数引入上述两个模型，既考虑了短期情况也考虑了长期情况[19]，且相关的研究表明，上述模型往往比传统模型更符合市场实际情况，能够得到更好的结果[20]。

1.3.2 数据挖掘方法

目前，上述模型方法已得到广泛应用，很多人仍在对其进行研究，并加以改进，以期能更贴切地描述市场的真实情况，但是这种方法仍然存在着非常明显的缺点。由于运用统计模型方法的前提是需要假设观测数据满足某一种模型，然后通过检验的方法判断模型选择是否合理。如果合理可以运用模型进行解释和预测[1]。如果在检验中发现模型不合理则需要分析影响因素对其进行调整。由于每种模型都有其自身的约束条件和适用范围，金融市场的情况又是非常复杂的，因而运用模型分析方法往往不能很好地对市场的真实情况进行描述和求解。

数据挖掘方法的运用弥补了上述不足，它利用各种不同的方法和技术(包括统计分析方法)从数据中抽取模式。同时，挖掘的过程建立在学习、归纳和推理的基础之上，不需要事先假设数据的分布情况，因而往往能得到隐藏在数据中的更加多样和丰富的规则。近年来，数据挖掘方法在金融数据分析中得到了广泛应用。

相对于传统的时间序列分析方法，金融时间序列数据中应用数据挖掘方法研究的主要问题如下：

1. 金融时间序列的降维去噪

金融时间序列往往维数较高，且含有一定的随机波动性。研究降维去噪的目的是选用合适的算法降低金融时间序列的维数，去除噪声点，从而提高后期数据分析的效率和质量。

2. 金融时间序列相似性查询和比较

相似性查询和比较的研究目的是以一定的效率，查找到与给定序列相似的目标序列。对于金融时间序列来说，这种相似性通常是指波动相似性，

即变化规律上的相似性。相似性查询和比较通常是进一步挖掘数据的基础，比如可以利用序列间的相似性程度进一步对序列进行分类或聚类等。

3. 金融时间序列的模式分类与聚类

分类是有监督的学习，而聚类是无监督的学习。虽然两者在学习过程和实现方法上存在差异，但是它们的目的都是将金融时间序列归类为少数的几个模式，从而便于人们对金融市场的规律进行理解和把握[13]。

4. 金融时间序列的分割

金融时间序列分割算法的研究，是为了利用合适、高效的算法将时间序列划分为多个子序列，从而降低时间序列的维度。因而它也可以算是一种预处理过程。

5. 金融时间序列的预测

预测是金融领域中一个常见的研究问题，由于金融市场是一个受外界因素影响较大的复杂系统，所以预测存在着很大的难点。金融时间序列具有非平稳、非线性特点，因而人们研究的重点是运用非线性预测技术和预测模型对金融数据进行预测。

目前在金融数据分析中运用数据挖掘方法的研究成果很多。例如Zhang等人[21]利用神经网络对未来股票趋势进行预测，进而判断买卖点。Kirkos等人[22]分别利用决策树、神经网络和贝叶斯置信网络识别虚假财务状况。Shi等人[23]分别利用集成学习方法以及支持向量机对金融数据进行分类。Lkhagva等人[24]利用一种扩展的符号聚合近似(ESAX)方法表示金融时间序列。谭等人[25]利用模糊关联规则算法对抽取股票市场的交易规则进行分析。

从上述内容可以看出，在金融数据分析中采用的数据挖掘技术有多种，包括支持向量机、机器学习(决策树、神经网络等)、粗糙集、群智能算法、聚类分析方法以及分类分析方法等。

随着数据结构的日益复杂，越来越多的研究者将一些新的理论引入数据挖掘中，这其中包括分形理论。分形理论应用于数据挖掘领域形成了一些新的算法和技术，我们可以称之为分形数据挖掘技术。利用分形数据挖掘技术进行数据分析的一个最大优势就是其可以处理高维、非线性数据，且由于分形维数的特殊意义往往使得挖掘的结果更加准确和易于解释。目前

在金融数据分析中，已有研究者利用分形数据挖掘技术对金融时间序列数据进行分割、趋势预测以及聚类等。

由于数据挖掘技术与传统的模型分析方法都有各自的优势，因而近年来有部分研究者将两者结合起来对金融数据进行分析，可以把这种方法称为混合集成方法[26]。例如：Li 等人将 ARMA 模型与广义回归神经网络(GRNN)相结合，对金融时间序列进行预测，并取得了较好的效果[27]，Pai 和 Lin 等人将 ARIMA 模型和支持向量机结合预测股票价格[28]。Hassan 等人将隐马尔可夫模型(HMM)、神经网络以及遗传算法相结合共同对股票市场进行预测[29]。

1.4 分形理论及其在金融数据分析中的研究现状

1.4.1 分形理论的提出

美籍法国数学家曼德尔布罗特(Mandelbrot)最早提出分形理论，随后他又提出了分形几何学的完整思想，并认为分维是用于研究许多物理现象的有力工具[30]。分形理论最初是用于描述自然界和非线性系统中的不光滑和不规则几何形体的[31]。这些被描述的几何形体在某种意义下图形或结构上具有自相似性。例如：英国海岸线崎岖不平，利用常规的尺度无法得出其长度，但是其具有自相似的特点，即整体的海岸线和局部的海岸线在曲折程度和复杂度上具有相似性，因而可以用分形维数来表征这一相似性。图 1-1 所示为典型的科赫(Koch)曲线。首先将一条直线分成三等分，然后将中间的$\frac{1}{3}$用等边三角形的两边取代，这样不断迭代下去，直至无穷，其构造过程体现了局部与整体严格的相似性。对于 Koch 曲线从其迭代过程可以计算出其分形维数值。在金融领域，这种自相似性主要体现在时间上。图 1-2 所示为上证综指 2000 年至 2008 年的日收益率曲线。从曲线图中可以看出日收益率随时间变化在形态上呈现出一定的相似性。

分形理论与传统的欧氏几何理论不同，分形理论的定量指标——分形