arXiv:2304.12085v1 [physics.soc-ph] 11 Apr 2023

Dangoron：基于滑动窗口的大规模时间序列数据网络构建

Yunlong Xu 罗切斯特大学
yxu103@u.rochester.edu

Peizhen Yang 罗切斯特大学
pyang11@u.rochester.edu

Zhengbin Tao 罗切斯特大学
ztao6@UR.Rochester.edu

CCS 概念

• **信息系统** → 数据管理系统；**流管理**；

关键词

时间序列、气候数据、相关矩阵、滑动查询

ACM 引用格式：

Yunlong Xu、Peizhen Yang 和 Zhengbin Tao。2023。Dangoron：基于滑动窗口的大规模时间序列数据网络构建。载于*2023*年国际数据管理会议伴会论文集（*SIGMOD-Companion '23*），*2023*年*6*月*18*日至*23*日，美国华盛顿州西雅图。ACM，纽约，美国，3 页。https://doi.org/10.1145/3555041.3589399

1 问题与动机

大规模时间序列数据的监测与分析已成为数据库领域备受关注的重要方向。构建动态相关性复杂网络，是从神经科学 [2, 4, 11], 气候科学 [3, 7], 和金融 [6, 12]等多个学科的大量时间序列数据中提取洞见的高效方法。在基于相关性的网络中，节点由各自对应的时间序列定义，边则表示指定时间范围内节点之间的相关性。通常，在大多数分析中，会根据查询时间范围、查询窗口大小和滑动步长，在滑动窗口内计算所有时间序列的两两相关性，从而生成一系列相关矩阵。皮尔逊相关系数是多个领域中最常用的度量指标之一 [1, 5, 8]。

动机示例。近年来，通过量化功能连接指标随时间的变化，血氧水平依赖功能性磁共振成像（BOLD-fMRI）的动态分析为了解脑网络性质提供了宝贵见解 [4]。此类分析依赖四维功能性磁共振成像数据，其中包含时间序列BOLD活动记录 [2]。每个三维功能磁共振成像数据集在每个时间点约包含 100K-10M 个脑区。传统的四维功能性磁共振成像数据分析会对基于脑区的连接矩阵进行脑区划分，这些矩阵涉及彼此不同、不连续但紧密交互的脑区分区 [11]。一种更优的替代方案是在体素级数据上构建图，并进行特征选择和图嵌入。然而，使用原始体素级数据构建网络在计算上

允许为个人或课堂使用而免费制作本作品的部分或全部数字版或纸质副本，但前提是不得为营利或商业利益而制作或分发副本，且副本必须包含此声明以及首页上的完整引文。本作品中第三方组件的版权必须得到尊重。如需其他用途，请联系版权所有者/作者。*SIGMOD-Companion '23, 2023*年*6*月*18*日至*23*日，美国华盛顿州西雅图© 2023 版权所有者/作者所有。ACM ISBN 978-1-4503-9507-6/23/06。https://doi.org/10.1145/3555041.3589399

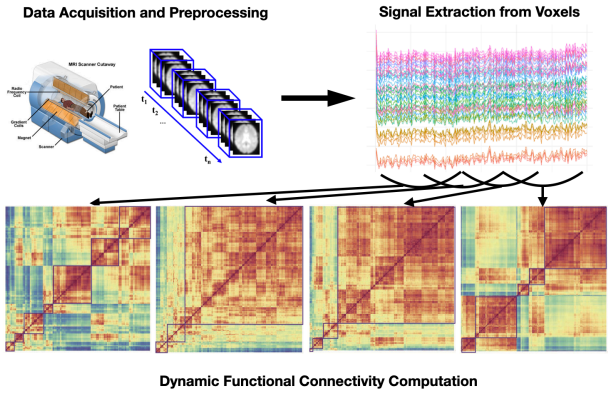

图1：跨滑动窗口的功能性磁共振成像网络构建

不可行，因为在滑动窗口内计算基于体素的功能连接矩阵需要耗时地进行两两相关性计算。本文介绍了在三维功能磁共振成像数据上为所有脑区计算一系列相关矩阵的技术

问题定义。给定一个由大小为 $N \times L$ 的矩阵 X 表示的时间序列集合。其中，共有 N 个时间序列，每一行表示一个长度为 L 的时间序列。X 中的每个元素 x_{ij} 表示在位置 i、时间 j 采集的某个变量的带时间戳的值。我们使用 X_i 表示第 i 行，该行对应于第 i 个时间序列。我们假设 X 中的所有时间序列均已同步，即每个时间序列在每个周期性时间间隔或时间分辨率上都有可用值。这可以通过对不同步的序列进行聚合和插值来实现。在查询时，用户定义查询范围 $r = (s, e)$，其中 s 和 e 分别为起点和终点，查询窗口大小为 l，滑动步长为 η，阈值为 β。该问题要求计算一系列相关矩阵 $\mathbf{C} = C^0, C^1, \cdots C^\gamma$，其中每个 C^k 都是 $\mathrm{X}k * \eta$ 的相关矩阵：$k * \eta + l$（即 X 从第 $k * \eta$ 列到第 $k * \eta + l$ 列的子矩阵）。在 C^k 中，每个元素 cij^k（$\geq \beta$）表示第 k 个滑动窗口范围内 X_i 与 X_j 之间的皮尔逊相关系数。如果 $c_{ij}^k < \beta$，则将其替换为 0。

核心任务是解决滑动窗口下大规模时间序列全对相关性计算的问题。主要挑战包括：1）确保大规模数据的网络构建与更新效率，以实现交互性；2）在分布各异的数据集上保持方法的稳健性；3）优化草图技术在后续网络分析中的有效性。

2 背景与相关工作

基于频率的变换 [9, 10, 16, 17] 和随机投影 [14, 15] 方法已被用于滑动窗口下的相关性计算。这两种方法都旨在降低时间序列的维度，并在低维空间中识别高度相关的序列对。基本窗口 [17]最初是为高相关性的时间序列对提出的，随后通过并行化得到改进 [9]。Parcorr [14]优化了增量计算，实现了最先进的效率。TSUBASA [13], 是最新的研究成果，它采用一种新颖的草图框架，在任意时间窗口上计算精确的成对相关性，但在处理滑动查询时效率不足。

除 TSUBASA 外，现有技术都因依赖数据而面临稳健性问题。基于频率的变换方法在对某些时间序列进行降维时表现不佳，只有当能量集中在少数几个域中时才能成功 [15]。迄今为止，还没有基准测试被提出以系统地测试解决方案的稳健性。

3 方法与独特性

在这项工作中，我们提出了 Dangoron，这是一个用于高精度、高效计算动态相关矩阵的框架；还提出了 Tomborg，这是一个用于生成时间序列数据集以测试框架稳健性的基准测试。

Dangoron 的核心概念是，转到下一个滑动窗口时，相关性相对稳定。通过计算相关性的未来上界和下界，我们可以优化计算。例如，如果当前相关性低于阈值，且下一个窗口的上界也低于阈值，我们就可以跳过下一时间步的计算。估计第 k 个窗口的边界，可以确定可跳过的步数，并直接跳转到目标窗口。

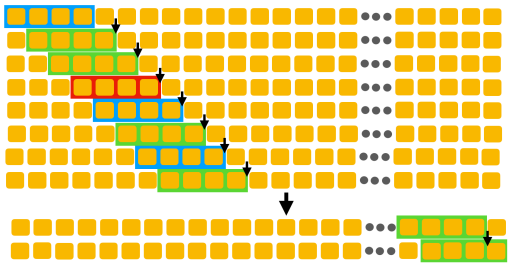

图2：Dangoron 的跳跃结构

我们采用基本窗口框架，将序列划分为更小的基本窗口，从而能够以更小的批次处理数据。参照**问题定义**，考虑来自 C 的 C_k，并将其记为 C。我们从 X 中选取 C 所指定时间范围内的第 i 行和第 j 行，并将它们标记为 $x = [\mathbf{x}_1、\ldots\ldots、\mathbf{x}_m]$ 和 $y = [\mathbf{y}_1、\ldots\ldots、\mathbf{y}_m]$。我们将基本窗口的大小定义为 $\mathbf{B} = [B_1, B_2, \ldots, B_m]$，其中 B_i 表示第 i 个基本窗口的大小。 x 与 y 之间的精确皮尔逊相关系数为：

$$Corr(x,y) = \frac{\sum_{j=1}^{n_s} B_j(\sigma_{x_j}\sigma_{y_j}c_j + \delta_{x_j}\delta_{y_j})}{\sqrt{\sum_{i=1}^{n_s} B_i({\sigma_{x_i}}^2 + \delta_{x_i}^2)}\sqrt{\sum_{i=1}^{n_s} B_i({\sigma_{y_i}}^2 + \delta_{y_i}^2)}} \quad (1)$$

$$\delta_{x_i} = \overline{x_i} - \frac{\sum_{k=1}^{n_s}\overline{x_k}}{n_s},\ \delta_{y_i} = \overline{y_i} - \frac{\sum_{k=1}^{n_s}\overline{y_k}}{n_s}$$

$\sigma_{x_i}(\sigma_{y_i})$ 表示 $x_i(y_i)$ 的基本窗口标准差， c_i 表示基本窗口 x_i 和 y_i 的相关系数， $x_i(y_i)$ 表示基本窗口 $x_i(y_i)$ 的均值， $\overline{n_s}$ 表示查询窗口中的基本窗口数量。利用公式1，我们可以预先计算并存储基本窗口统计量，并计算任意查询窗口和大小的相关系数。

图2展示了 Dangoron 的运行方式。每个黄色块代表一对基本窗口。首先，我们使用公式1计算精确相关性；如果结果低于阈值 β，则用蓝色块表示。接着，我们通过上界估计（例如公式2）确定 k，即使用二分搜索求解满足 $Corr_{i+k-1} < \beta$ 和 $Corr_{i+k} \geq \beta$ 的 k。在图2中，$k = 3$，我们用红色块表示高于 β 的上界。跳过 k 步（即跳过绿色块）后，我们会重新计算精确相关性。

$$Corr_{1+k} \leqslant Corr_1 + \frac{1}{n_s}(k - \sum_{i}^{k} c_i) = Corr_{i+k*} \quad (2)$$

我们已在每个基本窗口均来自某个样本分布的假设下证明了公式2，具体细节见我们的技术报告。对于更复杂的情况，我们推导了一般界。Dangoron 提出的另一项功能支持水平计算剪枝。给定时间序列 x、 y 和 z，如果已知成对相关性 c_{xz} 和c_{yz}，我们可以确定 c_{xz} 的范围为 $c_{xz}c_{yz} - \sqrt{(1-c_{xz}^2)(1-c_{yz}^2)} \leq c_{xy} \leq c_{xz}c_{yz} + \sqrt{(1-c_{xz}^2)(1-c_{yz}^2)}$。Tomborg 框架包括：(1) 根据用户指定的分布生成 C；(2) 在频域中生成 §；(3) 使用逆傅里叶变换将它们转换为 X。步骤 (2) 基于以下事实：离散傅里叶变换（DFT）保留系数与原始时间序列之间的距离。我们开发了逆 DFT 的实值变体，将数据从复空间转换到实空间（不同于原始的逆 DFT，后者将数据从复空间转换到复空间），并给出了证明。

4 结果与贡献

我们选取 TSUBASA [13] 作为基线。在 ***NCEA***数据集1 上，就纯查询时间而言，Dangoron 的速度比 TSUBASA 快一个数量级，且准确率达到 90% 以上，与 Parcorr 相当。在完成 Tomborg 的实现后，我们计划开展更多大规模实验。在本项工作中，我们做出以下贡献：1）我们介绍了 Dangoron，这是一个解决跨滑动窗口相关矩阵计算问题的框架，其速度至少比基线快一个数量级；2）我们提出了 Tomborg，这是首个针对相关矩阵计算问题的基准测试。

[1]https://www.ncei.noaa.gov/pub/data/uscrn/products/hourly02/2020/

参考文献

[1] Richard F Betzel。2022 年。网络神经科学与连接组学革命。载于 连接组学深部脑刺激。Elsevier，第 25–58 页。 [2] Joana Cabral、Morten L Kringelbach 和 Gustavo Deco。2017 年。功能连接在静态结构连接组上跨多个时间尺度动态演化：模型与机制。*NeuroImage* 160（2017），84–96。 [3] Avi Gozolchiani、Kazuko Yamasaki、Oz Gazit 和 Shlomo Havlin。2008 年。气候网络闪烁连接的模式遵循厄尔尼诺事件。*EPL*（*Europhysics Letters*）83，2（2008），28005。[4] R Matthew Hutchison、Thilo Womelsdorf、Elena A Allen、Peter A Bandettini、Vince D Calhoun、Maurizio Corbetta、Stefania Della Penna、Jeff H Duyn、Gary H Glover、Javier Gonzalez-Castillo 等。2013 年。动态功能连接：前景、问题与解释。*Neuroimage* 80（2013），360–378。 [5] CD Jayaweera 和 N Aziz。2018 年。主成分分析和皮尔逊相关系数在水处理厂人工神经网络模型开发中的应用可靠性。载于 *IOP*会议系列：材料科学与工程，第 458 卷。IOP Publishing，012076。 [6] Dror Y Kenett、Yoash Shapira、Asaf Madi、Sharron Bransburg-Zabary、Gitit Gur-Gershgoren 和 Eshel Ben-Jacob。2010 年。股市相关性的动态。*AUCO Czech Economic Review* 4，3（2010）。[7] Kyunghun Kim、Hongjun Joo、Daegun Han、Soojun Kim、Taewoo Lee 和 Hung Soo Kim。2019 年。考虑观测日降雨数据非线性的雨量站复杂网络构建。*Water* 11，8（2019），1578。[8] Anne C Krendl 和 Richard F Betzel。2022 年。社会认知网络神经科学。*Social Cognitive and Affective Neuroscience* 17，5（2022），510–529。 [9] Abdullah Mueen、Suman Nath 和 Jie Liu。2010 年。海量时间序列数据的快速近似相关。载于 *2010* 年 *ACM SIGMOD* 国际会议论文集数据管理会议. 171–182. [10] Yasushi Sakurai、Spiros Papadimitriou 和 Christos Faloutsos。2005。Braid：通过群组滞后相关进行流挖掘。载于 *2005*年 *ACM SIGMOD* 国际数据管理会议论文集。599–610. [11] Bertrand Thirion、Gaël Varoquaux、Elvis Dohmatob 和 Jean-Baptiste Poline。2014。哪种 f MRI 聚类能得到良好的脑分区？神经科学前沿 8 (2014), 167. [12] Oussama Tilfani、Paulo Ferreira、El Boukfaoui 和 My Youssef。2021。股票市场的动态互相关与动态传染：一种采用 DCCA相关系数 的滑动窗口方法。*Empirical Economics* 60, 3 (2021), 1127–1156.[13] Yunlong Xu、Jinshu Liu 和 Fatemeh Nargesian。2022。TSUBASA：基于历史与实时数据的气候网络构建。载于 *2022*年国际数据管理会议论文集。286–295.[14] Djamel Edine Yagoubi、Reza Akbarinia、Boyan Kolev、Oleksandra Levchenko、Florent Masseglia、Patrick Valduriez 和 Dennis Shasha。2018。ParCorr：跨滑动窗口识别相似时间序列对的高效并行方法。*Data Mining and Knowledge Discovery* 32 (2018), 1481–1507. [15] Tiancheng Zhang、Dejun Yue、Yu Gu、Yi Wang 和 Ge Yu。2009。带滑动窗口的流时间序列自适应相关分析。*Computers & Mathematics with Applications* 57, 6 (2009), 937–948. [16] Sheng Zhong、Vinicius MA Souza 和 Abdullah Mueen。2020。FilCorr：流式时间序列的过滤与滞后相关。载于 *2020*年 *IEEE* 数据挖掘国际会议（*ICDM*）。IEEE，1436–1441. [17] Yunyue Zhu 和 Dennis Shasha。2002。Statstream：对数千个数据流进行实时统计监测。载于*VLDB'02*：第*28*届国际超大型数据库会议论文集。Elsevier，358–369。