高技术通讯 2024 年 第34 卷 第6 期:578-589

doi:10.3772/j.issn.1002-0470.2024.06.003

时序数据的因果关系交互式可视分析①

丁伟杰②[*,**,******] 华 东[***] 袁 莹[**] 孙国道③[****] 尤芷芊[*****] 梁荣华[****]

([*]浙江工业大学信息工程学院 杭州 310023)

([**]浙江警察学院计算机与信息安全系 杭州 310053)

([***]浙江省公安厅情报指挥中心 杭州 310053)

([****]浙江工业大学计算机科学与技术学院 杭州 310023)

([*****]香港理工大学工程学院 香港 999077)

([******]基于大数据架构的公安信息化应用公安部重点实验室 杭州 310053)

摘 要 针对时序数据中因果关系检测算法效率低、错误率高、可解释性低的问题,本文提出一种新颖的用于时序数据的因果关系检测模型。该模型整合了泛函贪婪等价搜索(F-GES)模型与格兰杰(Granger)因果关系模型,展开因果关系的抽取和推断,并提出了因果关系可视分析方法,以交互式地分析时序数据中变量间的因果关系。可视分析方法形成了参数视图用于提高因果关系探索效率、因果关系树图用于直观有效地展示变量之间的因果关系、时间视图用于比较原始时序数据、堆叠流图用于帮助用户探索时序数据的层次演变以及平行坐标图用于进行相关性分析。基于真实数据形成的原型系统交互式地验证和总结时序数据中的因果关系,从而更高效地挖掘和理解时序数据中变量之间蕴含的因果规律以帮助决策。

关键词 因果关系;时间序列;可视分析;产业链

随着信息技术的变革和大数据技术的逐步成熟,时间序列的尺度和维度不断增大。变量之间互相影响,在医学、金融、教育等领域,通过检测海量的多元时序数据之间的因果关系,挖掘有用的信息、探索事物发生机制[1]以帮助决策具有重要的现实意义。挖掘事物之间深层次的关系,促使了因果关系分析的参与。因果关系分析的目的是从观察数据中推断事物产生的原因和结果。由于因果关系检测难度高以及成本高等问题,现有的工作缺乏对时序数据的因果关系研究。然而,研究者对时间序列数据的研究偏向于对相关性的探索和改进,一些研究采用基于相关性的分析来得出因果结论的方法是不合理的。相关性不是因果关系,相关性揭示了事物之间的关联关系,而因果关系揭示了事物之间更深层次的依赖和影响关系。一些研究者使用可视分析方法探索和挖掘某些特定数据之间的关系,然而,缺乏对时序数据的因果关系的可视分析。

目前,该领域所面临的主要挑战主要来自因果关系检测算法和可视呈现这 2 个方面。首先,针对因果关系检测算法,格兰杰(Granger)因果关系分析方法是针对时序数据的因果检测最广泛应用的方法,具有很强的可解释性,但是无法给出定量的因果强度。因此,相关研究者们提出了大量改进模型[2],更加精确的因果强度检测算法导致计算效率低,无法实时更新数据,导致无法加入人工修正的结果。所以为了提高运算速度,泛函贪婪等价搜索

① 教育部人文社会科学规划课题(22YJA840004)资助项目。

② 男,1981 年生,博士,副教授;研究方向:大数据分析及可视化,网络犯罪治理;E-mail:dingweijie@zjjcxy.cn。

③ 通信作者,E-mail:guodao@zjut.edu.cn。

(收稿日期:2023-02-13)

(functional greedy equivalence search,F-GES)算法[3]引入假设和并行计算技术来处理复杂的高维数据的因果关系检测。本文将 F-GES 模型与 Granger 因果关系模型相结合,在一定程度上缓解了因果关系检测算法效率低和可解释性低的问题。同时,数据驱动的因果关系检测方法会出现一定数量的错误率,针对此问题,本文设计了交互式因果关系可视分析系统,将人工反馈加入到算法检测的结果中,以提高正确率。因此,针对因果关系可视化,研究者们设计了一系列的因果关系可视隐喻来帮助用户进行高效的因果分析。最直观的方法是采用点线图表达因果关系,其中节点或者定制的图元表示一个变量或者变量的属性信息,连接表示 2 个标量之间存在因果关系,箭头表示因果关系的方向。点线图的重点是底层布局算法,研究者们针对不同的应用场景提出了大量的底层布局算法使得因果关系可视化更直观、高效、美观和丰富。

针对时序数据中因果关系分析效率低、错误率高和依赖自动算法提取因果关系存在可解释性低、关系错综复杂难以记忆等问题,本文提出一种新颖的用于时序数据的因果关系检测模型,该模型将 F-GES 模型与 Granger 因果关系模型相结合进行因果推断。同时为了更有效地分析和验证时序数据的因果关系,本文针对性地提出了因果关系可视分析方法,交互式地分析时序数据中变量间的因果关系,辅助数据深入洞察。可视分析方法形成了参数视图用于提高因果关系探索效率、因果关系树图用于直观有效地展示变量之间的因果关系、时间视图用于比较原始时序数据、堆叠流图用于帮助用户探索时序数据的层次演变以及平行坐标图用于进行相关性分析。基于真实数据形成的原型系统以交互方式验证和总结时序数据中的因果关系,利用可视分析技术实现对因果关系进行修正和可视化总结,以引导用户更高效地理解时序数据中的因果关系,填补自动检测因果关系算法的不足,帮助用户对时序数据中因果关系的归纳与应用。由于产业链上下游产品具有真实的因果关系,所以,为了验证本文提出的因果关系模型的有效性,本文基于真实的产业链数据,交互式地验证和总结时序数据的因果关系。

1 相关工作

1.1 时序数据因果关系算法

伴随机器学习的热潮,因果关系成为热门话题。时间序列数据之间存在复杂的关联关系,蕴含着大量的因果关系。检测多元时序数据之间的因果关系,发挥数据的潜在价值,对大数据在市场营销和医疗卫生等方面的应用具有重要的现实意义[4]。针对时间序列数据的因果关系研究,研究者们提出了大量的因果关系检测模型。Ren 等人[4]归纳了时序数据研究中主要的因果分析方法,包括 Granger 因果关系分析、基于信息理论的因果分析和基于状态空间的因果分析,其中由于 Granger 因果关系检测方法具有较强的可解释性,因此该算法成为了应用最广泛的方法。Granger 因果关系检测方法能够评估 2 个变量时间序列之间是否存在相互作用的因果关系。Granger 因果测试的结论是统计推断,能推断出 X 对 Y 的预测是有帮助的,但不能确定 X 和 Y 是否是因果关系。很多学者们根据场景特点,采用不同的数学统计检测方法来判断变量之间的因果关系。但是 Granger 因果关系检测方法只能给出定性的因果关系分析结果,即是否具有因果关系,不能检测变量之间具体的因果强度。又有很多学者们提出了用于检测因果强度的算法模型,来检测变量之间具体的因果强度[3]。

大部分因果关系检测的算法都是基于条件独立性[5]检测来挖掘变量之间的因果强度关系。文献[6]提出 GES 贪心搜索方法来解决网络搜索问题,并提出了评分函数作为因果网络的似然估计。然而,这些基于条件独立的方法效率低下,可扩展性低,无法适应于复杂的大规模数据。针对因果关系算法计算速度慢的问题,文献[3]通过引入假设和并行计算技术,提出了 F-GES 算法,解决了复杂的高维数据的因果关系检测过程中时间复杂度高的问题。

在探究因果关系时,可视化是非常重要的工具。因果关系往往涉及到多个变量之间的复杂交互,可视化可以帮助用户直观地展现这些关系,帮助用户更好地理解因果关系的本质。通过可视化工具,用

户可以将数据呈现为图表、图形和图像等形式,从而更清晰地观察和解释数据。这可以帮助用户更快速地发现因果关系中的问题和矛盾,以及减少数据解释的误差。可视化也能够帮助用户更好地进行因果推断。在因果关系检测过程中,用户需要识别哪些变量是原因,哪些变量是结果。通过可视化,用户可以更清晰地展现变量之间的关系,从而帮助用户确定原因和结果。同时,可视化工具也能够帮助用户更好地进行数据清洗和数据准备,减少因数据错误和噪音对结果的影响。除此之外,可视化还能够提高因果关系分析的可解释性。因果关系图的节点和边缘都可以进行标注和注释,使得用户能够更好地解释分析结果,并向其他人传达用户的分析和结论,这有助于提高因果关系分析的可信度和说服力。因此,可视化在因果关系检测中扮演了非常重要的角色。它能够帮助用户更好地理解因果关系、进行因果推断、减少数据解释误差,并提高分析的可解释性。可视化工具的运用,能够使用户更准确、更高效地发现因果关系,从而更好地进行决策和规划。

1.2 因果关系可视化

因果关系通常采用有向无环图(directed acyclic graph, DAG)表示,其中节点表示一个变量,连接表示 2 个标量之间存在因果关系,箭头表示因果关系的方向。随着多元数据的复杂度和关联性增加,学者开始从相关性分析迈入到因果关系的探究分析中,同时结合可视化分析对因果关系进行修正和解释。文献[7]提出了用动画、颜色和图案来表示因果关系的可视化技术,便于快速浏览。文献[8]展示了因果图(ICG)的交互式可视化,帮助用户理解大型数据集中的关系。文献[9]提出了一种用于基于时间的图形的焦点和上下文交互技术,它允许连续或离散地改变细节。文献[10]开发了一种用于表格数据因果分析的交互式可视化界面。文献[11]提出了二维图形可视化的因果网络和一套交互式工具,用户可以交互式地访问和验证由自动模型检测到的因果关系,允许用户对特定的子数据集进行因果探究。文献[12]提出使用文本叙述作为一种数据驱动的讲故事方法来增强因果关系的可视化。近些年来,国内也有很多可视化领域的专家和学者对因果关系进行研究。文献[13]介绍了一种基于地理空间数据的因果分析方法,包括一个可视化交互分析系统,允许用户对地理位置相关的隐含关系进行探索分析。文献[14]介绍了一种用于推测事件序列数据中的因果关系的联合检验方法,将时间序列数据和因果关系作为识别模型的层次,有效提升了对因果关系的识别性能。文献[15]提出了基于格兰杰因果关系的可视分析方法 Compass,用于检测城市因果关系、解释动态因果关系和修正因果关系。

本工作针对因果关系检测算法和可视呈现这 2 个方面展开研究,尝试缓解因果关系分析中存在的效率低下、可解释性低、错误率高、关系复杂等问题。具体地,本工作提出了一种新颖的因果关系检测模型,将 F-GES 模型和 Granger 因果关系模型相结合,以提高因果推断的效率和准确性;同时,本工作还设计了一系列因果关系可视隐喻,包括点线图、因果关系树图、时间视图、堆叠流图和平行坐标图等,以帮助用户更直观地理解和分析数据之间的因果关系。最终,通过真实的产业链数据验证了本工作提出的因果关系模型和可视化方法的有效性。

2 因果关系检测算法模型

本节主要将 F-GES 模型与 Granger 因果关系模型相结合,检测时序数据中变量之间的因果关系。

2.1 因果关系形式化定义

传统的因果的形式化定义:有向无环图(DAG),它提供了一种直观地表示和更好地理解因果关系、偏差等关键概念的方法,可以用于标记因果顺序[16]。

因果图是有向无环图的扩展,每个结构因果机制(structural causal models, SCM)可以表示为一个 DAG,其中一个节点代表一个数据维度,一个链接代表 2 个相连维度之间的依赖关系。因果图定义为 $G=(V,E)$,其中 V 代表节点、E 代表边,即因果关系,箭头的方向表示了因果关系的方向。例如,一个因果图中含有 3 个变量 X、Y、Z,根据皮尔森(Pearson)相关系数可以得到结论,如果 X 和 Y 之间不存

在边,则以 Z 为条件时 X 和 Y 独立。

2.2 因果推断模型

为了高效和准确地检测高维时序数据的因果关系,本文基于 F-GES 模型[3]检测高维数据的因果关系,采用贪心搜索算法来提高检测效率,同时结合 Granger 因果关系来提高因果关系检测模型的精确度。

F-GES 是一种基于分数的因果检测方法,每增加或减少一条边(即因果关系)会有一项分数来衡量因果图数据分布的程度,选择贝叶斯信息准则(Bayesian information criterion, BIC)进行评分,它能够有效地对时间序列的因果关系进行结构学习[17]。

本文采用 BIC 分数评价,如式(1)所示。

$$BIC = \ln(n)k - 2\ln(L) \tag{1}$$

式中, n 为样本量, k 为参数个数, $L = P(X \mid G)$ 为最大似然值。评分标准主要由因果图结构复杂性的惩罚和因果图与数据样本之间的适合度构成。该过程采用贪心搜索算法来提高检测效率。由此得到一个初始的有向无环图,如图 1 所示。

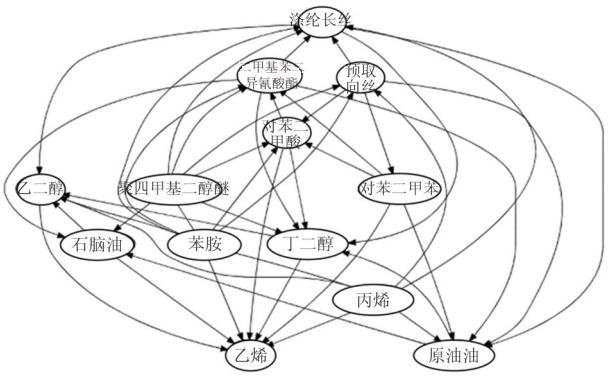


图 1 F-GES 检测的因果关系输出

以上基于有效成分数确定的方法不可避免会出现数据之间作用和抵消的现象,故因果关系图存在一定程度的误差。因此,本文基于 Granger 因果检验来进一步验证变量之间的因果关系。根据前文得到的因果关系,对每个因果关系建立双变量模型。以时间序列 X 和 Y 为例,检验 $x \rightarrow y$, 即 x 导致 y, 基于式(2)和(3)。

$$x_t = c_1 + \sum_{i=1}^{3}(\alpha_{1,i}y_{t-i}) + \sum_{i=1}^{3}(\beta_{1,i}x_{t-i}) + \epsilon_{x,t} \tag{2}$$

$$y_t = c_2 + \sum_{i=1}^{3}(\alpha_{2,i}y_{t-i}) + \sum_{i=1}^{3}(\beta_{2,i}x_{t-i}) + \epsilon_{y,t} \tag{3}$$

其中, c_1、c_2 是常数项,用于调整模型的基准; x_{t-i}、y_{t-i} 分别表示变量 x、y 的历史数据; $\alpha_{1,i}$ 和 $\alpha_{2,i}$ 是自回归系数; $\beta_{1,i}$ 和 $\beta_{2,i}$ 是交叉回归系数; $\epsilon_{x,t}$ 和 $\epsilon_{y,t}$ 是误差项。

由于时间序列存在滞后性,通过给定滞后参数 l 分别对全模型 $M_f^{[i,j]}$ 和简化模型 $M_r^{[i,j]}$ 2 个回归模型进行 Granger 检验, $M_f^{[i,j]}$ 和 $M_r^{[i,j]}$ 定义如式(4)和(5)所示。

$$y_t \sim \sum_{l=1}^{L} a_l \times y_{t-l} + \sum_{l=1}^{L} b_l \times x_{t-l} \tag{4}$$

$$y_t \sim \sum_{l=1}^{L} a_l \times y_{t-l} \tag{5}$$

其中, a_l 为是自回归系数, b_l 是交叉回归系数,分别表示变量 x、y 的历史数据。

可以看出,全模型使用了 X 和 Y 的滞后值,简化模型仅使用 Y 的滞后值。如果检验结果为 X 导致 Y, 多数情况下,全模型提供的预测会比简化模型更准确。

Granger 因果测试的结论是统计推断,能推断出 X 对 Y 的预测是有帮助的,但不能确定 X 和 Y 是否是因果关系。所以本文将其用于检验推断出的因果结论的可靠性。Granger 因果检验有多个数学统计检测方法,针对本文的时序数据连续的特性采用 F 检验(F-Test)数学检验方法。

$$F = \frac{(RSS_{\text{red}} - RSS_{\text{full}})/(r-s)}{RSS_{\text{full}}/(T-r)} \tag{6}$$

其中, RSS_{red} 和 RSS_{full} 分别是具有 r 和 s 参数的完整模型和简化模型的剩余平方和, T 是样本量。对于 F 检验,首先给定时间序列的间隔 $[i, j]$ 和对应的数据集 $\{y_k, x_{k1}, \cdots, x_{kp}\}_{j=i}^{j}$, 用 $M^{[i,j]}$ 表示普通最小二乘拟合得到的回归模型, $M^{[i,j]}$ 的误差平方和如式(7)所示。

$$SSE(M^{[i,j]}) = \sum_{t=1}^{T}(y_t - y_t')^2 \tag{7}$$

其中, y_t' 为 $M^{[i,j]}$ 根据 y_t 计算出的预测值, T 为时间序列的长度, F 检验如式(8)所示。

$$F = \frac{(SSE_r - SSE_f)/(d_f - d_r)}{SSE_f/(T - d_f - 1)} \tag{8}$$

其中, d_f 和 d_r 分别为 $M_f^{[i,j]}$ 和 $M_r^{[i,j]}$ 的模型自由度,即自变量的数量。本文算法实现过程中,将 d_f 和 d_r

分别设定为 $2L$ 和 L，所以时间区间 $[i:j+\delta]$ 的 F 检验统计值上界如式(9)所示。

$$\lceil F^{[i:j+\delta]} \rceil = \frac{(\lceil SSE(M_r^{[i:j+\delta]}) \rceil - \lfloor SSE(M_f^{[i:j+\delta]}) \rfloor)/L}{\lfloor SSE(M_f^{[i:j+\delta]}) \rfloor/((j+\delta-i+1)-2L-1)} \tag{9}$$

F 检验的统计值下界如式(10)所示。

$$\lfloor F^{[i:j+\delta]} \rfloor = \frac{(\lfloor SSE(M_r^{[i:j+\delta]}) \rfloor - \lceil SSE(M_f^{[i:j+\delta]}) \rceil/L}{\lceil SSE(M_f^{[i:j+\delta]}) \rceil/((j+\delta-i+1)-2L-1)} \tag{10}$$

其中，L 表示被比较参数组数量，$(j+\delta-i+1)$ 是模型中考虑的时间点总数，i 是起始时间点，$j+\delta$ 是结束时间点。

2.3　因果推断检测伪代码

算法的具体检测过程主要包含 2 个阶段：正向阶段和逆向阶段。在正向阶段，将一条新的边添加到现有的因果图 G 上，算法将为添加的边计算一个分数值，即因果图适合数据分布的程度。选择提高分数最高的一条边加入因果图 G 中，循环地进行边的添加，直到没有更多的边可以提高分数。正向阶段的正向等价搜索(forward equivalence search，FES)算法如算法 1 所示。

算法 1　正向等价搜索

FES(*sortedArrows*, *lookupArrows*, G)

1. While *sortedArrows* 不为空
   sortedArrows 分数差按降序排列的箭头列表，并且仅包含具有正分数差的箭头
2. 从 *sortedArrows* 中删除最高分数差异的边 $x \rightarrow y$ 对应的 Arrow A
3. if x 在 G 中不与 y 相邻 and A 在集合 *NaYX* 中 and A 中的 T 集合包含在 $x \rightarrow y$ 的 T 相邻中 then
   G 为正在构建的图，T 为 **y** 不与 x 相邻的所有邻点的集合，*NaYX* 是所有 z 节点的集合，使得 $z-x$ 且 z 与 y 相邻
4. if $NaXY \cup S$ 是一个图 and 不存在从 y 到 x 的半有向路径被 $NaXY \cup S$ 阻挡
   ▷ S 为正向 T 集合和逆向 H 集合的节点集合 then
5. 边 $x \rightarrow y$ 加入到 G and S 中的每个节点指向 y
6. $R \leftarrow$ ApplyMeekRulesLocally($\{x, y\}$)
   ▷ R 为一组节点集合，ApplyMeekRulesLocally 方法计算 Meek 规则参数
7. 从 R 中删除其邻居未被步骤 1 或未被 ApplyMeekRulesLocally 调用更改的节点。
8. 如果 x 和 y 不在 R 中则添加至 R
9. ReevaluateForward(R, G, *sortedArrows*, *lookupArrows*)
   ▷ ReevaluateForward 方法对正向搜索进行重新评估，主要过程为在 *sortedArrows* 中删除 *lookupArrows* 包含的所有箭头重新计算

正向阶段 FES 是一个连续的过程，在每一步都必须确定最大得分边缘，之后便进入逆向阶段(back equivalence search，BES)。逆向阶段与正向阶段相似，区别在于对边的添加改为对边的删除以达到逆向检验的目的。相比于正向阶段，逆向阶段只需要检查到目前已经包含在图中的边是否可能被移除，并且不需要考虑尚未添加到模型中的所有具有正分数差异的边。在每次迭代中，算法将改进得分最高的一条边删除，直至没有边可以删除。最终通过获得节点、边就得到了一个因果图。

F-GES 得到的因果关系既能拟合数据分布，又不会过度拟合。计算可以被分解来允许并行计算，并且计算可以在迭代中再利用。因此，F-GES 能够实现算法的高可伸缩性。

尽管 F-GES 因果算法是有效的，但实验过程中发现因果图中具有不确定性，难以保证因果关系都是真实可信的。因果关系可能随着时间的推移而发生变化，也可能存在滞后的因果效应。由于上述缺陷，因果检测的性能不能总是得到保证，所以本文引入了 Granger 因果算法(算法 2)加以验证。

算法 2　Granger 因果检验

输入：2 个时间序列 $X = x_1 x_2 \cdots x_T$，$Y = y_1 y_2 \cdots y_T$

输出：检验统计量和对应 p 值

1. For $i = 1$ to T do
2. $k_M \leftarrow i$，$k_H \leftarrow i$
3. For $j = i$ to T do
4. 根据式(3)～(8)用 $M^{[i:k_M]}$ 计算 $\lceil F_M^{[i:j]} \rceil$
5. if $\lceil F_M^{[i:j]} \rceil \leqslant c^*_{L,\, j-i+1-2L-1}$ then
6. continue
7. 根据式(3)～(9)用 $M^{[i:k_M]}$ 计算 $\lfloor F_M^{[i:j]} \rfloor$
8. if $\lceil F_M^{[i:j]} \rceil > c^*_{(L,\, j-i+1-2L-1)}$ then

9. 拟合全模型 $M_f^{[i,j]}$ 和简化模型 $M_r^{[i,j]}$
10. $k_M \leftarrow j$
11. $F_M^{[i,j]} \leftarrow \dfrac{(SSE(M_r^{[i,j]}) - SSE(M_f^{[i,j]}))/L}{(SSE(M_f^{[i,j]})/(j-i+1-2L-1))}$
12. if $\lceil F_M^{[i,j]} \rceil \leqslant c^*_{L,j-i+1-2L-1}$ then
13. continue

对于上述算法伪代码,首先枚举开始时间戳 i 和结束时间戳 j(第 1 ~ 3 行)。对于每个时间区间,使用回归模型 $M^{[i:k_M]}$ 计算 F 统计(F-statistic)的上界(第 4 行),如果上界小于等于对应的 F 分布临界值,则区间 $[i, j]$ 无法通过检验,则继续下一个区间 $[i, j+1]$ 的检验。否则,进一步计算 F-statistic 的下界(第 7 行),如果下界大于临界值,则区间 $[i, j]$ 通过检验,不需要拟合模型;如果小于等于临界值,就重新计算模型(第 9 行)并更新结束时间戳(第 10 行)。给定拟合模型后,计算当前的 F-statistic(第 11 行),如果区间 $[i, j]$ 没有通过检验,则继续下一个区间。拟合全模型需要 $O(T(2L)^2)$ 的时间,简化模型需要 $O(TL^2)$ 时间,所以这一算法的时间复杂度为 $O(TL^2)$。

F 检验生成具有相应 p 值的 F 检验统计量,如果 p 值小于显著水平(例如 $\alpha = 0.05$),那么拒绝原假设并得出结论,时间序列 X 导致时间序列 Y。当数据存在大量变量和滞后,F 检验可能会得出错误的结论,采用卡方分布[18]等其他检验方法得到检验的相关结果并存入数据库。根据格兰杰因果检验结果,将删除 p 值超过阈值的边,得出可靠的因果关系。

3 可视化系统设计与实现

因果关系可视化分析系统能够促进因果分析和推理的交互,帮助用户理解因果关系[19]。

3.1 系统总体架构

本文开发了一个新颖的交互式因果关系可视分析系统,整个系统的设计思路和框架如图 2 所示。

本文采用数据管理系统(MySQL)进行时序数据存储,采用 Python 进行数据处理和因果关系检测模型的实现,采用数据可视设计框架D3进行因果关系绘制,采用前后端连接框架 Django 作为开发框架是因为该框架简单易用,实现前后端连。开发了一个交互式因果关系可视分析系统,直观地展示时序数据中变量之间的因果关系,帮助用户理解并根据经验修正时序数据中的因果关系。本工作能够在数据更新时,实现对时序数据中变量之间的因果关系实时计算和更新。

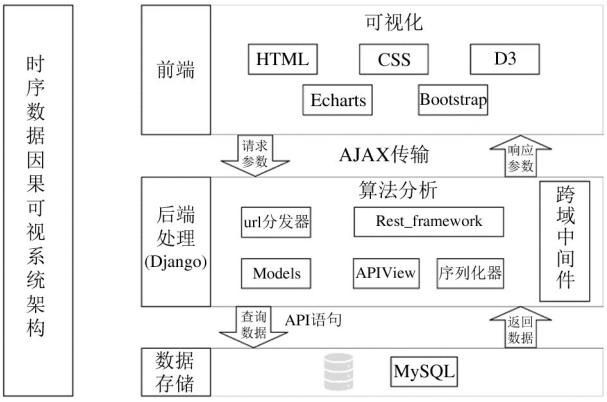


图 2 时序因果可视化交互系统总体架构

3.2 可视分析系统

本系统使用了 5 个主要可视化模块的设计与功能:(1)参数视图、关系视图呈现了因果树图;(2)数据视图通过平行坐标能够让用户对原始数据进行验证;(3)时间视图可以对多维度的数据进行趋势比较;(4)堆叠流图能够帮助用户进行时序层次演变的探索;(5)同时设计了一些交互操作,方便用户更好地理解、探索因果关系和数据。

参数视图如图 3(a)和(c)所示。图 3(a)某块支持用户动态调整参数,例如,系统支持交互式调节 F 检验的 p 值大小来筛选因果关系数据。图 3(c)某块为因果关系算法的计算结果表格,可以搜索相关元素或者排序,进而提高因果关系探索效率。

因果关系树图如图 3(b)所示,能直观有效地展示序列之间的因果关系。每个维度节点由一个圆表示,链接表示因果关系,方向一致。通过点击伸缩、悬浮展示关系等交互功能更好地理解因果关系。从左节点到右节点,用户可以将鼠标移动至节点或边查看相关属性和数据,也可以通过单击节点对此节点下属的子因果关系进行缩放,收缩的节点会以深色填充表示该节点下仍有子因果关系。

本文采用的树形图的布局，即每个节点的坐标主要采用 D3 库中的树形布局算法，输入参数是处于因果关系的各个节点和它们之间的关系，输出的返回值是表示所有节点的计算位置的数组。每个节点具有父节点属性和子节点数组、节点深度、节点的 x 坐标和节点的 y 坐标。为了实现缩放后的 x 和 y 坐标进行相应的移动以让布局更加合理，需要对节点的坐标重新进行计算。主要使用了树形布局(tree layout)和集群布局(cluster layout)算法实现对坐标的重新计算，树形布局算法通过传递根为节点建立 x 和 y 坐标，集群布局算法将树的每个叶节点定位在同一级别。

时间视图如图 3(d)所示，为原始数据的动态折线图，能够帮助用户进行各个元素的数据趋势比较。动态折线图直观地展示了随时间变化数据的发展趋势，将鼠标移动至某个节点会悬浮显示当日各个维度的价格情况并以竖线和节点的方式加以区别，供用户进行纵向比较；同时可以通过点击图例上的某个维度名称，描绘此维度的趋势供横向比较；根据图例所选的维度 y 轴的比例尺将动态变化至合适的范围。

堆叠流图如图 3(e)所示，其目的是帮助用户进行时序层次演变的探索。本文采用多层次堆叠图[20]来展示时序数据中所有变量的时序演变。每个时间序列被可视化为一个颜色图层，厚度表示给定时间步长的值。图层从左到右的方向表示随时间的演化，层的厚度反映了单个时间序列的总和。最高层表示层次结构的顶部，该层的厚度表示在每个给定时间步长的分组中时间序列的总和。全时序概览的时间步长 N 由多个变量的和得出：

$$N = c_1 + t_1 + d + t_2 + c_2 \tag{11}$$

其中，聚合层的时间步长定义为 c_i；过渡层的时间步长定义为 t_i；详细层的时间步长定义为 d；$i \in \{1, 2\}$，即为左右两侧。为了计算每个区域的相对水平长度，将每个层次的时间步长除以全概览的总步长 N，并乘以相应的投影因子，投影因子可以通过拖动竖线设置或自定义输入。详细层由 α 因子管理，过渡层由 β 因子管理，聚合层由 γ 因子管理，其中 $\{\alpha, \beta, \gamma\} \in N$。同时对投影因子进行限制 $\alpha > \beta > \gamma$，以保持这些区域时间步长之间的关系。聚合层对应的相对长度为 $RSC_i = \gamma \times \frac{c_i}{N}$，过渡层对应的相对长度为 $RST_i = \beta \times \frac{t_i}{N}$。

详细层对应的相对长度为 $PSD = \alpha \times \frac{d}{N}$，同样 $i \in \{1,2\}$，相对长度的总和用 RS 表示：

$$RS = RSC_1 + RST_1 + RSD + RST_2 + RSC_2 \tag{12}$$

堆叠流图的核心部分主要分为聚合层和详细层。该模块包括整个事件序列的高度抽象，支持查

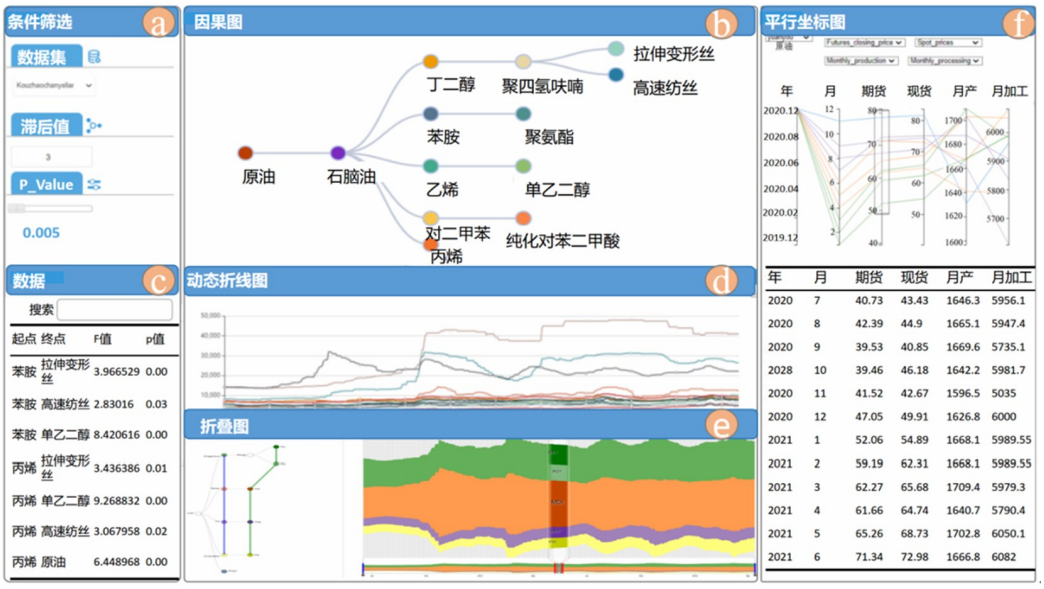


图 3　系统主界面

看整个流图的概览,控制滑块来放大区域进入详细数据分析界面,以查看详细层。

平行坐标图如图3(f)所示,其目的是帮助用户验证变量之间的因果关系。平行坐标图通过帮助用户理解变量之间的相关性来帮助用户更全面理解时序数据,进而帮助用户判断和验证变量之间的因果关系。平行坐标是一种与时间序列相关的用于可视化高维多变量数据的方法,它可以比较多个变量并查看数据之间的关系,每个垂直条代表一个变量,且所有轴均平行、垂直且等距放置。数据集的每个数据元素都通过连接的线段表示,这些线段源自一组连接的点,每个轴上都有一个点。最终将呈现一组线,每条线都是每条数据记录的多轴表示。通过选择数据维度和相关属性对原始多维数据进行探究,也可以通过控制坐标系的范围来筛选表格数据以进行交互分析。

4 案例研究

由于产业链数据中上下游产品间具有真实的因果关系,因此本节以口罩上游产业链中原材料的时序价格数据为例来验证因果模型有效性,同时阐述系统的实用性和可靠性。

4.1 数据预处理

本文研究的时间序列数据是在不同时间段记录的数据,按照其发生的时间顺序排列的一组数字序列[18]。时间序列的因果分析方法有基于信息论、条件独立检验、格兰杰因果检验等。由于控制实验的成本较高,现有的分析系统大多采用相关性分析来得出这种因果关系的结论。然而,相关性不是因果关系这一事实让更多学者开始研究因果分析,其目的是从大量数据中推断因果关系。

本文以2020年6月16日至2022年4月13日的口罩上游产业链中原材料的价格数据为例,进行因果关系分析。首先进行数据清洗和处理,然后将其存储在MySQL数据库中,部分属性如表1所示。

由于数据中聚四甲基二醇醚(PTMEG)维度的特征值较高,在同一坐标轴下则难以看出原油的特征值趋势。因此,本文采用归一化处理数据,各个维度的特征值将更加接近。接下来进行时序平稳化检验和处理,以原油从2020年6月16日至2022年4月13日期货价序列为例,用清洗完毕的原始数据进行平稳检验,包括ADF检验(augmented Dikey-Fuller test)和KPSS(Kwiatkowski-Phillips-Schmidt-Shin test)检验。最终检验结果如表2所示。

表1 产业链数据属性

国家	指标名称	频率	单位	时间区间	来源
美国	期货收盘价	日	美元/桶	20151125	洲际交易所(ICE)
美国	期货结算价	日	美元/桶	20151125	洲际交易所(ICE)
中国	期货收盘价	日	元/桶	20151125	上海国际能源交易中心
	原油现货价	日	美元/桶	20151125	同花顺金融
中国	原油	月	万吨	20151125	国家统计局
美国	期货成交量	日	手	20151125	洲际交易所(ICE)
美国	期货持仓量	日	美元/桶	20151125	石油输出国组织(OPEC)

根据上节介绍,ADF显著性检验的p值远远大于0.05,KPSS检验的p值小于0.05,且检验统计量同时大于3个不同等级的临界值,则该序列为非平稳序列,将通过下图差分方法将时间序列转换为平稳序列。

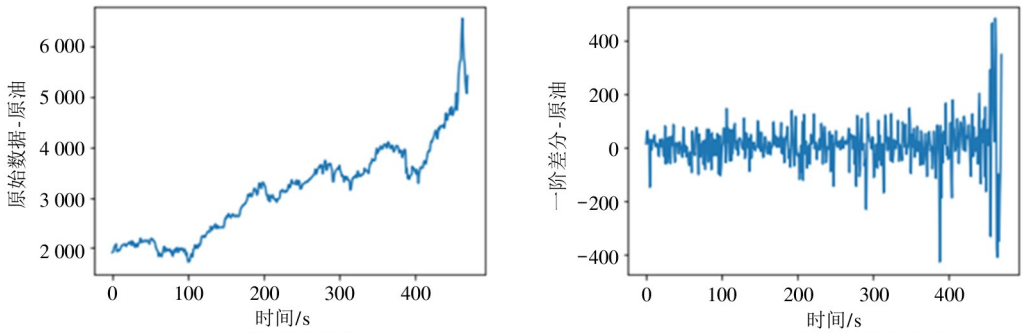
表2 原油期货价平稳性检验结果

检验算法	检验统计量	p值	临界值		
			1%	5%	10%
ADF	1.386 108	0.997 050	-3.445	-2.868	-2.570
KPSS	3.324 899	0.010 000	0.739	0.463	0.347

一阶差分法如式(13)所示。

$$\Delta y_t = y_t - y_{t-1} \tag{13}$$

其中，t 为序列时间节点的索引；Δy_t 为一阶差分的结果变量。对相邻时期作差得到一个新的序列，即用后一时期减去前一时期，让时间序列的波动曲线趋于平稳。使用差分法将非平稳序列转为平稳序列后如图 4 所示。将差分算法处理后的时间序列再次进行 ADF 和 KPSS 平稳性检验，结果如表 3 所示。



(a) 原始原油数据　　(b) 原油数据一阶差分结果

图 4　差分法效果图

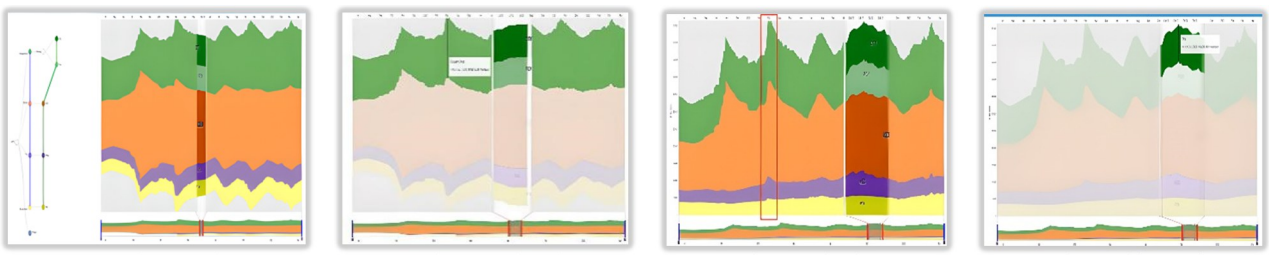
表 3　原油期货价差分后时间序列平稳性检验结果

检验算法	检验统计量	p 值	临界值		
			1%	5%	10%
ADF	−9.955 510	0.0	−3.445	−2.868	−2.570
KPSS	0.126 013	0.1	0.739	0.463	0.347

差分转换数据后，ADF 显著性检验的 p 值远低于 0.05，KPSS 检验的结果大于 0.05，且检验统计量同时小于 3 个临界值，因此，拒绝原假设，得出当前数据序列是平稳的结论。至此，时间序列已经从非稳态成为稳态序列，为后续的因果检验提供更可靠的数据。

4.2　产业链原材料时序数据分析

图 5 展示了口罩上游产业链中材料价格的时序数据分析结果。其中图 5(a)右侧的流图中的每一个条带代表产业链中的一种原材料，颜色编码用于区分 5 个子产业链，丁二醇、苯胺、乙烯、对二甲苯和丙烯为子产业链的上游节点，由图 5(a)左侧的垂直虚线链接，表示上游层次，同时也表示右边聚合层的组成元素。每个层次的厚度表示该产业链的原始数值大小，由图 5(a)可以看出苯胺子产业链的原始价格较高。从图 5(b)全时序概览层和详细层可以观察到，在 2020 年 10 月之前所有的数据维度都保持平稳状态，但在 2020 年 10 月、2021 年 3 月和 10 月以及 2022 年 3 月都分别达到了峰值，可以分析这段时间口罩上游产业链原材料价格有所上升，推测此产业链存在一定的周期性，高峰期为 3 月和 10 月。通过切换至堆叠图，观察到不同子产业链间的层次波动关联，以图 5(c)的 2021 年 3 月数据为例，在这段时间内各详细层次都有明显的凸起，即达到了价格的峰值，进一步证实了详细层之间的关联关系。通过控制滑块查看放大区域，图 5(d)选择了 2021 年 10 月的时间区间进行原材料详细数据分析，进一步探究层次之间的关联影响。可以观察到在10月



(a) 子产业链分析　(b) 产业周期性分析　(c) 产业关联分析　(d) 产业原材料详情

图 5　时序数据分析

19 日各个数据维度的价格都达到了所选时间的顶峰,并且呈现相同的趋势。由于产业链中的上游产品与下游产品之间存在因果关系,所以,当产业链中上游产品的价格波动时,下游产品的价格也会出现波动的现象。

4.3 因果关系及数据分析

以上述数据为基础,将因果算法参数滞后值设置为 3,p 值为 0.00 ~ 0.05 进行因果关系及数据分析。图 6 所示为算法得出的口罩上游产业链因果关系图,原油为产业链的最上游节点,下游为石脑油,其有多个子产业链,点击节点缩放子产业链,显示其因果关系子图;通过因果关系分析出原油的价格影响其下游石脑油的价格;另外,在此产业链中上游原油、石脑油等原始化学混合物,加工为丁二醇、苯胺等化学材料,再通过加工工艺生产出口罩上游产业链原料,如纯化对苯二甲酸(PTA)和单乙二醇(MEG),最后生产出拉伸变形丝(DTY)、高速纺丝(POY)等化学纤维制品。通过动态折线图可以比较多个维度数据的数值比较,通过比较时间序列的趋势来验证因果关系的可靠性。图 7 展示了丁二醇和 PTMEG 的数据,观察到这 2 个维度数据具有相似的趋势,而根据前文推断的因果关系丁二醇为 PTMEG 的上游,支持了因果算法的可靠性。通过因果数据模块可以查看计算出的F方差和p值结论,实时搜索某个维度或者排序。例如,观察原油对石脑油的因果计算结果,p 值小于 0.05,处于筛选范围内,表明原油为因,石脑油为果,存在单向因果关系。

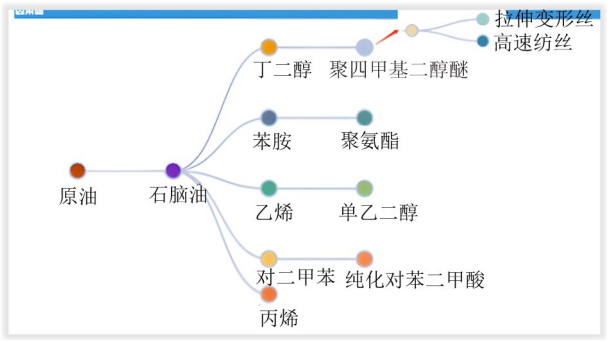


图 6 因果关系分析

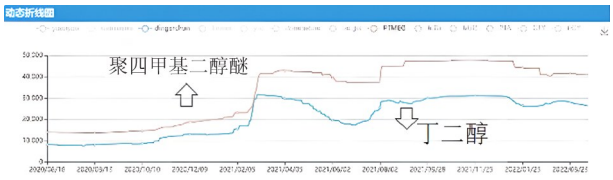
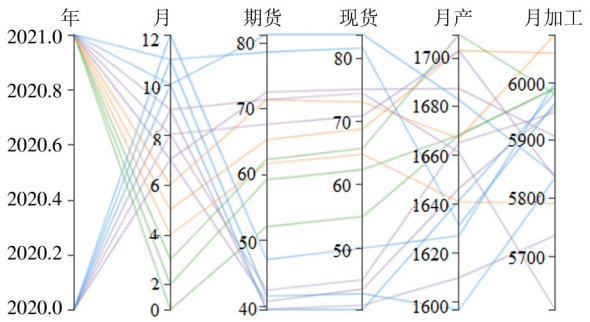
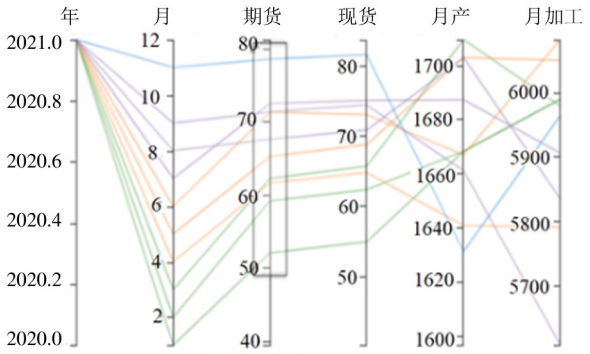


图 7 丁二醇、聚四甲基二醇醚数据比较

最后采用平行坐标进行相关性分析,如图 8 所示。选择原油的期货价、现货价、月产量和月加工值,比较原始数据的相关性。期货价和现货价之间多为平行线,表示两者关联性较大,成正相关,而月产量和月加工值之间交点较多,表明这 2 个属性之间不存在明显的关联性。图 8(b)在平行坐标系上将期货价的范围限制在 50 ~ 80 之间;由图 8(c)的表格数据和年轴可以看出,这个范围对应的时间均



(a)原油平行坐标关联性



(b) 原油期货价范围选择

年	月	期货	现货	月产	月加工
2021	1	52.06	54.89	1 668.1	5 989.55
2021	2	59.19	62.31	1 668.1	5 989.55
2021	3	62.27	65.66	1 709.4	5 979.10
2021	4	61.66	64.74	1 640.7	5 790.40
2021	5	65.26	68.73	1 702.8	6 050.10
2021	6	71.34	72.98	1 666.8	6 082.00
2021	7	72.53	75.04	1 687.2	5 905.60
2021	8	67.63	70.81	1 702.8	5 834.50
2021	9	71.37	74.36	1 661.0	5 607.20
2021	11	78.60	81.52	1 631.1	5 964.30

(c) 原油期货价表格数据

图 8 平行坐标相关性分析

在 2021 年,说明 2021 年相比 2020 年价格整体有所上涨。产业链中上游产品原油的价格直接影响了下游产品石脑油的价格,两者具有因果关系,验证了本文提出的因果算法的可靠性。

5 结 论

本文提出一种新颖的用于时序数据的因果关系检测模型,该模型整合了 F-GES 模型与 Granger 因果关系模型,展开因果关系的抽取和推断,并提出了交互式因果关系可视分析方法分析时序数据中变量间的因果关系,支持交互式可视操作,验证和总结时序数据中的因果关系。该方法有助于完成因果关系探索和假设分析,在已有经验的基础上完善和验证因果关系,得出有效的结论。同时用可视化来解释因果检测的过程和结果,帮助用户理解检测过程,为验证提供指导,也为因果关系研究方向提供了新思路。

然而,本工作存在一定程度的不足。首先,因果关系检测模型存在不确定性和误差。用户在模型检测结果的基础上进一步改正结果,无法将用户反馈参与到因果检测算法中。其次,系统探索的提示和导向不足。无先验知识的非专业用户无法判断因果关系的正确性,进而无法有效地使用该系统。在未来的工作中,将尝试将人工反馈参与到因果推断的算法中来改进因果关系检测模型,并改进该因果关系可视分析系统,将探索应用更加广泛的因果推断算法。

参考文献

[1] 蔡瑞初, 谢泳, 陈薇, 等. 面向社交媒体的直接因果网络发现算法[J]. 计算机应用研究, 2020,37(9):2689-2693.

[2] CUENCA E, SALLABERRY A, WANG F Y, et al. Multistream: a multiresolution streamgraph approach to explore hierarchical time series[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(12):3160-3173.

[3] RAMSEY J, GLYMOUR M, SANCHEZ-ROMERO R, et al. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images[J]. International Journal of Data Science and Analytics, 2017,3(2):121-129.

[4] REN W, HAN M. Survey on causality analysis of multivariate time series[J]. Acta Automatica Sinica, 2021,47(1):64-78.

[5] DORIS E, HOYER P. On causal discovery from time series data using FCI[EB/OL]// Proceedings of the 5th European Workshop on Probabilistic Graphical Models. Helsinki, Finland: HIIT, 2010:121-128.

[6] MAXWELL C. Optimal structure identification with greedy search[J]. Journal of Machine Learning Research, 2002(3):507-554.

[7] ELMQVIST N, TSIGAS P. Growing squares: animated visualization of causal relations[C]// Proceedings of the ACM Symposium on Software Visualization. San Diego, USA: ACM, 2003:1-11.

[8] NEUFELD E M, KRISTTORN S K, GUAN Q, et al. Exploring causal influences[C]// Proceedings of the Visualization and Data Analysis. San Jose, USA: SPIE, 2005, 5669:52-62.

[9] MORAWA R, HORAK T, KISTER U, et al. Combining timeline and graph visualization[C]// Proceedings of the 9th ACM International Conference on Interactive Tabletops and Surfaces. Dresden, Germany: Interactive Tabletops and Surfaces, 2014:345-350.

[10] WANG J, MUELLER K. The visual causality analyst: an interactive interface for causal reasoning[J]. IEEE Transactions on Visualization and Computer Graphics, 2015,22(1):230-239.

[11] WANG J, MUELLER K. Visual causality analysis made practical[C]// Proceedings of the IEEE Conference on Visual Analytics Science and Technology. Phoenix, USA: IEEE, 2017:151-161.

[12] CHOUDHRY A, SHARMA M, CHUNDURY P, et al. Once upon a time in visualization: understanding the use of textual narratives for causality[J]. IEEE Transactions on Visualization and Computer Graphics, 2020,27(2):1332-1342.

[13] 陈为, 朱标, 张宏鑫. BN-Mapping:基于贝叶斯网络的地理空间数据可视分析[J]. 计算机学报, 2016,39(7):1281-1293.

[14] 张义杰, 李培峰, 朱巧明. 面向事件时序与因果关系的联合识别方法[J]. 计算机工程, 2020,46(7):65-71.

[15] DENG Z, WENG D, XIE X, et al. Compass: towards better causal analysis of urban time series[J]. IEEE Transactions on Visualization and Computer Graphics, 2021,28(1):1051-1061.

[16] 陈加略, 姜远. 基于标记因果顺序挖掘的多标记分类方法[J]. 软件学报, 2022,33(4):1267-1273.

[17] 王双成, 郑飞, 张立. 基于贝叶斯网络的时间序列因果关系学习[J]. 软件学报, 2021,32(10):3068-3084.

[18] 付超, 胡旭洁, 于红艳, 等. 基于卡方分布的统一多学科可靠性分析[J]. 计算机集成制造系统, 2017,23(7):1439-1446.

[19] 李超, 求文星. 基于机器学习的因果推断方法研究进展[J]. 统计与决策, 2021,37(11):10-15.

[20] CUENCA E, SALLABERRY A, WANG F Y, et al. Multistream: a multiresolution streamgraph approach to explore hierarchical time series[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(12): 3160-3173.

Interactive visual analysis of causality in temporal data

DING Weijie[*,**,******], HUA Dong[***], YUAN Ying[**], SUN Guodao[****], YOU Zhiqian[******], LIANG Ronghua[****]

([*]College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023)

([**]Department of Computer and Information Security, Zhejiang Police College, Hangzhou 310053)

([***]Information Command Center, Zhejiang Provincial Public Security Department, Hangzhou 310053)

([****]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023)

([*****]College of Engineering, Hongkong Polytechnic University, Hong Kong 999077)

([******]Key Laboratory of Public Security Information Application Based on Big-data Architecture, Ministry of Public Security, Hangzhou 310053)

Abstract

As data storage technology is increasingly improving, the correlations of variables in time series data are more complex. It is difficult to artificially speculate on the causalities based on previous accumulated experience to support the exploration of deeper relationships. The use of machine algorithms to detect the causality between multivariate time series data and exert the potential value of data has important practical significance for the application of big data in marketing and health care. Aiming at low efficiency issues, high error rate and low interpretability of causality models in time series data, this paper combines the functional greedy equivalence search (F-GES) model with the Granger causality model for causal inference, and proposes an interactive causality visual analysis approach, which includes the parameter view to improve the efficiency of causality exploration, the causality tree to visually display the causalities, the time view to compare the original time series data, and the streamgraph view for users to explore the hierarchical evolution of raw dataset, and parallel coordinate to analyze correlations among variables. This system supports interactive visual manipulation, verification, and summarization of causal relationships in time series data. Thus, mining causalities between variables in time series data can help users for decision-making.

Key words: causality, time series, visual analysis, industry chain