第 34 卷 第 5 期 计 算 机 工 程 2008 年 3 月
Vol.34 No.5 Computer Engineering March 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)05—0185—03 文献标识码：A 中图分类号：TP181

基于商拓扑结构的序列构成和预测

何富贵[1]，张燕平[1,2]，赵 姝[1,2]，杨雪洁[1]，陈 洁[1,2]，张 铃[1,2]
(1. 安徽大学计算智能与信号处理实验室，合肥 230039；2. 安徽大学人工智能研究所，合肥 230039)

摘 要：对多变量时间序列进行分析有利于更好地了解各时间序列的特性。根据相关性的时间序列在商空间模型中，可依据信息相关性，该文综合利用多个相关序列提供的信息对其中一个序列进行了预测，通过商空间理论的分解和合成法减小信息不完备产生的影响，从而获得更多准确信息和规则。

关键词：商空间；商拓扑；时间序列；预测模型

Constituting and Forecasting of Series Based on Quotient Topology

HE Fu-gui[1], ZHANG Yan-ping[1,2], ZHAO Shu[1,2], YANG Xue-jie[1], CHEN Jie[1,2], ZHANG Ling[1,2]
(1. Key Lab of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039;
2. Institute of Artificial Intelligence, Anhui University, Hefei 230039)

【Abstract】 Researching these time-series which are interdependent as integration, namely multi-variable time-series analysis, the properties of these time-series can be realized better. For the forecasting question of time-series, in this paper, the relative time-series are dealt synthetically. In quotient space model, based on relativity of information, in order to reduce the impact of the incomplete information, the decomposition and synthesis method of quotient space theory is introduced to obtain more accurate information and rules. The result of experiments shows that the method is effective. Therefore, the solution to the problems is undoubtedly a reasonable approach.

【Key words】 quotient space; quotient topology; time-series; forecasting model

1 概述

对于随时间变化的动态数据，在建立动态系统模型时一般选择时间进程，即以时间作为基本自变量的分析方法，构成时间序列。选取一定的时段依次形成序列固然方便，但这可能掩埋了事物内在的拓扑关系。为此，本文依据对象论域的商拓扑结构来考虑时间序列的构成，将需要预测的因素作为决策属性处理，相关的关联因素作为序列的条件属性，按决策属性与条件属性影响的时间顺序，以及一定的时段来构成序列。

在时间序列分析中，文献[1]利用指数自回归成功地进行了经济混沌时序的预测。混沌吸引子具有十分复杂的几何结构且其内在行为具有相当的不规则性。通常，不同的混沌实测数据应该建立不同的混沌模型。因为混沌现象具有对初始条件的敏感依赖性，只要初始条件稍有差别或微小扰动，则会使系统的最终状态出现巨大的差异。所以混沌系统的长期演化行为是不可预测的。当用基于 BP 神经网络来实时地预测时序时，其结果不是很理想，主要是由传统的神经网络的学习算法而引起的，影响预测模型的可靠性和准确性。

小波神经网络是在小波分析的基础上提出的一种前馈网络，由于它引入了两个新的参变量，即伸缩因子和平移因子，因此小波神经网络具有比小波分解更多的自由度，从而使其具有更灵活有效的函数逼近能力，在塔式分解并选择性重构、滤波的基础上，留下时间序列的主要趋势[2]，为之后的神经网络训练并预测作了准备工作，经过筛选恰当的各个参数，通过小波神经网络能达到较佳的预测效果。但小波分析虽然可平滑变化曲线，去除干扰信号，但难以同时反映多个因素共同作用产生的结果。

现实世界的许多事件不是孤立的，相互间往往具有一定的联系。例如空气质量的变化、股票价格的波动等，所形成的序列之间都具有一定的相关性，目前人们已经认识到，复杂的实际问题需要智能系统对各种不同来源的信息进行综合。将具有一定相关性的多个时间序列作为一个整体进行研究，即进行多变量时间序列分析，这将有利于更好地了解各时间序列的特性。因此对于时间序列的预测问题，本文综合考虑具有相关性的时间序列。

2 商拓扑的序列构成

2.1 商拓扑

张钹院士和张铃教授在研究问题求解时，引入商集概念，创新地建立了商空间理论。该模型是用一个三元组(X,F,T)来描述一个问题，从不同的粒度(角度、层次)考察问题(X,F,T)，是指给定论域 X 的一个等价关系 R，并由 R 产生商集$[X]$，然后研究相应问题$([X],[F],[T])$，其中 F,T 分别是 X 上的属性函数和拓扑结构，$[F],[T]$分别是商集$[X]$上对应的商属性函数和

基金项目：国家“973”计划基金资助项目(2004CB318108)；国家自然科学基金资助项目(60675031, 60475017)；教育部博士点基金资助项目(20040357002)；安徽省自然科学基金资助项目(0504200208)；安徽省教育厅重点自然科学研究基金资助项目(2006KJ015A)；安徽省教育厅自然科学研究基金资助项目(2005KJ053)

作者简介：何富贵(1982－)，男，硕士研究生，主研方向：智能计算，人工神经网络；张燕平，教授、博士；赵 姝，博士；杨雪洁，硕士研究生；陈 洁，硕士；张 铃，教授、博士生导师

收稿日期：2007-08-19 **E-mail**：fuguihe@163.com

商结构。称([X],[F],[T])为(X,F,T)的商空间。X 的所有不同的商集及其对应的商空间，就构成了问题(X,F,T)的不同粒度世界[3]。

对于论域X，设对象(X,F,T)，对$A \subset X$具有性质H，若在[X]中引入结构[T]，使得在([X],[F],[T])中也具有性质H，则称[T]是(X,F,T)关于性质H的商结构。当(X,T)是半序集，则([X],[T])是关于元素的具有保序性(H)的商结构[3]。在商空间理论中应用商结构拓扑，可以使得商空间粒度与粒度之间转换得到一定的发展。

2.2 商拓扑序列

例如空气质量、股市走势这类数据，虽然一直处于变化的过程中，但它们都带有明显的时序性。时间序列从直观上说，数据值的变化直接受到时间参数的影响，因此直接利用原始数据不仅简单明了，而且容易发现数据瞬时变化的特性[4]。但仅以某一固定时段如天来构成基本的样本向量，则未有效考虑时序数据的前后关系和相互作用。

为了能恰当地反映时序性质的影响，每一个训练样本中的输入数据X_i都是一个由若干个顺序采样的多变量数据$X_i(t),X_i(t\text{-}1),\cdots,X_i(t\text{-}j)$按一定的关系及预测目标组成的子集，其中时间t的单位是时段，$X_i(t)$是该时段内的一组相关变量，如大气质量预测中的气温、气压、相对湿度、总云量、低云量、降水量、风速 7 个相关量，和股市走势预测中的开盘指数、最高指数、最低指数、收盘指数、成交金额这 5 个相关量。这样，下 1 个输入数据子集X_{i-1}将是$X_{i-1}(t\text{-}1),X_{i-1}(t\text{-}2),\cdots,X_{i-1}(t\text{-}j\text{-}1),X_{i-1}(t\text{-}j)$等，依此类推。若以日为时段单位，则为输入数据集将从当前日开始，向后每间隔 1 日，以j日为时间窗口，采集j个数据作为同一指标的输入。先后 2 个样本间，将有j-1 日数据重叠，相应的输入神经元的数目也将比单日扩大j倍，即训练数据集可表示为

$X_1(t) = \{X_{11}, X_{12}, X_{13},\cdots, X_{1m}\}$

$X_2(t) = \{X_{21}, X_{22}, X_{23},\cdots, X_{2m}\}$

…

$X_n(t) = \{X_{n1}, X_{n2}, X_{n3},\cdots, X_{nm}\}$

其中，X_{ij}为第i个时间序列的第j个观察值。

每一向量对应的决策属性为预测值，即对大气质量决策属性势 j+1 日的大气质量、对股市走势是 j+1 日的走势。这种序列构成，具有一定相关性的多个时间序列作为一个整体进行研究，即进行多变量时间序列分析，这将有利于更好地了解各时间序列的特性。

关于 j 值的确定，一般由预测的时段确定。例如是日大气质量预测、日股市走势预测，参考人对同类事件的时段考虑，j 为周或旬较合理；对周、旬、月的预测，一般采用混合粒度的序列构成，即序列中有近几日的日数据，也有以周或旬为粒度的数据，整个序列不是单一粒度的。

2.3 决策属性的变化序列

金融中的数据，尤其是股票数据的特点是交易频繁，如果单单按固定时间抽取价格数据，不能完全反映股市交易量单调增加，量推动价的股市中的这类现象，故本文将股市中的原始数据如开盘指数、最高指数、最低指数、收盘指数、成交金额不按固定时段整理，而是按固定成交量、变时段的方式整理成交易量序列数据集。

为了能恰当地反映出交易量单调增加的影响，每一个训练样本中的输入数据X_i都是一个由固定交易量而顺序采样时段、开盘指数、最高指数、最低指数、收盘指数的多变量数据$X_i(l),\ X_i(l\text{-}1),\cdots,\ X_i(l\text{-}j)$按一定的关系及预测目标组成的子集，其中，l的单位是一交易量；Xi(j)是该交易量内的一组相关变量。这样，下一个输入数据子集X_{i-1}将是$X_{i-1}(l\text{-}1),X_{i-1}(l\text{-}2),\cdots,X_{i-1}(l\text{-}j\text{-}1),X_{i-1}\ (l\text{-}j)$等，依此类推。即训练数据集可表示为

$X_1(l) = \{X_{11}, X_{12}, X_{13},\cdots, X_{1m}\}$

$X_2(l) = \{X_{21}, X_{22}, X_{23},\cdots, X_{2m}\}$

…

$X_n(l) = \{X_{n1}, X_{n2}, X_{n3},\cdots, X_{nm}\}$

其中，X_{ij}为第i个交易量序列的第j个观察值。

为了预测股市行情变化趋势，本文使用以一定交易量为准的移动平均值。其中，采用的交易量对上证大盘有 50 亿元、200 亿元；对各股(深发展)有 100 000 手、200 000 手，一手=100 股。对每一等交易量，做一次输入数据采样。为了能恰当地反映交易量性质的影响，每一个训练样本中的输入数据X_i都是一个由 8-21 个依费波里奇数变化的顺序采样数据，如取费波里奇数=8，则有$X_i(t\text{-}7),X_i(t\text{-}6),\cdots,X_i(t\text{-}1),X_i(t)$组成一个样本。

商拓扑模型表示的决策型序列数据，对多变量关联的数据构成的预测模型，能够有效地预测趋势的本质在于：序列的构成过程中考虑了信息粒度的影响，和信息发生前后的拓扑关系对决策属性的影响，及多种形式序列对同一决策属性的合成关系。

商空间理论能够通过分解或合成有效减少信息不完备或推理规则不明确的影响，由于我们对时间序列的构成不是以一个变量对应一个时间即一对一为基本拓展为多维序列，而是将一个时间对应的多个包含着各态信息变量即一对多为基本拓展为多维序列，所以使预测更符合实际，避免了只考虑其中一维变量而忽略了其他变量的影响，对构成的序列采用构造性学习方法[5]将数据本身相似的时间序列进行了划分、分类，能捕捉到数据本质的特性，而序列粒度的变化，对应于不同的短、中期的预测，通用性较好。

3 实验结果

3.1 大气质量预测

大气质量指的是低层大气中含的有害物质与气体(如：pm10、SO_2、NO_2)含量多就质量差，含量少就质量好。本文预测大气质量是指预测低层大气中pm10 的值。

基本信息由 2003 年 1 月、2 月和 12 月，2004 年 1 月、2 月和 12 月，2005 年 1 月、2 月和 12 月共 271 天的气温、气压、相对湿度、总云量、低云量、降水量、风速和 pm10 值八项数据。按日粒度计算形成序列 1，构成 271 个样本，每个样本 8 个属性，其中 7 个是特征属性，最后一个是决策属性，它是根据 pm10 的值来确定优、良、轻度、中度、重度 5 个等级，结果如表 1 所示 (其中，2003 年 1 月、2 月和 12 月，2004 年 1 月、2 月和 12 月，2005 年 1 月共 212 个样本作为学习样本，预测 2005 年 2 月的 pm10 等级)。

表 1 序列 1 的结果

算法	识别率/(%)	覆盖数	拒识个数	训练时间/s	测试时间/s
覆盖	72.881 4	118	8	0.093	0
SVM	66.101 7	/	/	1.125	0

考虑前 5 日和前 10 日数据对决策值的影响，依次构成序列 2(35 个特征属性，1 个决策属性，2003 年 1 月、2 月和 12 月，2004 年 1 月、2 月和 12 月，2005 年 1 月和 2 月共 228 个样本作为学习样本，预测 2005 年 12 月)；序列 3(70 个特

征属性，1 个决策属性，2003 年 1 月、2 月和 12 月，2004 年 1 月、2 月和 12 月，2005 年 1 月和 2 月共 213 个样本作为学习样本，预测 2005 年 12 月)，结果如表 2、表 3 所示。

表 2　序列 2 的结果

算法	识别率/(%)	覆盖数	拒识个数	训练时间/s	测试时间/s
覆盖	92.598 6	123	12	0.125	0
SVM	81.481 5	/	/	0.219	0.047

表 3　序列 3 的结果

算法	识别率/(%)	覆盖数	拒识个数	训练时间/s	测试时间/s
覆盖	100	73	7	0.093	0
SVM	100	/	/	0.125	0

可见，序列的构成过程中考虑了信息粒度和信息发生前后的拓扑关系对决策属性的影响，可有效提高预测的准确性。

3.2　股市走势预测

上证指数时间序列的预测模型由下周趋势、明日趋势两块 6 个预测目标组成。下周趋势由大盘五日均线预测、大盘 5 日后的走势预测、各板块 5 日均线预测 5 个投影组成，3 投影面的合成结果并适度考虑政策面和经济环境的影响，就是下周预测的结果；明日趋势由大盘 3 日均线预测、大盘 2 日均线预测、大盘半日数据的明日走势预测 3 个投影组成，3 投影面的合成结果并适度考虑政策面的影响，就是明日预测的结果。

对每一投影面的预测，采用的方法是将对应粒度的原始数据按目标需求整理成一个高维样本集，而后按多侧面递进算法[6]将其分解，以提高其识别力和泛化力。例如对明日趋势的 3 日均线预测，首先，以 1991 年 1 月 21 日～2001 年 6 月 7 日的 2 550 个交易日的数据转化为一个 100 维的训练集，即为 100×2550 个样本，对 2001 年 6 月 8 日～2002 年 5 月 17 日的 222 交易日的走势进行预测。本文按多侧面递进算法将其分解成 5 组 25 维的向量集，得到的实验结果如表 4 所示。

表 4　前 20 个交易日的测试结果

样本	正确	错误	拒识	正确率/(%)	识别率/(%)
1 个 100×2550	115	29	78	115/144=79.9	144/222=65
5 个 25×2550	167	13	42	167/180=92.8	180/222=81
1 个 100×2550	16	4	0	16/20=80	16/20=80
5 个 25×2550	18	2	0	18/20=90	18/20=90

3 日均线走势预测曲线与实际曲线的如图 1 所示。由图 1 可知，对紧邻样本集的预测，预测的效果比较理想，离开已知的样本集越远，预测的效果越差，半年之后，几乎都是拒识。这也符合实际情况，若仅依据半年甚至一年前的数据，就可判断出半年后的走势，这对瞬息万变的股市是没有实际意义的，在此，只是作为一种测试的方法。

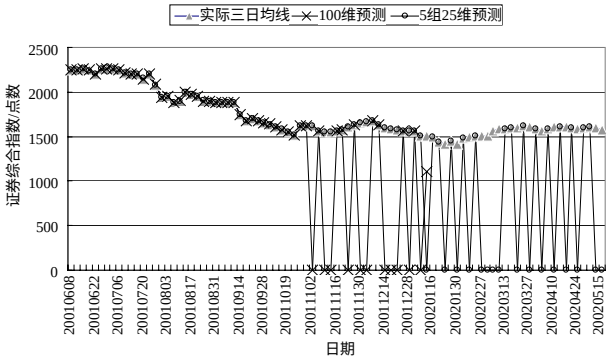

图 1　三日均线预测

可见，混合粒度模型可充分利用信息粒度的不同层次，即尽量综合利用多个相关序列提供的信息对其中一个序列进行预测。

4　结束语

本文综合考虑具有相关性的时间序列，依据信息的相关性，一部分作为同一粒度下依据论域的时间拓扑关系构成不同序列进行组合，一部分作为信息粒度的不同层次而被充分利用，使信息不完备的影响通过商空间理论中的分解和合成法尽可能地减少，从而获得尽可能多的准确信息和规则。相关的实验数据验证了这种方法的有效性。

参考文献

[1] 马军海, 陈予恕, 刘曾荣. 动力系统实测数据的非线性混沌模型重构[J]. 应用数学和力学, 1999, 20(11): 1128-1134.

[2] Daubechies I. The Wavelet Transform — — Time-frequency Localization and Signal Analysis[J]. IEEE Trans. on Info. Theory, 1990, 36(5): 427

[3] 张　铃, 张　钹. 问题求解理论及应用——商空间粒度计算理论及应用[M]. 2 版. 北京：清华大学出版社, 2007-03.

[4] Harris L, Gurel E. Price and Volume Effects Associated with Changes in the S&P500 List: New Evidence for the Existence of Price Pressures[J]. Journal of Finance, 1986, 41(4): 815-829.

[5] 张　铃, 张　钹. 多层前向网络的交叉覆盖设计算法[J]. 软件学报, 1999, 7(10): 737-742.

[6] 张燕平, 张　铃, 吴　涛. 机器学习中的多侧面递进算法[J]. 电子学报, 2005, 33(2): 327-331.

（上接第 184 页）

[6] Jain S. Digital Watermarking Techniques: A Case Study in Fingerprints & Faces[C]//Proc. of Computer Vision, Graphics and Image Processing Conf. Bangalore, India: [s. n.], 2000: 139-144.

[7] Ratha N K, Connell J H, Bolle R M. Secure Data Hiding in Wavelet Compressed Fingerprint Bangalore, India[C]//Proc. of ACM Multimedia. New York, USA: [s. n.], 2000: 127-130.

[8] Pankati S, Yeung M M. Verification Watermarks on Fingerprint Recognition and Retrieval[C]//Proc. of SPIE. San Jose, USA: [s. n.], 1999, 3657: 66-78.

[9] Gunsel B, Uludag U, Teklap A M. Robust Watermarking of Fingerprint Images[J]. Pattern Recognition, 2002, 35(12): 2739-2747.

[10] Bender W, Gruhl D, Morimoto N, et al. Techniques for Data Hiding[J]. IBM Systems Journal, 1996, 35(3/4): 313-336.

[11] Kutter M, Jordan F, Bossen F. Digital Signature of Color Images Using Amplitude Modulation[C]//Proc. of SPIE. San Jose, USA: [s. n.], 1997, 3022: 518-526.

[12] Daugman J. How Iris Recognition Works[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2004, 14(1): 21-30.

[13] Ye Xueye. Iris and Face Based Multi-biometrics Identification and Fusion Algorithm[D]. Hefei, China: University of Science and Technology of China, 2006.

[14] 中国科学技术大学智能信息处理实验室虹膜数据库[Z]. (2006-08-10). ftp://202.38.78.224/Sample/.