基于扰动的亚复杂动力系统因果关系挖掘

郑皎凌[1),2)] 唐常杰[1)] 乔少杰[3)] 杨 宁[1)] 李 川[1)] 陈 瑜[1)] 王 悦[1)]

[1)](四川大学计算机学院数据库与知识工程研究所 成都 610065)

[2)](成都信息工程学院软件工程系气象信息共享与数据挖掘实验室 成都 610225)

[3)](西南交通大学信息科学与技术学院 成都 610031)

摘 要 传统因果分析方法主要是基于具有分布预设的概率模型,但动力系统通常是存在反馈的非线性系统,不适合采用概率方法进行分析.针对这一问题,该文提出了基于扰动的亚复杂动力系统因果分析方法,主要工作包括:(1)采用基因表达式编程的函数拟合方法对动力系统时间序列进行差分方程拟合,减免了关于数据分布模型的预设;(2)基于得到的拟合函数,通过对自变量的扰动来计算因变量的相应波动,提出了根据扰动和波动的数值关系来判断自变量和因变量之间因果关系的判断准则,并基于该准则提出了因果关系挖掘算法和挖掘结果可信度验证方法;(3)在合成数据和真实数据上进行了翔实实验,结果表明该文所提出的算法能挖掘出合理因果关系,在不同数据规模情况下能得到一致挖掘结果.与两种基于概率统计的因果分析方法进行了对比实验,结果表明当系统要素多于两个时,该文的算法仍然能够得到多个要素间正确的因果关系,而两种基于概率统计的方法则无法挖掘出正确的因果关系.

关键词 亚复杂动力系统;因果关系分析;扰动;波动;函数拟合;数据挖掘

中图法分类号 TP311 **DOI 号** 10.3724/SP.J.1016.2014.02548

Mining Causality in Sub-Complex Dynamic System Based on Perturbation

ZHENG Jiao-Ling[1),2)] TANG Chang-Jie[1)] QIAO Shao-Jie[3)]

YANG Ning[1)] LI Chuan[1)] CHEN Yu[1)] WANG Yue[1)]

[1)](*Institute of Database and Knowledge Engineering, School of Computer Science, Sichuan University, Chengdu* 610065)

[2)](*Laboratory of Meteorological Information Sharing and Data Mining, Software Engineering Department, Chengdu University of Information Technology, Chengdu* 610225)

[3)](*School of Information Science and Technology, Southwest Jiaotong University, Chengdu* 610031)

Abstract Traditional causality analyzing methods are based on probability models with predefined distributions. However, as dynamic systems are usually non-liner systems with feedback loops, the probability methods are not suitable for analyzing dynamic systems. In order to deal with this problem, this study proposes a new method to analyze the causal relationship between elements in sub-complex dynamic system. Main contributions include: (1) this study uses Gene Expression Programming to regress dynamic systems' differential equations, and thus, avoid predefining the probability distribution function of the data; (2) based on the differential function, calculates the fluctuate value of the response variable by perturbing the independent variables. This study judges the causal relationships between the independent variables and the response variable by analyzing the numerical relationship between the response variables' fluctuate value

收稿日期:2012-03-20;最终修改稿收到日期:2014-06-27.本课题得到国家自然科学基金青年基金(61202250,61203172,61100045)、四川省教育厅青年基金(11ZB088)、成都信息工程学院中青年学术带头人科研基金(J201208)、成都信息工程学院引进人才项目(KYTZ201110)、高等学校博士学科点专项科研基金(20110184120008)及教育部人文社会科学研究青年基金(14YJCZH046)资助.**郑皎凌**,女,1981 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为数据库与知识工程、复杂系统与社会计算. E-mail: zjl9191@163.com. **唐常杰**,男,1946 年生,教授,博士生导师,主要研究领域为数据库与知识工程. **乔少杰**,男,1981 年生,博士,副教授,主要研究方向为数据库与知识工程. **杨 宁**,男,1974 年生,博士,讲师,主要研究方向为数据库与知识工程、数据流挖掘. **李 川**,男,1974 年生,博士,副教授,主要研究方向为数据库与知识工程. **陈 瑜**,男,1974 年生,博士,讲师,主要研究方向为进化计算与智能计算. **王 悦**,男,1981 年生,男,博士,讲师,主要研究方向为数据库与知识工程、机器学习.

and the independent variables' perturbation value. Based on the judging criterion, this study proposes causality mining algorithm and causality trustiness evaluation algorithm; (3) conducts experiments on synthesized and real datasets. These results show that the algorithm can find reliable causal relationships which accord with social and natural principles. The algorithm can get the same results with different data scales. This study conducts comparative experiments with two causality analyzing methods which are based on probability analysis. These results show that when the system contains more than two elements, our method can still obtain the correct causal results, while the two comparative methods cannot find the correct causal results.

Keywords　sub-complex dynamic system; causality analysis; perturbation; fluctuate; function regression; data mining

1　引　言

1.1　研究动机

复杂系统是目前很多领域的研究课题，如复杂网络、非线性系统和混沌系统等. 亚复杂系统是对复杂系统进行特征提取，忽略次要因素得到的一个较简单且有可能作出工程性解决方案的系统，简称亚复杂系统(Sub-Complex System，SCS). 亚复杂动力系统因果关系挖掘旨在从动力学关系上找出系统各要素间的因果关系.

此项研究有下列背景：(1) 为亚复杂系统干预规则挖掘提供可靠的因果关系支持，保证干预规则的合理性和正确性. 亚复杂系统干预规则挖掘旨在描述人工干预下亚复杂系统的动力学行为，解释干预的普适规律[1]. 而要制定干预规则必须首先弄清系统中各个要素间的因果关系，之后才能施加干预，也只有这样得出的决策才是可信的；(2) 在动力系统理论中，判断一个要素对另一个要素的影响是通过建立要素间的微分或差分方程，然后分析方程的特征属性来进行的. 但现实中很多要素间并没有直接的方程可循，例如人们认为地球的潮汐和温度可能与太阳黑子有关，但目前尚无相关的动力学方程. 找出这类动力系统的因果关系将为建立更准确的模型奠定基础；(3) 现有的因果关系研究对数据做了大量假设，如数据呈正态分布或没有环路因果链等，这些假设在实践环境中不一定成立，故需要新的方法来分析动力系统的因果关系. 针对上述问题，本文探索如何挖掘亚复杂动力系统因果关系.

由于亚复杂动力系统的运行机制总是与时滞和反馈相关的，所以进行分析的对象是系统中各个要素所产生的时间序列数据. 下面的实例可充分说明本文研究的对象和目标.

例 1.　有某段公路上随时间变化(每小时)的 4 个要素：汽车流量(car(t))，平均风速(wind(t))，二氧化氮浓度(NO_2(t))，温度(temp(t)).

(1) 挖掘对象，如图 1 所示，挖掘对象是由这 4 个要素产生的时间序列数据；

(2) 挖掘目标，显然这 4 个要素可以构成 $2\times C_4^2=12$ 对因果关系，挖掘目标是从这 12 对关系中排除掉虚假的因果关系，从而得到真实的因果关系. 如图 1 中 Car→NO_2 √ 表示 car 会影响 NO_2 是真实因果关系，而 NO_2→Car× 则表示 NO_2 会影响 car 是虚假的因果关系.

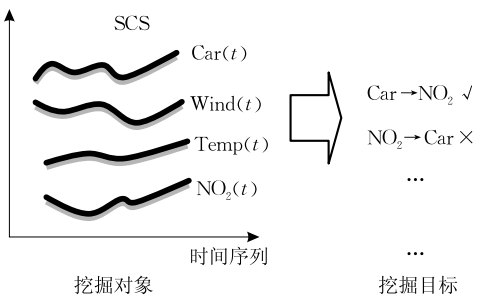


图 1　SCS 因果关系挖掘的对象和目标

下面对本文挖掘对象和目标进行形式化描述.

(1) 亚复杂动力系统因果挖掘对象 SCS. SCS 是系统中各个要素产生的所有时间序列数据，形式化表达成四元组$\langle m,k,T,E_{m\times k}\rangle$，$m$ 是系统中要素个数，k 是每个要素存在时滞个数，T 表示时间序列的长度，$E_{m\times k}$表示系统中所有要素在所有时滞上的集合.

如例 1 亚复杂动力系统的挖掘对象$SCS=\langle m,k,T,E_{m\times k}\rangle$，则 $m=\{\text{car},\text{wind},NO_2,\text{temp}\}$，若$k=1$，则 $E_{m\times k}=\{\text{car}(t),\text{car}(t-1),\text{wind}(t),\text{wind}(t-1),NO_2(t),NO_2(t-1),\text{temp}(t),\text{temp}(t-1)\}$，若 $T=$

200 d，则表示 SCS 中含有 200 d 的时间序列数据.

(2) 亚复杂动力系统因果挖掘目标. 设有 $SCS=\langle m,k,T,E_{m\times k}\rangle$，$m=\{x_1,x_2,\cdots,x_m\}$，则该系统的候选因果关系共有 $2\times C_m^2$ 个，记为 $Candidate\_Set=\{x_i\rightarrow x_j \mid 1\leq i,j\leq m, i\neq j\}$. 本文旨在挖掘出 $Candidate\_Set$ 子集 $Causality\_Set\in Candidate\_Set$，使得任意 $\langle x_j\rightarrow x_i\rangle\in Causality\_Set$ 都是系统中真实因果关系. 为描述简洁，令 $x_j\rightarrow x_i$ 表示 x_j 是 x_i 的原因.

1.2 问题的难点

(1) 基于概率对 SCS 进行挖掘的难点

传统的对时间序列进行因果挖掘的方法主要是基于概率的，如两个时间序列之间的相关系数，互信息，转移熵等. 而 SCS 各要素间和要素自身通常存在时滞和反馈，即整个系统模型主要是基于差分或微分方程，而不是基于概率状态的转移. 所以，我们认为采用概率方法进行 SCS 因果挖掘是不合适的.

(2) 基于回归对 SCS 进行挖掘的难点

传统的因果关系挖掘，如因果贝叶斯网和格兰杰因果检验等，都可以基于回归拟合来进行[2-4]，文献[5-6]采用了遗传算法来对因果贝叶斯网的概率参数进行拟合. 进化算法，特别是基因表达式编程(Gene Expression Programming, GEP)，具有强大的函数发现能力，给定预先假设好的因果关系，GEP 通常都能够拟合出适应度较高的函数，同时 GEP 还能够拟合出多种适应度较高的函数，显然这就使得这种基于拟合函数的因果分析方法可信度较低. 例 2 表明基于回归的因果关系发现方法可信度较低.

例 2. 例 1 的动力系统中有 4 个要素，从实际生活经验出发可知 car→NO_2 是可信因果关系，而 NO_2→car 则肯定是虚假的因果关系，因为汽车流量一般是不受道路上二氧化氮浓度影响的. 但在具体实验中，采用 GEP 进行函数回归，发现无论是以 car(t) 作为因变量，还是以 NO_2(t) 作为因变量，都能拟合出适应度很高的函数，故采用函数回归得到的因果关系实际上是不可信的.

(3) 存在噪声的可信因果关系挖掘

即使采用回归分析能得到较准确的因果关系，当数据中含有噪声数据时，挖掘结果也会受到影响，下面举例说明噪声数据的含义及其带来的影响.

例 3. 考虑例 1 的动力系统. 假设 car→NO_2 是真实的因果关系，而气温 Temp 和风速 Wind 对 NO_2 没有影响. 如果把 Car，Temp 和 Wind 作为自变量，把 NO_2 作为因变量进行函数拟合，来分析哪些要素是 NO_2 的原因，有可能得出错误结论，因为此时 Temp 和 Wind 是噪声数据，会对分析形成干扰(噪声数据对因果挖掘的干扰形式可参见 4.1 节例 6).

2 相关工作

如 1.2 节所述，传统的在时间序列上的因果关系挖掘主要是基于概率的，实践表明，SCS 的时间序列数据主要是通过差分或微分方程产生的，而不是基于状态间的概率转移的，基于此，本文方法与传统方法有较大不同. 本节旨在介绍传统的因果挖掘方法.

典型的因果分析方法有格兰杰因果检验[2-3]，因果贝叶斯网络[4-5]，基于信息理论的因果分析等. 按照研究对象分类，包括时间序列因果分析，非线性时间序列的因果分析和生物信号网络的因果分析.

格兰杰因果检验既考察变量间的相互关系又考虑其自身变化，格兰杰检验判断变量 X 是否能预测变量 Y，若不能，则认为 X 不能导致 Y，反之亦然. 但格兰杰检验需要假设所检验的序列是平稳的，而实际情况是很多宏观和金融序列都是非平稳的，文献[6]讨论了这个问题. 同样，对于因果系统的检验也是基于函数回归的，这种理论认为只要回归函数中自变量的时间早于因变量的时间，就称该系统是因果的，我们认为这是不太合理的，因为进化计算特别是 GEP 有着强大的函数发现能力，给定预先假设好的因果关系，GEP 通常都能够拟合出适应度较高的函数，显然这就使得这种对拟合函数的参数进行分析的方法可信度较低. 并且上述方法主要针对的都是线性系统，很难推广到非线性系统.

因果贝叶斯网络[7]能够学习变量间的概率依存关系及其随时间变化的规律，通过隐变量集合表达时间序列蕴含的潜在信息，通过构造一个有向无环图来反映一系列变量间的概率依存关系. 网络结构[8-9]是动态贝叶斯网络中一个关键部分，它表明了系统各元素间的因果关系. 目前主要基于进化算法来学习网络结构，这又会存在上述问题，即通过进化算法可得到多种不同且适应度都较高的网络结构，选取其中一种而放弃另一种都无法给出很好的解释. 同时，动态贝叶斯网络主要是进行状态间的因果分析，很难运用到数值型非线性系统因果挖掘中.

基于信息转移的因果分析利用信息论工具，如信息熵，动态熵，互信息和转移熵等指标来表示两个

信号序列间信息的传递方向，文献[10-11]考虑了因果方向. 由于互信息[12-13]是一种静态且对称的指标，不适合衡量信息流动的动态过程，文献[14]提出了转移熵来计算信息流间驱动和响应的关系. 但是，由于当时间序列的 n 维分布函数的维数较多或对连续取值区间的划分较细时，都会引起概率论中的“维数灾”. 如果预先对时间序列的性质作出某些假设，如服从马尔科夫性或泊松分布，都会降低分析准确性.

非线性时间序列的因果分析起源于线性时间序列，文献[15]采用搜索算法对多种时间要素构成的有向无环图进行搜索，来发现线性系统中的偏序关系和依赖关系. 文献[16]克服了上述算法只针对线性系统的弱点，提出了快速因果推理算法(Fast Causality Inference)，能够挖掘存在隐变量(latent variables)和反馈(feedback)的因果系统. 另一些研究基于一组已有的模型族，通过贝叶斯后验方法对这些模型进行打分，最后通过这些模型挖掘出系统的因果关系[17]. 这些方法依赖于一些先验的假设，比如系统中的各要素之间是条件独立的，或者服从正态分布等，从很大程度上限制了其应用范围.

在生物信号时间序列的分析中，主要采用各种滤波，自回归模型及相关性分析方法. 如采用相关性来分析神经信号与癫痫病发作的关系[18]，采用滤波来分析脑电波与各种行为的关系[19]以及采用自回归模型来分析心血管和心肺间各种生物信号的交互关系[20-21]. 目前研究的较多的蛋白质信号网络的因果分析则主要采用贝叶斯因果网络来进行分析. 如文献[22-23]提出了基于因果贝叶斯网的方法来分析人类 T 细胞中蛋白质信号网络的因果关系，其主要思路是通过操纵蛋白质中某种化学物质的含量(如磷和磷脂)来控制该蛋白质的活跃(active)或不活跃的状态(non-active)，继而来判断其对整个蛋白质信号网络的因果影响.

从最新的研究成果来看，Shi 等人[24]提出了基于时间序列相似性来判断复杂系统中各要素因果关系的方法，并成功地用于股票因果分析中；Snowsill 等人[25]基于重用度来分析 Web 上各个网页之间的引用因果关系，即把原创页面作为原因，把转载页面作为结果；Liu 等人[26]基于频繁项集来挖掘交通数据流中的时空因果干预关系. 这些算法都能够出色地分析出特定领域中各要素之间的因果关系，但对于复杂系统中存在环路、时滞的情况就不是很合适.

本文第 3 节介绍亚复杂动力系统因果挖掘过程，包括 3.2 节介绍 GEP 对时间序列数据的拟合，3.3 节介绍对拟合函数的扰动，3.4 节介绍基于扰动结果进行因果分析，并提出了整篇文章所依据的假设和定理；第 4 节提出基于该假设的两种因果挖掘算法和挖掘结果可信度判断方法；第 5 节给出在合成数据和真实数据上的实验结果.

3　基于扰动的 SCS 因果关系挖掘机制

3.1　总体挖掘流程

如前所述，本文的因果挖掘摆脱了传统的基于概率状态转移的思路，而主要是基于动力学中的扰动理论，本节旨在概括描述基于扰动的 SCS 因果关系挖掘的过程.

因果关系挖掘过程如图 2 所示. 设有亚复杂动力系统 $SCS=\langle m,k,T,E_{m\times k}\rangle$，$m=\{x_1,x_2,\cdots,x_m\}$，不失一般性，假设要分析 x_j 是否是 x_i 原因，即 $x_j\rightarrow x_i$，则挖掘流程如下：(1) 采用 GEP 以 $x_i(t)$ 为因变量，以 $E_{m\times k}-x_i(t)$ 为自变量进行函数拟合，得到函数 $x_i(t)=f(\{E_{m\times k}-x_i(t)\})$；(2) 基于扰动(后面将详细介绍)对 f 中的自变量 x_j 施加扰动 δ，并计算出因变量 x_i 的相应波动值 δf；(3) 根据因果判断原则判断 $x_j\rightarrow x_i$ 的真实性.

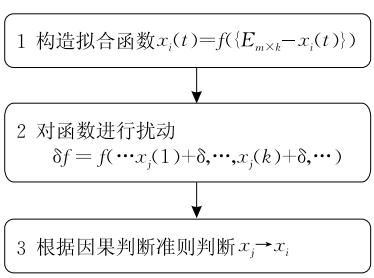


图 2　因果关系总体挖掘流程

3.2　基于 GEP 的函数拟合

本节旨在完成图 2 中的第 1 步，即构造拟合函数. 设有亚复杂动力系统 $SCS=\langle m,k,T,E_{m\times k}\rangle$. 则挖掘因果关系的第 1 步是构造以每个要素 $x_i\in m$ 作为因变量，以 $E_{m\times k}-x_i(t)$ 作为自变量的函数. 下面给出拟合函数的形式化定义.

定义 1.　ε-拟合函数 $f_{x_i(t)}$. 给定 $SCS=\langle m,k,T,E_{m\times k}\rangle$，$\varepsilon$-拟合函数 $f_{x_i(t)}$ 使得 $x_i'(t)=f_{x_i(t)}(\{E_{m\times k}-x_i(t)\})$，并且 $|(x_i'(t)-x_i(t))/x_i(t)|<\varepsilon$，其中 $x_i(t)$ 是原时间序列的值，$x_i'(t)$ 是通过 $f_{x_i(t)}$ 得到的拟合值，$\{E_{m\times k}-x_i(t)\}$ 是 $f_{x_i(t)}$ 的自变量集合.

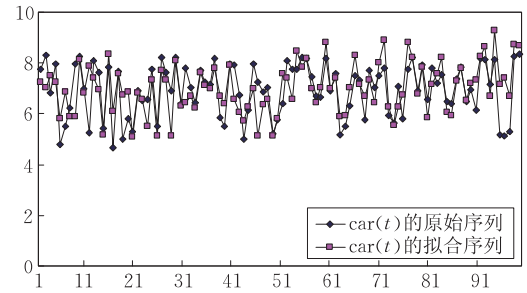
值得指出的是，进行函数拟合时，若因变量是 $x_i(t)$，则自变量是 $E_{m\times k}-x_i(t)$. 易知 $x_i(t-1)\sim$

$x_i(t-k)$也属于自变量，这是因为我们需要用$x_i(t-1)\sim x_i(t-k)$来构造关于要素x_i动力系统差分方程.下文中都按照上述方式进行函数拟合.由于GEP是遗传算法和遗传编程的结合算法，有强大函数发现能力，故这里采用GEP来进行函数挖掘，关于GEP的详细介绍参见文献[27].

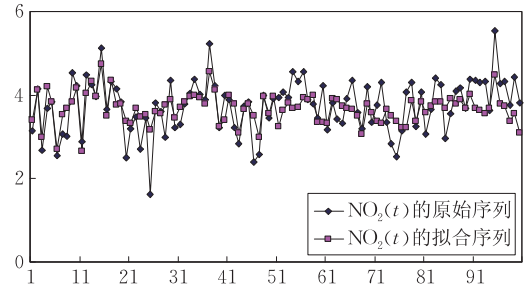
3.3 对拟合函数进行扰动

本节目的是解决第2节问题难点中的第一个难点，也即完成图2中的第2步.由于对真实和虚假的因果关系，GEP都能拟合出适应度较高的函数，无法直接基于拟合结果进行因果分析，我们在拟合结果的基础上，基于扰动和波动的数值关系进行因果判断，下面先举例说明.

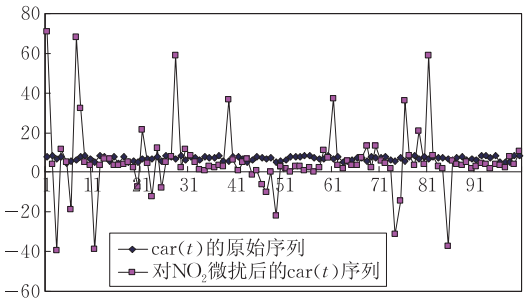
例4. 考虑例1中的动力系统.判断car与NO_2的因果关系，即判断是车流量引起二氧化碳浓度的变化，还是相反.(1)首先分别以二者作为因变量进行函数拟合设为$\mathrm{car}(t)=f(\cdots NO_2(t)\cdots)$，$NO_2(t)=f(\cdots \mathrm{car}(t)\cdots)$.从图3(a)、图3(b)可以看到，二者都有较好的拟合效果；(2)分别对拟合出的函数在自变量上施加一定的扰动，即$\mathrm{car}(t)=f(\cdots NO_2(t)+\delta\cdots)$，$NO_2(t)=f(\cdots \mathrm{car}(t)+\delta\cdots)$，然后观察因变量的波动情况.图3(c)、图3(d)显示了扰动后因变量的波动情况.可以看到图3(d)中$NO_2(t)$的波动很小，而图3(c)中car(t)式的波动很大.我们认为现实中的动力系统总是具有一定惯性的，当自变量发生微小扰动时，因变量会由于惯性不会产生很大的波动.所以，如果拟合函数反映的是动力系统中真实的因果关系，那么当对自变量进行微扰时，因变量的波动不大.反之，若拟合函数所反映的因果关系在现实系统中并不存在，那么对自变量的微扰可能会引起因变量的强烈变化，第3.4节给出具体解释.



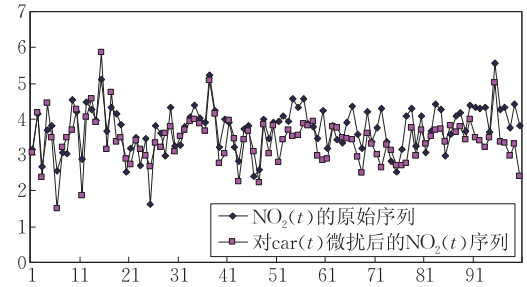
(a) car(t)的原始序列和拟合序列



(b) $NO_2(t)$的原始序列和拟合序列



(c) car(t)的原始序列和微扰后的序列



(d) $NO_2(t)$的原始序列和微扰后的序列

图3 真实因果关系和虚假因果关系的扰动效果(横坐标都表示以小时为单位的时间顺序，纵坐标表示在每个时间点上的汽车流量Car(t)或者是二氧化氮浓度$NO_2(t)$)

下面形式化描述对拟合函数自变量的扰动及因变量的相应波动.

定义2. 给定$SCS=\langle m,k,T,E_{m\times k}\rangle$及拟合函数$f_{x_i(t)}$.

函数的扰动$\delta f_{x_i(t)}(x_j,\delta,t_0)$：$\delta f_{x_i(t)}(x_j,\delta,t_0)$是挡在$t=t_0$时刻，当自变量$x_j(t_0)\sim x_j(t_0-k)$取得增量$\delta$，而其他自变量固定时，$f_{x_i(t)}$取得的增量，即$\delta f_{x_i(t)}(x_j,\delta,t_0)=f_{x_i(t)}(\cdots x_j(t_0)+\delta,\cdots,x_j(t_0-k)+\delta\cdots)-f_{x_i(t_0)}(\cdots x_j(t_0),\cdots,x_j(t_0-k)\cdots)$.

波动$\sum f_{x_i(t)}(x_j,\delta,t)$：波动是扰动在整个时间序列上的平方和，即$\sum f_{x_i(t)}(x_j,\delta,t)=\delta f_{x_i(t)}(x_j,\delta,t_1)^2+\cdots+\delta f_{x_i(t)}(x_j,\delta,t_T)^2$，其中$T$就是$SCS$中时间序列的长度.

例5. 设有$SCS=\langle m,k,T,E_{m\times k}\rangle$，$m=\{x,y\}$，$k=1$，$T=5$，$E_{m\times k}=\{x(t),x(t-1),y(t),y(t-1)\}$，

$x(t)$和 $y(t)$的时间序列如表 1 所示.

表 1　$x(t)$和 $y(t)$的时间序列

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
$x(t)$	1	3	7	10	15
$y(t)$	2	5	3	7	6
$y(t)+\delta$	4	7	5	9	8

设得到以 $x(t-1)$, $y(t-1)$为自变量，以 $x(t)$为因变量的拟合函数 $f_{x(t)}$: $x'(t)=x(t-1)+y(t-1)$($x(t)$是原时间序列的值，$x_i'(t)$是通过 $f_{x_i(t)}$得到的拟合值)，则当自变量 y 取得增量 $\delta=2$ 后，$f_{x(t)}$在时间序列 2～5 上的波动$\sum f_{x(t)}(y,\delta,t)=\delta f_{x(t)}(y,\delta,1)^2+\delta f_{x(t)}(y,\delta,2)^2+\delta f_{x(t)}(y,\delta,4)^2+\delta f_{x(t)}(y,\delta,4)^2$, $\delta f_{x(t)}(y,\delta,2)=|x'(2)-x(2)|^2=|(x(1)+(y(1)+\delta))-x(2)|^2=|1+4-3|^2$; $\delta f_{x(t)}(y,\delta,3)=|x'(3)-x(3)|^2=|(x'(2)+(y(2)+\delta))-x(3)|^2=|1+4+7-7|^2$; $\delta f_{x(t)}(y,\delta,4)=|x'(4)-x(4)|^2=|(x'(3)+(y(3)+\delta))-x(3)|^2=|1+4+7+5-10|^2$; $\delta f_{x(t)}(y,\delta,5)=|x'(5)-x(5)|^2=|(x'(4)+(y(4)+\delta))-x(3)|^2=|1+4+7+5+9-15|^2$.

3.4　因果判断准则

本节将基于函数扰动和波动概念，提出本研究依据的因果判断准则(图 2 中第 3 步). 为清晰表达，先给出本节用到的符号(表 2)和必要而合理的假设.

表 2　符号表

符号	解释
Max($f(x)$)	$f(x)$在整个时间序列上的最大值
Min($f(x)$)	$f(x)$在整个时间序列上的最小值
C	常数
δ_x	对 x 的扰动常数
$\delta f_y(x,\delta_x,t_0)$	函数的扰动，见定义 2
$\sum f_y(x,\delta_x,t)$	函数的波动，见定义 2

假设 1.　设动力系统含有两个要素 x, y，其中 x 影响 y，且有 $y=f(x)$; x 是随机时间序列，通过函数拟合，得到 $x=g(y)$，由于在实际系统中 $g(y)$具有时滞和其他自变量，故 $g(y)\neq f^{-1}(y)$. 并设 x_0, y_0分别是 x 和 y 的任意一个取值，$f'(x)$, $g'(y)$分别是 $f(x)$和 $g(y)$的一阶导函数，则有下列假设：

(1) $f(x)$, $g(y)$在其自变量的范围内具有各阶导数；

(2) $|f'(x_0)\delta_x|<C\times|\text{Max}(f(x))-\text{Min}(f(x))|$;

(3) $|g'(y_0)\delta_y|\gg C\times|\text{Max}(g(y))-\text{Min}(g(y))|$.

解释：对于(1)，动力系统函数对自变量各阶可导的要求在实践中容易满足. 对于(2)、(3)，我们以实验 5.2.1 节中的 box-jenkins 时间序列数据的一部分(如图 4(a)、图 4(b)来进行说明. 由于 $CO_2(t)$具有一定的惯性，图 4(a)中点 x_0 处作切线，有 $\Delta_1=f'(x_0)\delta_1$和 $\Delta_2=f'(x_0)\delta_2$，并且由于 $f(x)$波动较小，当 δ 较大时仍有 $|f'(x_0)\delta_x|<C\times|\text{Max}(f(x))-\text{Min}(f(x))|$. 而 Gas$(t)$是人为随机控制，波动较大，在图 4(b)中点 y_0 处作切线，有 $\Delta_1=f'(x_0)\delta_1$和 $\Delta_2=f'(x_0)\delta_2$，但当 δ 较大时，有 $|g'(y_0)\delta_y|\gg C\times|\text{Max}(g(y))-\text{Min}(g(y))|$.

值得指出的是，假设中存在两个参数，即对 x 施加扰动的大小 δ_x和常数 C. 我们认为 δ_x取 $x(t)$在整个时间序列上的平均值是比较合适的，C 一般取 10.

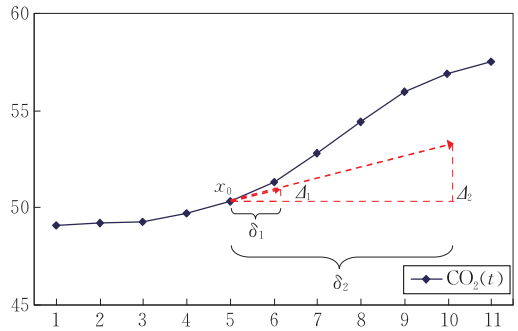
定理 1.　设动力系统含有两个要素 x, y，其中 x 影响 y，且有 $y=f(x)$; x 是随机时间序列，且时间序列的长度为 T，通过函数拟合，得到 $x=g(y)$，则当 δ_x和 δ_y取较大值时有：

(1) $|\delta f_y(x,\delta_x,t_0)|<C\times|\text{Max}(f(x))-\text{Min}(f(x))|$;

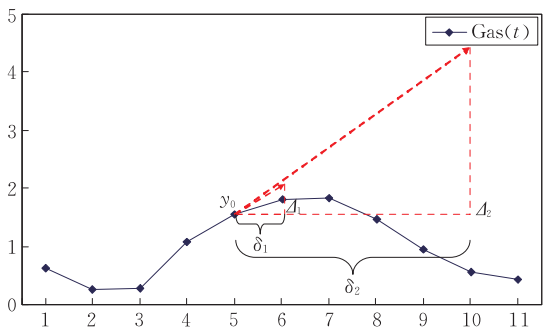
(2) $|\delta g_x(y,\delta_y,t_0)|\gg C\times|\text{Max}(g(y))-\text{Min}(g(y))|$;

(3) $\sum f_y(x,\delta_x,t)<C\times T\times|\text{Max}(f(x))-\text{Min}(f(x))|^2$;

(4) $\sum g_x(y,\delta_y,t)\gg C\times T\times|\text{Max}(g(y))-$



(a) 对$CO_2(t)$在不同尺度上进行扰动后的波动结果



(b) 对Gas(t)在不同尺度上进行扰动后的波动结果

图 4　在不同时间序列上进行扰动的结果(横坐标都表示以小时为单位的时间顺序，纵坐标表示在每个时间点上的二氧化碳浓度 $CO_2(t)$或者是天然气浓度 Gas(t))

$\mathrm{Min}(g(y))|^2$.

证明. 见附录 1.

基于假设 1 和定理 1 有以下因果判断定理 2.

定理 2. 设有 $SCS=\langle m,k,T,E_{m\times k}\rangle$,有任意两个要素 $x,y\in m$,通过函数拟合得到拟合函数 $y=f(x)$,如果 $\sum f_y(x,\delta_x,t)\gg C\times T\times|\mathrm{Max}(f(x))-\mathrm{Min}(f(x))|^2$,那么 x 不是 y 的原因.

证明. 反证,设 x 是 y 的原因,由定理 1 有 $\sum f_y(x,\delta_x,t)<C\times T\times|\mathrm{Max}(f(x))-\mathrm{Min}(f(x))|^2$ 矛盾,命题得证. 证毕.

即当 $\sum f_y(x,\delta_x,t)\gg C\times T\times|\mathrm{Max}(f(x))-\mathrm{Min}(f(x))|^2$ 时,可得 x 不是 y 的原因,但无法证明当 $\sum f_y(x,\delta_x,t)<C\times T\times|\mathrm{Max}(f(x))-\mathrm{Min}(f(x))|^2$,$x$ 就是 y 的原因. 故第 4 节的算法都是基于排除法来判断因果关系的.

4 亚复杂动力系统因果关系挖掘算法及分析

本节将描述具体的挖掘算法,并解决第 2 节的第二个难点,即含噪声数据的因果挖掘. 其中朴素算法只通过一次扰动来判断因果关系,而抗噪算法可通过多次回归和扰动来逐步排除虚假因果关系.

4.1 基于扰动的朴素因果挖掘算法

算法如算法 1,例 6 给出了说明.

算法 1. *Naïve_Causality_Mining*( )

输入:$SCS=\langle m,k,T,E_{m\times k}\rangle$ 动力系统时间序列

输出:挖掘出的可信因果关系集合 *Causality_Set*

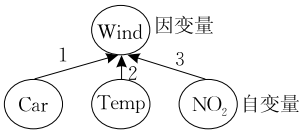
```
1. Causality_Set=null; //初始为空
2. For each x_i in m
3.   target=x_i(t);
4.   variables=E_{m×k}-x_i(t);
5.   f_{x_i(t)}=Generate_Function (target, variables);
6.   For each x_j(t)~x_j(t-k) in variables
7.     Calculate δf_{x_i(t)}(x_j,δ);
8.     If δf_{x_i(t)}(x_j,δ) don't satisfy Lemma 2
9.       Causality_Set=Causality_Set-⟨x_j→x_i⟩;
10.    End If
11.  End For
12. End For
13. Return Causality_Set( );
```

(1)第 5 行,*Generate_Function*(*target*,*variables*)表示以 *target* 为因变量,以 *variables* 为自变量进行基于 GEP 的函数拟合,得到函数 $f_{x_i(t)}$;

(2)第 6 行,表示依次对 $f_{x_i(t)}$ 自变量集合 *variables* 中的每个要素进行扰动. 由于系统的时滞为 k,故对每个要素 x_j 所有时间序列上的值 $x_j(t)\sim x_j(t-k)$ 都进行扰动;

(3)第 7~10 行,基于扰动计算出波动 $\delta f_{x_i(t)}(x_j,\delta)$(第 7 行),然后根据波动大小进行因果判断(8~10 行).

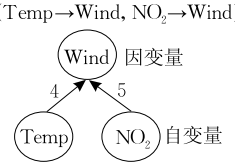
例 6. 考虑例 1 中的动力系统,分析影响 Wind 的要素. 考虑图 5 及表 3,通过朴素因果挖掘算法来分析影响 Wind 的要素只需一次函数拟合,而自变量是所有其他要素. 表 3 中自变量,因变量及扰动,波动的含义见定义 2,如表 3 第 2 行扰动为 Temp+(−3),表示对自变量 Temp 时间序列实施的扰动为 Temp=Temp+(−3). 参考值表示由自变量拟合出的关于因变量的函数在整个时间序列上的最大值与最小值差值的平方. 然后根据定理 2 来排除虚假因果关系. 可以看到,假设 $C=10$,那么表 3 第 1 行的因果关系就可以直接被定理 2 排除,因为波动$\gg C\times T\times$参考值. 而 2,3 行的因果关系不能被定理 2 排除,所以算法 1 的挖掘结果为{Temp→Wind,NO_2→Wind}.



算法1对Wind=f(Car,Temp,NO_2)进行扰动分析

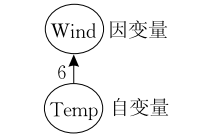
算法1排除Car→Wind

算法1挖掘结果{Temp→Wind, NO_2→Wind}



去掉自变量Car后,算法2对Wind=f(Temp,NO_2)进行扰动分析

算法2排除NO_2→Wind



去掉自变量NO_2后,算法2对Wind=f(Temp)进行扰动分析

算法2挖掘结果Temp→Wind

图 5 因果挖掘算法示意图

表 3 因果挖掘算法扰动结果

	自变量	因变量	扰动	波动	参考值
1	Car,Temp,NO_2	Wind	Car+(−6)	Infinity	37974
2	Car,Temp,NO_2	Wind	Temp+(−3)	79099	37974
3	Car,Temp,NO_2	Wind	NO_2+(−3)	15909	37974
4	Temp,NO_2	Wind	Temp+(−3)	6574	37974
5	Temp,NO_2	Wind	NO_2+(−3)	939678	37974
6	Temp	Wind	Temp+(−3)	3332	37974

4.2 基于扰动的抗噪因果挖掘算法

如例 6 所示,在朴素算法的挖掘过程中,表 3 的

第 1 行已经判断出 Car 不是 Wind 的原因，但在表 3 的 2、3 行中仍然采用含有自变量 Car 的拟合函数来分析 Temp 和 NO_2 对 Wind 的影响，这显然是不合理的，因为没有去掉噪声 Car，下面的抗噪因果挖掘算法旨在解决上述问题.

算法 2. *Noise_Robust_Causality_Mining*( ).

输入：$SCS=\langle m,k,T,E_{m\times k}\rangle$ 动力时间序列数据

输出：挖掘出的可信因果关系集合 *Causality_Set*

1. *Causality_Set*=null;//初始为空
2. *Completeness*=false;//因果关系集合的完备性验证结果初始为假
3. For each x_i in m
4. $target=x_i(t)$;
5. $variables=E_{m\times k}-x_i(t)$;
6. $variables=Clear\_Noise(target,variables)$;
7. If ($variables\neq$null)
8. For each $x_j(t)\sim x_j(t-k)$ in *variables*
9. $Causality\_Set=Causality\_Set+\langle x_j\rightarrow x_i\rangle$;
10. End For
11. End If
12. End For
13. If ($variables\neq$null)
14. $Completeness=Judge\_Trustiness(target,variables)$;
15. End If
16. Return *Causality_Set*( );

过程 1. *Clear_Noise*(*target*,*variables*).

输入：因变量 *target*，自变量集合 *variables*

输出：降噪后的自变量集合

1. $f_{x_i(t)}=Generate\_Function(target,variables)$;
2. While ($variables\neq$null) Do
3. For each $x_j(t)\sim x_j(t-k)$ in *variables*
4. Calculate $\delta f_{x_i(t)}(x_j,\delta)$;
5. If !($\delta f_{x_i(t)}(x_j,\delta)$ satisfy causality scenario)
6. $variables=variables-\{x_j(t)\sim x_j(t-k)\}$;
7. End If
8. End For
9. If (*variables* is changed) and ($variables\neq$null)
10. $f_{x_i(t)}=Generate\_Function(target,variables)$;
11. End If
12. Else break;
13. End While
14. Return *variables*;

过程 2. *Judge_Trustiness*(*target*,*variables*).

输入：经过 *Clear_Noise* 进行降噪后的自变量集合

输出：因果关系集合的完备性验证结果

1. $f_{x_i(t)}=Generate\_Function(target,variables)$;
2. Calculate $\delta f_{x_i(t)}(variables,\delta)$;
//对自变量集合 *variables* 中的所有变量同时进行
//扰动，并计算相应拟合函数的波动
3. If ($\delta f_{x_i(t)}(variables,\delta)$ satisfy causality scenario)
//如果对所有变量同时进行扰动时，拟合函数相应
//波动幅度仍然在因果判断准则所规定的范围内，
//则说明算法所挖掘出的结果是可信的
4. Return true;
5. Else
6. Return false;
7. End If

(1) 算法 *Noise_Robust_Causality_Mining*() 的核心是第 6 行降噪函数 *Clear_Noise* 以及第 14 行的结果完备性判断函数 *Judge_Trustiness*，函数 *Clear_Noise* 主要是循环去掉 *variables* 集合中所有不是因变量的原因的要素，函数 *Judge_Trustiness* 主要是进一步判断算法挖掘出的因果关系集合的可信性，如果函数的返回值为真，那么在很大程度上算法挖掘出的因果关系集合是可信的.

(2) 函数 *Clear_Noise*(*target*,*variables*). 第 1 行 *Generate_Function*(*target*, *variables*) 表示以 *target*(即 x_i) 为因变量，以 *variables* 为自变量进行基于 GEP 的函数拟合，得到函数 $f_{x_i(t)}$. 第 5 行基于因果判断准则将 *variables* 集合中不是 x_i 原因的要素 x_j 去掉，整个过程循环进行，直到 *variables* 集合为空或不发生变化.

(3) 函数 *Judge_Trustines*(*target*,*variables*) 的核心是 2,3 行，其含义是对自变量集合 *variables* 中的所有变量同时进行扰动并计算相应拟合函数的波动，如果在对自变量集合 *variables* 中的所有变量同时进行扰动的情况下，拟合函数的相应波动幅度仍然在因果判断准则所规定的范围内，则说明算法所挖掘出的结果是可信的，否则是不可信的. 因为现实中的系统在其要素都发生适当扰动时通常是能够保持一定的稳定性的，如果拟合函数在其自变量同时发生扰动后也具有较强的稳定性，则可以进一步说明拟合函数能够较好地符合现实中动力系统的性质，进一步证实结果的可信性.

下面通过例 7 和例 8 来举例说明算法 2 的运行过程，其中函数 *Clear_Noise* 的运行过程通过例 7 说明，函数 *Judge_Trustiness* 的运行过程通过例 8 说明.

例 7. 考虑图 5 及表 3，背景和术语与例 6 相同.

(1) 根据算法 2 的 3～5 行，执行图 5 中的 1～3 步(对应表 3 的 1～3 行)，此时 $variables=\{$Car, Temp, $NO_2\}$；(2) 当发现 Car 不是 Wind 的原因时(表 3 第 1 行)，将 Car 从 *variables* 中删掉，即此时 $variables=\{$Temp, $NO_2\}$，然后重新进行函数拟合，

并重新扰动 Temp，NO_2，如图 5 第 4，5 步（表 3 第 4，5 行）；(3) 此时发现 NO_2 也不是 Wind 的原因（表 3 第 5 行），故继续将 NO_2 从 *variables* 中删掉，即此时 *variables*={Temp}，又重新进行函数拟合并重新扰动 Temp，如图 5 第 6 步（表 3 第 6 行）所示. 因为 $3332<10\times37974$，故 NO_2 是 Wind 的原因.

例 8. 考虑表 4，背景和术语与例 6 相同. 表 4 的目标是判断算法 2 的挖掘结果是否可信. 通过表 4 可以看到，当以要素 Car，Temp 为自变量，以要素 NO_2 为因变量进行函数拟合后，分别单独对自变量 Car（扰动结果如表 4 第 1 行）和自变量 Temp（扰动结果如表 4 第 2 行）进行扰动时，其波动的幅度符合因果判断准则的范围（即 $6459<10\times1439$，$64<10\times1439$），故算法 2 认为 Car，Temp 都是 NO_2 的原因. 而通过 *Judge_Trustines* 同时对 Car，Temp 进行扰动，得到拟合函数波动的幅度仍然近似符合因果判断准则的范围，因为 118 与 $10\times11=110$ 的数值相差很小，则函数 *Judge_Trustines* 认为{Car→NO_2，Temp→NO_2}是可信的因果关系.

表 4 因果挖掘算法扰动结果

	自变量	因变量	扰动	波动	参考值
1	Car，Temp	NO_2	Car+(−6)	6459	1439
2	Car，Temp	NO_2	Temp+(−3)	64	1439
3	Car，Temp	NO_2	Car+(−3)，Temp+(−3)	118	11

4.3 算法分析

4.3.1 GEP 生成拟合函数的一致性

算法 1，2 都要先采用 GEP 生成时间序列的拟合函数，而同一组数据可拟合出多个函数. 由于算法 1，2 都只基于其中一个拟合函数进行扰动，本节将用定理 3 证明只要拟合函数达到足够高的精度，那么无论选取哪一个函数进行扰动，算法的挖掘结果是相同的. 为了证明定理 3，需要有下面合理的假设 2.

假设 2. 当拟合精度足够高时，拟合函数的值和原函数的值相差很小有$|f_0(x)-f_1(x)|<\varepsilon$.

定理 3. 设有亚复杂动力系统 $SCS=\langle m,k,T,E_{m\times k}\rangle$，有任意两个要素 $x,y\in m$，x 是随机时间序列，y 是受 x 影响的因变量，设 $y=f_0(x)$ 是 x,y 真正符合的函数，$y=f_1(x)$ 是采用 GEP 生成的任意拟合函数，则当拟合精度足够高时，有

(1)$|f_0'(x)|\approx|f_1'(x)|$，$f_0'(x)$，$f_1'(x)$分别是 $f_0(x)$，$f_1(x)$的导数.

(2) 如果$\sum f_0(x,\delta_x,t)<C\times T\times|\mathrm{Max}(f_0(x))-\mathrm{Min}(f_0(x))|^2$，那么$\sum f_1(x,\delta_x,t)<C\times T\times|\mathrm{Max}(f_1(x))-\mathrm{Min}(f_1(x))|^2$，$\delta_x$是 x 时间序列的平均值.

证明. 见附录 1.

4.3.2 算法的完备性讨论

算法 1，2 都是采用排除法，去掉非因果关系，则剩下的关系就是挖掘结果，下面对排除法的完备性作简单讨论.

排除法指出如果 X,Y 的时间序列数据关系符合定理 2，则 X,Y 不具有因果关系，但排除法只能排除具备定理 2 中数据性质的非因果关系，不能排除具备其他类型数据性质的非因果关系. 故排除法是不完备的，即算法去掉的都是非因果关系，但并不是所有的非因果关系算法都能排除.

但本文所提出算法 2 的 *Judge_Trustiness* 函数对挖掘出的因果关系集合的可信度进行了进一步的判断，因为该函数提出一个很强的扰动和波动规则，即当所有自变量同时发生扰动时，如果拟合函数的波动范围仍然在规定范围内，算法才认定所挖掘出的因果关系集合可信，所以函数 *Judge_Trustiness* 的使用提高了算法的完备性.

在社会，经济以及自然科学中对因果关系的大量研究说明，如果动力系统中两个要素 X,Y 不具备因果关系，那么在很大程度上 X,Y 所体现出的数据关系都是符合定理 2 的. 比如，文献[24]就指出两个时间序列如果呈现出相似性，那么这两个时间序列就可能含有因果关系. 显然，本文所提出的相似性要求更强，因为本文认为除了两个时间序列本身呈现出相似性或可回归性外，并且在微扰作用下两个时间序列的变化规律也呈现出相似性. 这就说明了定理 2 中的排除法能够排除大部分非因果关系，故定理 2 是具备一定的普适性和完备性的.

4.3.3 算法复杂度分析

从算法 1，2 可知，算法的时间主要消耗在用 GEP 进行函数拟合的过程中，如果以 GEP 函数拟合一次作为单位时间，则有下面的算法时间复杂度.

设有动力系统的时间序列数据 $SCS=\langle m,k,T,E_{m\times k}\rangle$，则

(1) 朴素因果挖掘算法的时间复杂度为 $O(|m|)$，因为对 m 中的每个要素，算法只进行一次 GEP 拟合.

(2) 抗噪因果挖掘算法的时间复杂度为 $O(|m|^2)$，首先来分析最坏情况下所需的降噪次数，因为对 m 中的每个要素 x_i，算法在最坏情况下需要对 m 中除 x_i 外的其他的 $|m|-1$ 个要素依次降噪，故需进行 $|m|-1$ 次 GEP 拟合，则对 m 中的所有要素共需进

行$|m|\times|m-1|$次 GEP 拟合. 但在最好情况下的降噪次数为 0，即进行 GEP 拟合的次数与朴素算法相同为$|m|$次. 我们认为需要$|m|$次与需要$|m|\times|m-1|$次情况出现的概率相同，则抗噪算法平均需要的 GEP 拟合次数为$(|m|\times1+|m|\times2+\cdots+|m|\times|m-1|)/(|m|-1)=|m|^2/2$，故抗噪算法的平均时间复杂度仍为$O(|m|^2)$.

5　实验和性能分析

实验数据包括 1 组合成数据和 3 组真实数据. 数据量分别为 100，200，100 和 5000，都是亚复杂动力系统若干要素组成的时间序列，分别采用朴素和抗噪因果挖掘算法挖掘，及可信度判断函数来验证最终结果的可信性. 所有实验的实验过程如图 6 所示.

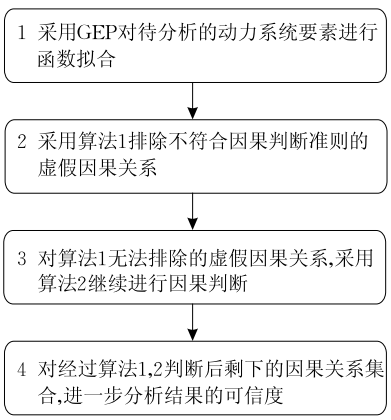


图 6　因果关系挖掘过程

本节中表 6、10、14 的背景和术语与例 6 相同，$C=10$. 并且 ε-拟合函数中 ε 的平均值小于 0.1.

同时，我们选择了两种现有的因果分析技术，包括格兰杰因果分析和基于转移熵的因果分析，来与本文的因果分析结果进行对比. 这两种方法都能够进行两个时间序列的因果方向分析，当实验数据集中的要素个数大于两个时，实验结果列出了要素集合两两组合的因果分析结果.

5.1　在合成数据上进行的实验

构造了如式(*)的一个动力系统，存在 3 个要素 x,y,u，该系统时间序列数据的形式化描述为 $SCS=\langle m,k,T,E_{m\times k}\rangle$，$m=\{x,y,u\}$，$k=1$，$T=100$，$E_{m\times k}=\{x(t),x(t-1),y(t),y(t-1),u(t),u(t-1)\}$. 数据按如下方式生成：(1)令 $x(0)=0.5$，$y(0)=0.5$，$u(t)$是 0 和 0.5 之间的随机时间序列；(2)将 $x(t-1)$，$y(t-1)$，$u(t)$的值代入式(*)迭代运行 100 次，得到含有 100 个数据的时间序列.

$$\begin{cases}u(t)=Random(0,1)\times0.5\\x(t)=f=u(t)\times(x(t-1)+1/(x(t-1)+\\\qquad y(t-1)+u(t)))\\y(t)=g=x(t-1)/(x(t-1)+y(t-1)+u(t))\end{cases}\quad(*)$$

该系统共 6 个候选因果关系$\{u\to x,u\to y,x\to y,x\to u,y\to x,y\to u\}$. 表 5 给出了如何通过算法 1，2 以及可信度验证函数来判断和验证该系统的因果关系. 表 5 中算法 1 和 2 得出判断结论的实验数据见表 6 的 1～6 行. 表 5 中可信度验证函数得出判断结论的实验数据见表 6 的 7，8 行. 其中，表 5 的第 6 列“依据”表示通过了表 6 中的哪几行得到. 如第 1 行“1，$x\to u$，假，—，—，2”表示因果关系 $x\to u$ 通过算法 1 判断为假，该结论通过表 6 第 2 行得到.

表 5　实验结果

	候选因果关系	算法 1	算法 2	可信度验证	依据
1	$x\to u$	假	—	—	2
2	$y\to u$	假	—	—	1
3	$x\to y$	真	真	真	3,4,7
4	$u\to y$	真	真	真	3,4,7
5	$y\to x$	真	真	真	5,6,8
6	$u\to x$	真	真	真	5,6,8

表 6　扰动及波动结果

	自变量	因变量	扰动	波动	参考值
1	x,y	u	$y+0.5$	Infinity	16
2	x	u	$x+0.5$	Infinity	16
3	x,u	y	$u+0.5$	53	15
4	x,u	y	$x+0.5$	62	15
5	y,u	x	$u+0.5$	25	15
6	y,u	x	$y+0.5$	59	15
7	x,u	y	$x+0.5$，$u+0.5$	57	66
8	y,u	x	$y+0.5$，$u+0.5$	19	1.7

在表 5 的 1，2 行中，因算法 1 已经判断出 $x\to u$，$y\to u$ 是虚假因果关系，故不需要继续使用算法 2 和可信度验证函数来进一步判断. 而表 5 的 3～6 行中，算法 1，2 判断的结果均为真，故需使用可信度验证函数来进一步判断. 最终发现的因果关系为$\{x\to y,y\to x,u\to y,u\to x\}$，可见所有正确的因果关系都能被挖掘出来.

表 7 给出采用格兰杰方法对 x,y,u 三个要素两两进行因果分析的结果，原假设表示要素 1 不是要素 2 的原因，当 p 值小于 0.05 时拒绝原假设. 格兰杰因果分析时两个时间序列时滞为 1 个时间单位. 格兰杰方法发现的因果关系包括$\{x\to y,y\to x,u\to y\}$，格兰杰方法未发现的因果关系包括$\{u\to x\}$.

表 7 格兰杰因果判断结果

原假设	p 值
y 不是 x 的原因	3.E−07
x 不是 y 的原因	5.E−35
u 不是 x 的原因	0.2366
x 不是 u 的原因	0.1116
u 不是 y 的原因	6.E−32
y 不是 u 的原因	0.0527

表 8 给出了采用转移熵指标(Transfer Entropy, TE)对 x,y,u 三个要素两两进行因果分析的结果，如 $TE(x,y)=1.1185$，两个时间序列的时滞为 1 个时间单位. 对任意两个要素 a,b，我们认为当 $TE(a,b)>0.5$ 或 $TE(a,b)/TE(b,a)>5$ 时，表示 $a\rightarrow b$，即 a 是 b 的原因，其他基于转移熵的实验仍然以此为因果判定标准. 转移熵指标发现的因果关系包括$\{x\rightarrow y, y\rightarrow x, u\rightarrow y\}$，转移熵指标未发现的因果关系包括$\{u\rightarrow x\}$. 两种已有方法均未能发现因果关系 $u\rightarrow x$.

表 8 转移熵因果判断结果

转移熵(TE)	变量 x	变量 y	变量 u
变量 x		1.1185	0.0367
变量 y	0.5395		0.3928
变量 u	0.0083	1.1572	

5.2 在真实数据上进行的实验

5.2.1 box-jenkins 数据

这组数据来自文献[28]，是由 box-jenkins 在进行时间序列的分析时提出的，包含两个要素，即一个火炉中输入天然气流量 Gas 的时间序列和天然气燃烧后得到的 CO_2 浓度的时间序列，共 200 组数据. 显然这是一个动力系统，并且有因果关系 Gas→CO_2，数据的形式化定义为 $SCS=\langle m,k,T,E_{m\times k}\rangle$，$m=\{\text{Gas},CO_2\}$，$k=5$，$T=200$，$E_{m\times k}=\{\text{Gas}(t),\text{Gas}(t-1),\text{Gas}(t-2),\text{Gas}(t-3),\text{Gas}(t-4),\text{Gas}(t-5),CO_2(t),CO_2(t-1),CO_2(t-2),CO_2(t-3),CO_2(t-4),CO_2(t-5)\}$. 可知，该系统只有 2 个候选因果关系$\{\text{Gas}\rightarrow CO_2, CO_2\rightarrow\text{Gas}\}$，表 9 给出了如何通过算法 1,2 以及可信度验证函数来判断和验证该系统的因果关系. 表 9 的第 6 列“依据”表示通过了表 10 中的哪几行得到. 如第 1 行“1，Gas→CO_2，真，—，—，1,2”表示因果关系 Gas→CO_2 通过算法 1 直接判断为真，该结论通过表 10 中 1,2 行得到.

在表 9 第 1 行中，因系统只包含一个原因要素和一个结果要素，不存在噪声数据，故不需要使用算法 2 去噪，也不需使用可信度判断函数来判断结果可信性. 而在表 9 第 2 行中，由于算法 1 已经判断出 $CO_2\rightarrow$Gas 是虚假的因果关系，故不需要继续使用算法 2 和可信度验证函数来进一步判断. 很明显，算法 1 的结论是正确的，因为 Gas 燃烧是产生 CO_2 的原因.

表 9 实验结果

	候选因果关系	算法 1	算法 2	可信度验证	依据
1	Gas→CO_2	真	—	—	1,2
2	CO_2→Gas	假	—	—	3,4

表 10 扰动及波动结果

	自变量	因变量	扰动	波动	参考值
1	Gas	CO_2	Gas+0.2	1942	556545
2	Gas	CO_2	Gas+(−0.2)	373	556545
3	CO_2	Gas	CO_2+50	796	2199
4	CO_2	Gas	CO_2+(−50)	Infinity	2199

表 11 给出了格兰杰方法的分析结果，两个时间序列的时滞为 5 个时间单位，当 p 值小于 0.05 时拒绝原假设. 可以看到，格兰杰方法能够发现因果关系 Gas→CO_2.

表 11 格兰杰因果判断结果

原假设	p 值
Gas 不是 CO_2 的原因	5.E−39
CO_2 不是 Gas 的原因	0.8192

表 12 给出了转移熵指标的分析结果，两个时间序列的时滞为 5 个时间单位. 可以看到，转移熵指标能够发现因果关系 Gas→CO_2.

表 12 转移熵因果判断结果

转移熵(TE)	变量 Gas	变量 CO_2
变量 Gas		0.1854
变量 CO_2	0.0101	

5.2.2 空气污染数据

数据来自于 http://lib.stat.cmu.edu/datasets/NO2.dat，例 1 对其进行了说明，原始数据是包含 8 个要素的动力系统时间序列，但其中有 4 个要素与整个系统关系很小，故在数据预处理时将其去掉了. 为了保证时间单位一致，每一天只保留一个数据，共 100 组的数据. 形式化描述为 $SCS=\langle m,k,T,E_{m\times k}\rangle$，$m=\{\text{Car},\text{Temp},NO_2,\text{Wind}\}$，$k=1$，$T=100$，$E_{m\times k}=\{\text{Car}(t),\text{Car}(t-1),\text{Temp}(t),\text{Temp}(t-1),NO_2(t),NO_2(t-1),\text{Wind}(t),\text{Wind}(t-1)\}$. 可知，该系统共有 12 个候选因果关系$\{\text{Car}\rightarrow NO_2, \text{Temp}\rightarrow NO_2, \text{Car}\rightarrow NO_2, \text{Car}\rightarrow\text{Temp}, NO_2\rightarrow\text{Temp}, \text{Wind}\rightarrow\text{Temp}, \text{Car}\rightarrow\text{Wind}, NO_2\rightarrow\text{Wind}, \text{Temp}\rightarrow\text{Wind}, NO_2\rightarrow\text{Car}, \text{Temp}\rightarrow\text{Car}, \text{Wind}\rightarrow$

Car}. 表 13 给出了如何通过算法 1,2 以及可信度验证函数来判断和验证该系统的因果关系. 表 13 中算法 1 和 2 得出判断结论的实验数据见表 14 的 1～14 行. 表 13 中可信度验证函数得出判断结论的实验数据见表 14 的 15 行. 其中,表 13 的第 6 列"依据"表示通过了表 14 中的哪几行得到的表 13 中的结论. 如第 1 行"1,C→N,真,真,真,4,5,15"表示因果关系 C→N 通过算法 1,2 及可信度验证都被判断为真实的,判断过程通过表 14 中的 4,5,15 行得到. 第 3,6,12 行表示在进行 GEP 函数拟合时,箭头左边的自变量无法出现在拟合函数中,所以也就自动的被判断为虚假因果关系. 为节省篇幅将表 13、表 14 中 Car,Temp,NO_2,Wind 分别缩写为 C,T,N,W.

表 13　实验结果

	候选因果关系	算法 1	算法 2	可信度验证	依据
1	C→N	真	真	真	4,5,15
2	T→N	真	真	真	4,5,15
3	W→N	—	—	—	—
4	C→T	真	假	—	13,14
5	N→T	假	—	—	12
6	W→T	—	—	—	—
7	T→W	真	真	真	7,9,11
8	N→W	真	假	—	8,10
9	C→W	假	—	—	6
10	N→C	假	—	—	1
11	T→C	真	假	—	2,3
12	W→C	—	—	—	—

表 14　扰动及波动结果

	自变量	因变量	扰动	波动	参考值
1	N,T	C	N+(−3)	Infinity	4912
2	N,T	C	T+(−3)	117	4912
3	T	C	T+(−3)	7610394	4912
4	C,T	N	C+(−6)	6459	1439
5	C,T	N	T+(−3)	64	1439
6	C,T,N	W	C+(−6)	Infinity	37974
7	C,T,N	W	T+(−3)	79099	37974
8	C,T,N	W	N+(−3)	15909	37974
9	T,N	W	T+(−3)	6574	37974
10	T,N	W	N+(−3)	939678	37974
11	T	W	T+(−3)	3332	37974
12	C,W,N	T	N+(−3)	94300	163
13	C,W,N	T	C+(−6)	8.7	163
14	C,W	T	C+(−6)	6215	163
15	C,T	N	C+(−3), T+(−3)	118	11

在表 13 的 5,9,10 行中,由于算法 1 已经判断出这些关系是虚假的因果关系,故不需要继续使用算法 2 和可信度验证函数来进一步判断. 在表 13 的 4,8,11 行中,算法 1 判断出的结果为真,而算法 2 判断出的结果为假,则又将排除这几行中的因果关系,而不需要继续使用可信度验证函数来进一步判断. 表 8 的 1,2,7 行中,算法 1,2 判断出的结果均为真,需要继续用可信度验证函数验证. 验证过程由表 14 的 15 行表示. 可以看到,当同时对 Car 和 Temp 产生−3 的扰动时,NO_2 的波动幅度仍然近似符合因果准则,即虽然 $118>10\times11=110$,但是 118 与 110 实际上很接近,故仍然说明 Car→NO_2,Temp→NO_2 的因果关系是可信的.

最终得到的因果关系为{Car→NO_2,Temp→NO_2,Temp→Wind},由于大气中 NO_2 的一个重要来源是汽车排放,光化学烟雾的产生就有 NO_2 的参与,所以第 1 个因果关系是符合实际情况的. 对第 2 个因果关系,文献[29]证实了随大气温度的上升,NO_2 浓度有上升趋势. 对第 3 个因果关系,因为温度的变化会导致空气的流动从而产生风,该因果关系也是符合实际情况的.

表 15 给出了采用格兰杰方法对该数据集进行两两因果分析的结果,其中原假设表示要素 1 不是要素 2 的原因,当 p 值小于 0.05 时拒绝原假设. 格兰杰因果分析两个时间序列的时滞为 5 个时间单位. 格兰杰方法发现的因果关系只有{NO_2→Car},显然这不符合实际情况,因为公路上的车流量不会受到 NO_2 浓度的影响,而主要和是否为上下班高峰期有关. 其他几个因果关系格兰杰方法都未能发现.

表 15　格兰杰因果判断结果

原假设	p 值
CAR 不是 NO_2 的原因	0.4339
NO_2 不是 CAR 的原因	0.0168
TEMP 不是 NO_2 的原因	0.5871
NO_2 不是 TEMP 的原因	0.7320
WIND 不是 NO_2 的原因	0.8447
NO_2 不是 WIND 的原因	0.0679
TEMP 不是 CAR 的原因	0.8793
CAR 不是 TEMP 的原因	0.6872
WIND 不是 CAR 的原因	0.4338
CAR 不是 WIND 的原因	0.2061
WIND 不是 TEMP 的原因	0.7006
TEMP 不是 WIND 的原因	0.4715

表 16 给出采用转移熵(Transfer Entropy,TE)对数据两两因果分析的结果,两个时间序列时滞为 5 个时间单位. 转移熵指标未能发现任何因果关系.

表 16　转移熵因果判断结果

转移熵(TE)	变量 NO_2	变量 car	变量 temp	变量 wind
变量 NO_2		0.3460	0.2393	0.2813
变量 car	0.2700		0.2689	0.2412
变量 temp	0.3434	0.4006		0.3002
变量 wind	0.3284	0.3973	0.3003	

5.2.3 睡眠窒息症病人数据

实验旨在验证算法在不同数据规模上的效率和结果的一致性.数据来自文献[30],取 B2.dat 中前 5000 条记录,对数据进行了 0 均值和单位方差正规化.

该数据包含两个要素,分别是患有睡眠窒息症的病人的心率(Heart Rate, HR)和呼吸率(Breath Rate, BR)的时间序列数据,形式化描述为 $SCS=\langle m,k,T,E_{m\times k}\rangle$, $m=\{\mathrm{HR},\mathrm{BR}\}$, $k=2$, $T=5000$, $E_{m\times k}=\{\mathrm{HR}(t),\mathrm{HR}(t-1),\mathrm{HR}(t-2),\mathrm{BR}(t),\mathrm{BR}(t-1),\mathrm{BR}(t-2)\}$.图 7~图 9 给出了实验结果,从中可得出以下结论.

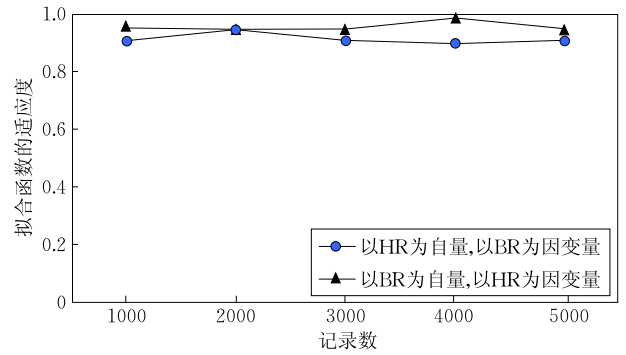


图 7 GEP 拟合函数的适应度

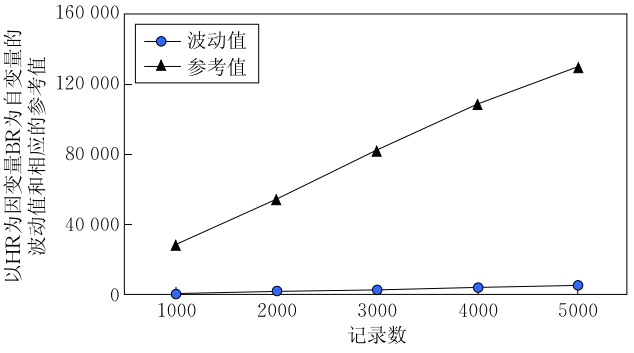


图 8 以 HR 为因变量进行扰动的结果

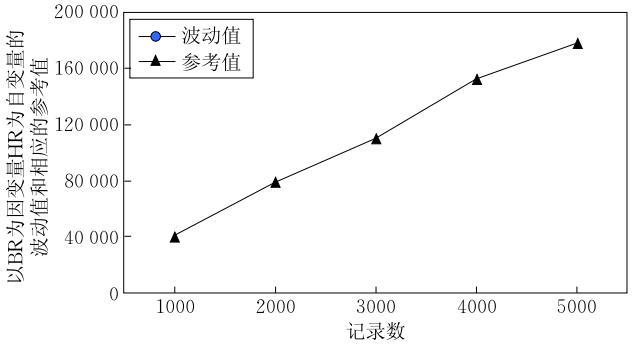


图 9 以 BR 为因变量进行扰动的结果

(1) GEP 对不同规模的数据均能拟合出适应度很高的函数.图 7 是当记录数从 1000~5000 时,分别以 HR 和 BR 为因变量,以另一要素为自变量进行函数拟合,得到的平均适应度与最优适应度的比值.可知在不同规模下,GEP 均能达到很高适应度,这肯定了算法基于 GEP 进行函数拟合的可用性.

(2) 算法的挖掘结果对不同规模的数据具有一致性.图 8、图 9 表示记录数从 1000 条逐步增加到 5000 条时,分别对以 HR 和 BR 为因变量的拟合函数进行扰动后得到的因变量的波动值以及相应的参考值,参考值表示由自变量拟合出的关于因变量的函数在整个时间序列上的最大函数值与最小函数值差值的平方.从图 8 可知,当以 HR(t)为因变量时,均有波动值小于参考值.而在图 9 中由于每次实验得到的波动值都趋近于无穷大,故只有参考值随记录数增大的变化曲线,当以 BR(t)为因变量时,均有波动值≫参考值,由定理 2 排除因果关系 HR→BR.故各种数据规模下算法均可得到因果关系 BR→HR.由于患有睡眠窒息症病人的呼吸是不正常的,故导致心率变化,故 BR 影响 HR 的结论是合理的.

6 结论及进一步工作

实验结果表明对 5.1 节中的合成数据,本文所提出的方法能够挖掘出所有正确的因果关系(因为数据是合成的,所以正确因果关系就是已知的),而格兰杰因果分析方法和基于转移熵的方法均少发现了一对因果关系.由于 5.2.1 节中的实验数据只包含两个要素,所以本文的算法和两种对比算法均找出了正确的因果关系.值得指出 5.2.2 节中的数据是包含 4 个要素的空气污染数据,本文的算法所发现的 3 个因果关系的正确性在相关领域有专业的文献支持,或者是目前公认的环境科学的一些结论.而格兰杰因果分析方法只得到了一个因果关系,并且能够易知是不符合实际情况的,转移熵方法则未能找出该数据集中的任何因果关系.

综上,本文提出的算法能够正确地挖掘出亚复杂动力系统各要素间的因果关系,特别是当系统要素大于两个时,本文的方法仍然能够得到正确的结论,而概率方法则不行.

从实验分析中可以看出本文算法在以下几个方面具有较好的应用效果:首先,算法可以应用于气象复杂系统的因果分析中,5.2.2 节只是对简单的气象要素进行了因果分析,还可以进一步对复杂气象要素间的因果关系进行分析,比如气象要素间的遥相关分析,气象极值要素之间的因果分析等.其次,由于生物体本身也是一个复杂系统,故算法还可以

对生物信号进行因果分析，5.2.3 节对人体体征要素的时间序列进行了因果分析，还可对人体蛋白质序列，脑电波序列等各种生物信号要素进行分析.

下一步工作主要是进一步提高抗噪算法对噪声要素的识别精度以及将因果分析结论运用到亚复杂动力系统的干预分析中.

参 考 文 献

[1] Tang Chang-Jie, Zhang Yue, Tang Liang, et al. A survey on mining kinetic intervention rule from sub-complex systems. Journal of Computer Applications, 2008, 28(6): 2732-2736 (in Chinese)
(唐常杰，张悦，唐良等. 亚复杂系统中动力学干预规则挖掘技术研究进展. 计算机应用，2008，28(6)：2732-2736)

[2] Ashley R, Granger C W J, Schmalensee R. Advertising and aggregate consumption: An analysis of causality. Econometrica, 1980, 48(5): 1149-1167

[3] Ostermark R, Aaltonen J. Comparison of univariate and multivariate Granger causality in international asset pricing: Evidence from finnish and Japanese financial economies. Applied Financial Economics, 1999, 9(2): 155-165

[4] Mascherini M, Stefanini F. M-ga: A genetic algorithm to search for the best conditional gaussian bayesian network// Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce. Vienna, Austria, 2005: 61-67

[5] Wong M L, Leung K S. An effecient data mining method for learning Bayesian networks using an evolutionary algorithm-based hybrid approach. IEEE Transactions on Evolutionary Computation, 2004, 8(4): 378-404

[6] Gelper S, Aurelie L, Christophe C. Consumer sentiment and consumer spending: Decomposing the Granger causal relationship in the time domain. Applied Economics, 2007, 39 (1): 1-11

[7] Nielsen U H, Pellet J P, Elisseeff A. Explanation trees for causal bayesian networks//Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence. Helsinki, Finland, 2008: 427-434

[8] Ay N, Polani D. Information flows in causal networks. Advances in Complex Systems, 2008, 11(1): 17-41

[9] Yuan C, Lu T C. Finding explanations in bayesian networks//Proceedings of the 18th International Workshop on Principles of Diagnosis. Nashville, USA, 2007: 414-419

[10] Rosenblum M G, Pikovsky A S. Detecting direction of coupling in interacting oscillators. Physical Review E, 2001, 64(10): 1-4

[11] Arnhold J, Grassberger P, Lehnertz K. A robust method for detecting interdependences: Application to intracranially recorded EEG. Physica D: Nonlinear Phenomena, 1999, 134(4): 419-430

[12] Darbellay G A. An estimator of the mutual information based on a criterion for independence. Computational Statistics, 1999, 32(1): 1-17

[13] Palus M. Detecting nonlinearity in multivariate time series. Physics Letters A, 1996, 213(3): 138-147

[14] Schreiber T. Measuring information transfer. Physical Review Letters, 2000, 85(2): 461-464

[15] Bessler D, Yang J, Wongcharupan M. Price dynamics in the international wheat market: Modeling with error correction and directed acyclic graphs. Journal of Regional Science, 2003, 43(1): 1-33

[16] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. New York: MIT Press, 2000

[17] Drton M, Richardson T S. Graphical answers to questions about likelihood inference for gaussian covariance models. Washington: Department of Statistics, University of Washington, Technical Report: 467, 2004

[18] Franaszczuk P J, Bergey G K. Application of the directed transfer function method to mesial and lateral onset temporal lobe seizures. Brain Topography, 1998, 11(1): 13-21

[19] Korzeniewska A, Kasicki S, Kamiński M. Information flow between hippocampus and related structures during various types of rat's behavior. Journal of Neuroscience Methods, 1997, 73(1): 49-60

[20] Faes L, Porta A, Cucino R. Causal transfer function analysis to describe closed loop interactions between cardiovascular and cardiorespiratory variability signals. Biological Cybernetics, 2004, 90(6): 390-399

[21] Nollo G, Faes L, Porta A. Exploring directionality in spontaneous heart period and systolic pressure variability interactions in humans: Implications in the evaluation of baroreflex gain. American Journal of Physiology-Heart and Circulatory Physiology, 2005, 288(4): 1777-1785

[22] Sachs K, Perez O, Peter D. Causal protein-signaling networks derived from multiparameter single-cell data. Science, 2005, 308(5721): 523-529

[23] Pearl J. Causality: Models, Reasoning and Inference. Cambridge: MIT Press, 2000

[24] Shi X, Fan W, Zhang J. Discovering shakers from evolving entities via cascading graph inference//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 1001-1009

[25] Snowsill T M, Fyson N, De Bie T, Cristianini N. Refining causality: Who copied from whom?//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 466-474

[26] Liu W, Zheng Y, Chawla S. Discovering spatio-temporal causal interactions in traffic data streams//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 1010-1018

[27] Ferreira C. Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence: Studies in Computational Intelligence. New York: Springer-Verlag, 2006

[28] Box G E P, Jenkins G M. Time series analysis: Forecasting and control. San Francisco: Holden-Day, 1970

[29] Zuo Hao-Yi, Gao Jie, Cheng Juan. Measurement of NO_2 concentration in atmosphere in Chengdu by solar spectra. Spectroscopy and Spectral Analysis, 2006, 26(7): 1356-1359 (in Chinese)
(左浩毅, 高洁, 程娟. 太阳光谱方法测量成都地区大气二氧化氮浓度. 光谱学与光谱分析, 2006, 26(7): 1356-1359)

[30] Makridakis S. Time Series Prediction: Forecasting the Future and Understanding the Past. MA, USA: Addison-Wesley Publishing Company, 1994

附录 1. 正文中定理 1 和定理 3 的证明过程.

定理 1. 设动力系统含有两个要素 x,y,其中 x 影响 y,且有 $y=f(x)$;x 是随机时间序列,且时间序列的长度为 T,通过函数拟合,得到 $x=g(y)$,则当 δ_x 和 δ_y 取较大值时有:

(1) $|\delta f_y(x,\delta_x,t_0)|<C\times|\mathrm{Max}(f(x))-\mathrm{Min}(f(x))|$;

(2) $|\delta g_x(y,\delta_y,t_0)|\gg C\times|\mathrm{Max}(g(y))-\mathrm{Min}(g(y))|$;

(3) $\sum f_y(x,\delta_x,t)<C\times T\times|\mathrm{Max}(f(x))-\mathrm{Min}(f(x))|^2$;

(4) $\sum g_x(y,\delta_y,t)\gg C\times T\times|\mathrm{Max}(g(y))-\mathrm{Min}(g(y))|^2$.

证明. (a) 对于(1)、(2). 因假设 1 中已指出 $f(x)$, $g(y)$在其自变量范围内具有各阶导数,并且都是有界的,所以对 x 和 y 的任意取值 x_0 和 y_0,$f(x)$和 $g(y)$可以展开成 x_0 和 y_0 的泰勒级数

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+f''(x_0)(x-x_0)^2/2!+\cdots+f^{(n)}(x-x_0)(x-x_0)^n/n!+\cdots \quad (1)$$

$$g(y)=g(y_0)+g'(y_0)(y-y_0)+g''(y_0)(y-y_0)^2/2!+\cdots+g^{(n)}(y-y_0)(y-y_0)^n/n!+\cdots \quad (2)$$

式(1),(2)去掉 2 阶以上的可以忽略的数量值可得

$$f(x)=f(x_0)+f'(x_0)(x-x_0) \quad (3)$$

$$g(y)=g(y_0)+g'(y_0)(y-y_0) \quad (4)$$

又根据函数扰动的定义有

$$\delta f_y(x,\delta_x)=f_y(x+\delta_x)-f_y(x) \quad (5)$$

$$\delta g_x(y,\delta_y)=g_x(y+\delta_y)-g_x(y) \quad (6)$$

在式(5)、(6)中为了描述简洁,去掉了函数扰动定义中的时间 t_0,因为 t_0 仅表示某一特定时刻自变量的取值,所以与证明无关. 将式(3)、(4)分别代入式(5)、(6)得

$$\delta f_y(x,\delta_x)\approx f(x_0)+f'(x_0)(x+\delta_x-x_0)-f(x_0)+f'(x_0)(x-x_0) \quad (7)$$

$$\delta g_x(y,\delta_y)\approx g(y_0)+g'(y_0)(y+\delta_y-y_0)-g(y_0)+g'(y_0)(y-y_0) \quad (8)$$

将式(7)、(8)整理后可得

$$\delta f_y(x,\delta_x)\approx f'(x_0)\delta_x \quad (9)$$

$$\delta g_x(y,\delta_y)\approx g'(y_0)\delta_y \quad (10)$$

根据假设 1 及式(9)、(10),可知定理 1(1)、(2)成立.

(b) 对于(3)、(4). 因为$\sum f_y(x,\delta_x,t)=\delta f_y(x,\delta_x,t_1)^2+\cdots+\delta f_y(x,\delta_x,t_T)^2$,并且$\delta f_y(x,\delta_x,t)<C\times|\mathrm{Max}(f(x))-\mathrm{Min}(f(x))|$.

则 $\delta f_y(x,\delta_x,t)^2<C\times|\mathrm{Max}(f(x))-\mathrm{Min}(f(x))|^2$,故有$\sum f_y(x,\delta_x,t)<C\times T\times|\mathrm{Max}(f(x))-\mathrm{Min}(f(x))|^2$,同理有$\sum g_x(y,\delta_y,t)\gg C\times T\times|\mathrm{Max}(g(y))-\mathrm{Min}(g(y))|^2$,故定理 1(3)、(4)成立. 证毕.

定理 3. 设有亚复杂动力系统 $SCS=\langle m,k,T,E_{m\times k}\rangle$,有任意两个要素 $x,y\in m$, x 是随机时间序列,y 是受 x 影响的因变量,设 $y=f_0(x)$是 x,y 真正符合的函数,$y=f_1(x)$是采用 GEP 生成的任意拟合函数,则当拟合精度足够高时,有

(1) $|f_0'(x)|\approx|f_1'(x)|$, $f_0'(x)$,$f_1'(x)$分别是 $f_0(x)$, $f_1(x)$的导数.

(2) 若 $\sum f_0(x,\delta_x,t)<C\times T\times|\mathrm{Max}(f_0(x))-\mathrm{Min}(f_0(x))|^2$,则$\sum f_1(x,\delta_x,t)<C\times T\times|\mathrm{Max}(f_1(x))-\mathrm{Min}(f_1(x))|^2$,$\delta_x$是 x 时间序列平均值.

证明. (1)ε-拟合函数定义有式(11)、(12),因 ε 很小,有式(13)

$$|(f_0(x)-f_1(x))/f_0(x)|<\varepsilon \quad (11)$$

$$|(f_0(x+\delta)-f_1(x+\delta))/f_0(x+\delta)|<\varepsilon \quad (12)$$

$$|(f_0(x)-f_1(x))/f_0(x)|\approx|(f_0(x+\delta)-f_1(x+\delta))/f_0(x+\delta)| \quad (13)$$

不失一般性,令所有绝对值符号中的值均大于 0,则式(13)可得式(14)、(15),通过恒等变形可得式(16),通过式(16)可得式(17),由假设 2 知$|f_0(x)-f_1(x)|<\varepsilon$,则通过式(17)可得$|f_0'(x)|\approx|f_1'(x)|$,命题(1)得证.

$$f_1(x+\delta)f_0(x)-f_1(x)f_0(x)\approx f_1(x)f_0(x+\delta)-f_1(x)f_0(x) \quad (14)$$

$$f_0(x)(f_1(x+\delta)-f_1(x))\approx f_1(x)(f_0(x+\delta)-f_0(x)) \quad (15)$$

$$f_0(x)(f_1(x+\delta)-f_1(x))/\delta\approx f_1(x)(f_0(x+\delta)-f_0(x))/\delta \quad (16)$$

$$F_0(x)\,f_1'(x)\approx f_1(x)f_0'(x) \quad (17)$$

(2) 由定义 2 有$\sum f_0(x,\delta_x,t)=\delta f_0(x,\delta_x,t_1)^2+\cdots+\delta f_0(x,\delta_x,t_T)^2$,因为定理 1 已经证明$\sum f_0(x,\delta_x,t)<C\times T\times|\mathrm{Max}(f_1(x))-\mathrm{Min}(f_1(x))|^2$,故 $\delta f_0(x,\delta_x,t_1)^2+\cdots+\delta f_0(x,\delta_x,t_T)^2<C\times T\times|\mathrm{Max}(f_1(x))-\mathrm{Min}(f_1(x))|^2$. 根据定理 1 中式(9)有 $\delta f_0(x,\delta_x)\approx f'(x_0)\delta_x$. 又因为(1)已证明$|f_0'(x)|\approx|f_1'(x)|$故 $\delta f_0(x,\delta_x)\approx f'(x_0)\delta_x\approx f_1'(x)\times\delta_x\approx\delta f_1(x,\delta_x)$. 故 $\delta f_0(x,\delta_x,t_1)^2+\cdots+\delta f_0(x,\delta_x,t_T)^2\approx\delta f_1(x,\delta_x,t_1)^2+\cdots+\delta f_1(x,\delta_x,t_T)^2$. 因为 $\delta f_0(x,\delta_x,t_1)^2+\cdots+\delta f_0(x,\delta_x,t_T)^2<C\times T\times|\mathrm{Max}(f_1(x))-\mathrm{Min}(f_1(x))|^2$,故 $\delta f_1(x,\delta_x,t_1)^2+\cdots+\delta f_1(x,\delta_x,t_T)^2<C\times T\times|\mathrm{Max}(f_1(x))-\mathrm{Min}(f_1(x))|^2$,故$\sum f_1(x,\delta_x,t)<C\times T\times|\mathrm{Max}(f_1(x))-\mathrm{Min}(f_1(x))|^2$,命题(2)得证. 证毕.

ZHENG Jiao-Ling, born in 1981, Ph.D., associate professor. Her research interests include database and knowledge engineering, complex system, social computing.

TANG Chang-Jie, born in 1946, professor, Ph.D. supervisor. His research interests include database and knowledge engineering, data mining.

QIAO Shao-Jie, born in 1981, Ph.D., associate professor. His research interests include database and knowledge engineering, data mining.

YANG Ning, born in 1974, Ph.D., lecturer. His research interests include database and knowledge engineering, evolutionary computing.

LI Chuan, born in 1974, Ph.D., associate professor. His research interests include database and knowledge engineering.

CHEN Yu, born in 1974, Ph.D., lecturer. His research interests include database and knowledge engineering, evolutionary computing.

WANG Yue, born in 1981, Ph.D., lecturer. His research interests include database and knowledge engineering, machine learning.

Background

This study tries to find causal relationships between elements in sub-complex dynamic systems. Traditional causality analyzing methods are based on probability models with predefined distributions. The probability based methods have difficulty when dealing with dynamic systems. This paper tries to mine reliable causality from sub-complex system based on dynamic system's perturbation theory.

This study is supported by the National Natural Science Foundation of China (Nos. 61202250, 61203172, 61100045), Youth Foundation of Sichuan Educational Committee (No.11ZB088), Research Fund for Young Academic Leaders of Chengdu University of Information Technology (No. J201208), the Introduction of Talent Project of Chengdu University of Information Technology (No. KYTZ201110), the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20110184120008), and the Youth Foundation for Humanities and Social Sciences of Ministry of Education of China (No. 14YJCZH046).

Social community large scale cooperation cognition rules mining targets at mining social community's large scale cooperation rules by integrating the existing achievements in cognitive science with intelligent computing. This project tries to deal with NP-Hard problems by utilizing the complex system's emergence phenomenon.

The theoretical results of this paper can provide solid foundation for the projects' further research.