Advances in Philosophy 哲学进展, 2024, 13(5), 905-909
Published Online May 2024 in Hans. https://www.hanspub.org/journal/acpp
https://doi.org/10.12677/acpp.2024.135135

Hans汉斯

大数据视角下的因果性与相关性问题研究

李明伟，刘正平

昆明理工大学马克思主义学院，云南 昆明

收稿日期：2024年4月25日；录用日期：2024年5月15日；发布日期：2024年5月24日

摘　要

大数据研究中对因果性与相关性的探讨已成为当前的焦点议题，因果性的追求曾在科学的发展过程中扮演了重要角色，但在现代科学实践中却遭遇了许多挑战。与此同时，相关性凭借其高效的数据处理优势，展现出广泛的社会应用前景。本文深入探讨了科学领域中因果性与相关性的本质，并对其理论局限性进行了深刻反思。最终，在大数据的语境中揭示了因果性与相关性之间的内在联系，并据此提出了一种整合因果性与相关性的策略。

关键词

因果性，相关性，大数据，整合

A Study on Causality and Correlation from the Perspective of Big Data

Mingwei Li, Zhengping Liu

School of Marxism, Kunming University of Science and Technology, Kunming Yunnan

Received: Apr. 25th, 2024; accepted: May 15th, 2024; published: May 24th, 2024

Abstract

The discussion of causality and relevance in big data research has become the current focus of issues. The pursuit of causality has played an important role in the development process of science, but it has encountered many challenges in modern science practice. At the same time, relevance, with its advantages of efficient data processing, shows a wide range of social application prospects. This paper explores the nature of causality and relevance in science and deeply reflects on its theoretical limitations. Finally, the internal link between causality and correlation is revealed in the context of big data, and a strategy to integrate causality and correlation is proposed accordingly.

文章引用: 李明伟, 刘正平. 大数据视角下的因果性与相关性问题研究[J]. 哲学进展, 2024, 13(5): 905-909.
DOI: 10.12677/acpp.2024.135135

Keywords

Causation, Relevance, Big Data, Integration

1. 引言

2023 年 2 月，中共中央与国务院印发了《数字中国建设整体布局规划》，《规划》将数据要素放到一个更为宏大的“数字中国”图景中，并进行了周密的规划与部署。大数据正在深刻改变社会生活以及我们观察、理解世界的方式[1]。在众多领域内，诸如在线购物、医疗保健及人口管理等，运用大数据分析结果优化决策的做法，已经展示出其非凡的优越性。基于此，本文试图在大数据时代整合因果性和相关性，为未来的数据分析提供新的方法和洞察力，以更好地应用数据来推动中国的数字化转型和社会进步。

2. 科学领域中的因果性与相关性

2.1. 科学领域中的因果性

因果性是科学领域中一个极为核心且复杂的概念，它强调的是一个事件和第二个事件之间的作用关系。即前一事件被认为是后一事件的原因，后一事件被认为是前一事件的结果[2]。因果性在人类理解世界和揭示其本质的过程中发挥了不可替代的认知作用，科学研究的一个主要目标就是揭示和理解自然界中事物之间的因果关系。

科学领域中，古代原子论的支持者们坚信一切现象都是依据因果法则产生的。德谟克利特坚决否认任何事物会因机缘巧合而发生，原子论的另一位代表性人物留基波也曾持有类似的看法：“没有什么是可以无端发生的，万物都是有理由的，而且都是必然的”[3]。这些观点都坚信一切事件都有其明确的先前原因，这些原因决定了事件的不可避免发生。牛顿就宣称：若已知系统的起始状态和内部相互作用，便能准确预测其过去和未来的状态变化。因果性在经典物理学中被认为是确定性的，这种确定性的因果关系为科学预测提供了坚实的基础。19 世纪初的法国天文学家和数学家拉普拉斯态度则更加激进，他认为倘若我们掌握了宇宙某一特定时刻所有粒子速度与位置的完整信息，并且完全通晓自然界的运作法则，那么我们就有能力精确地推断出宇宙在任何过去或将来时刻的具体状态，这一观点被后人称为“拉普拉斯之妖”。拉普拉斯强调了科学界对揭示因果关系和确立事物确定性的不懈追求，体现了人们对科学有能力完全认知和预测自然现象的积极信仰。

在经典力学盛行的时代，人们相信只要我们明确了事物发展背后的原因，我们就能精准地预知和操控其结果。但随着实证科学的不断进步，因果观念在本质上的理解开始显露其局限性，面临着统计学、统计物理学，特别是量子力学的深刻挑战。

2.2. 科学领域中的相关性

相关关系是指客观现象之间确实存在的，但在数量上表现为不确定的相互依存关系[4]。相关性在科学研究中扮演着重要的角色，它帮助研究者揭示变量之间的关联关系，为深入研究和理解复杂现象提供

了基础支撑。随着科学技术的发展尤其是大数据时代的到来，相关性理论在当代受到前所未有的重视。

在量子力学领域，粒子的行为特征是借助波函数进行描述的，波函数把粒子的状态体现为多种潜在状态的叠加，这种状态的叠加反映了量子力学内在的复杂关联性以及其根本的不确定性。在量子世界中，实验设置在内的各种因素及其相互作用都变得极其复杂，我们不再能够将两个相同的实验设置视为等价，这导致了每次观测都具有其独特的个体性。最终我们只能借助概率论的框架来进行预测和解释，由此导致的结果是，科学度量在理论方面具有局限性，因果关系只具有概率性的意义[5]。我们在科学上不得不寻求相关性的转向。海森堡的测不准原理则进一步强调了量子力学中的相关性，该原理表明我们不能同时准确地测量一个粒子的位置和动量。这种复杂迷离的现象不仅揭示了量子世界中不确定性，而且还说明在描述粒子行为中，相关性比因果性扮演更为核心的角色，从而为我们重新理解和审视世界提供了一个以相关性为中心的新视角。

经典统计学的奠基人皮尔逊认为：“因果性的观念开始崩解，尽管它曾为物理学家带来无限的利益……今后，有关宇宙的哲学观将是一种相关的变量系统的哲学观，接近但绝非达到完全的相关即绝对的因果性”[6]。在皮尔逊的统计学视角下，因果性被当作相关性的一种特殊情形，虽然因果性在解释某些现象时是不可或缺的，但在许多情况下，通过研究变量间的相关性就足以揭示出其背后的规律和机制。

而且，在当今大数据时代的背景下，科学思维方式也产生了深刻的影响和转变，我们正在逐渐从深究因果性转向关注相关性。由于人类每一时刻都在产生大量复杂且多源异构的数据，为了更快更好地适应生活，我们完全有可能不去过分深究隐藏在数据背后的因果性，而是通过充分挖掘和揭示海量数据间的相关性来实现我们的目标。换言之，这代表着告别总是试图了解世界运转方式背后的深层原因的态度，而走向仅仅需要弄清现象之间的联系以及利用这些信息来解决问题[7]。

3. 大数据应用中因果性和相关性的局限

3.1. 因果性的局限

通过大数据分析，我们可以评估不同干预措施或事件对特定结果的影响，并进一步揭示其背后的因果联系。例如，在医疗领域，通过分析患者的治疗历程和健康状况数据，我们能够确定不同治疗方法对患者健康状况的具体影响。

然而，在探讨大数据应用中的因果性问题时，我们也不可避免地面临一系列局限性。首先，大数据分析通常侧重于数据之间的相关性，这导致了因果性的真正确定面临挑战，因为仅凭相关性无法建立因果性。其次，反向因果性是大数据中的一个常见挑战，这意味着两个变量可能相互影响，而不是一个单方面导致另一个。这种情况下，因果性的方向可能不明确，容易导致误导性的分析结果。此外，大数据处理中存在众多混淆因素，这些因素作为因果变量的选择十分困难，选择不慎很可能对因果性研究的准确性产生负面影响。因此，在因果性分析中需要仔细考虑和控制混淆因素对处理数据的污染。最后，数据采样偏差和伦理隐私问题也是因果性研究的限制因素，因为它们可能导致研究结果不具有普适性或受到伦理约束。

为了克服上述这些局限性，研究人员需要综合运用领域专业知识、因果推断方法和实验设计，以确保因果性的可靠性和准确性。但在实际应用中，由于探寻因果性的机制十分复杂，过分地追求因果的确定性和精确度，不仅难以取得因果性地成果，反而陷入了低效率的困境。

3.2. 相关性的局限

在大数据中，我们能够通过揭示出事物之间的相关性，来达到一些特定的目的。比如，基于聚类的大数据驱动相关分析可以揭示能源消耗与产品产量之间的潜在关系，从而提高能源和资源的效率[8]。然

而，在许多情况下，仅凭相关性进行推断并不总是可靠的，在大数据的研究过程中相关性也遭遇了种种困境。

首先，在大数据的背景下，数据的种类繁多且复杂交杂，其中包括了有价值的数据、伪造数据、噪声数据和毫无意义的数据，这些问题严重影响相关性度量的准确性，数据集中的噪声、异常值和缺失数据需要通过严格的数据清洗和预处理来管理。其次，数据中变量的随机性使得基于相关性的预测变得不稳定，人们很难准确无误地评估变量间的联系和发展趋势，尤其是在复杂的干扰因素作用下，对变量间相关性的评判就变得更加困难。再次，大数据的相关性有着追求全样本分析的趋势，计算相关性可能涉及昂贵的计算成本，并需优化算法以处理大规模数据。最后，处理包含敏感信息的大数据时，必须严格遵守数据隐私和安全标准，这对全样本的相关性分析来说也是一项巨大的挑战。总之，相关性分析在实际应用中有着各种局限性。相关变量的遗漏会影响预测模型的准确性，数据集和变量的选择至关重要，既要最大限度地提高预测准确性，又要评估基于特定特征的歧视可能性与公平性[9]。为保证数据处理结果的准确性和可靠性，我们要对相关性做出足够的反思。

4. 大数据中因果性与相关性的整合

在大数据的背景下，我们发现无论因果性还是相关性在实际运用中都有其局限性，要想全面地揭示现象的深层本质，仅仅依赖单一的理论框架是远远不够的，不同理论各有其显著优势。

相关性分析作为一种快速直观的方法，使我们能够通过对大量数据的处理和解读，探寻事件和现象之间的关联，从而对过去和现在的状态做出准确判断，并且还能预测未来的发展趋势。但也要看到，相关性分析虽然在大数据应用中具有极高的价值，但它仅仅是认识和理解世界的第一步。因果性研究能够揭示事件和现象之间的内在联系和作用机制，帮助我们从根本上理解事物的发展规律。科学领域中相关关系仍然无法取代因果关系，仅凭相关关系并不能说明应当如何成功干预现实世界[10]。尽管因果性研究可能更为复杂和费时，但它提供了一种更为深刻和全面的认识世界方式，学者周涛更是坚定地认为：“若人们忽视或放弃了对因果关系的追寻，这将是人类自我堕落的开始”[11]。

因此，在认识数据世界时，我们可以通过收集数据并运用技术手段分析其中的相关性，从而找到理解数据世界因果性的新途径，而不仅仅是在数据分析中寻找和验证预先设定的因果性。我们主张对大数据进行相关性分析并非完全放弃对因果性的追求。“其所关注的相关性是对因果关系的逼近和靠拢，是在无法确定因果关系时的一种折中”[12]。重新厘清因果关系和相关关系之间的关系是大数据时代的核心问题[13]。在大数据的背景下，我们需要对因果性与相关性进行有机整合，构建一个包含两者的综合研究框架，减弱因果性与相关性各自的局限性，这将对大数据的发展与应用大有裨益。

5. 结语

在当今以大数据为主导的时代，数据的广泛应用正深刻地影响着人类的生产方式和生活质量。因果性和相关性的研究不应该是相互孤立的，而是应该相辅相成，共同推进。通过因果性的揭示能够验证和深化通过相关性分析得到的结论，提高我们认识世界的准确性和可靠性。而通过相关性分析，我们可以快速找到潜在的关联和模式，为因果性的深入研究提供线索和方向。综上所述，大数据时代要求我们对因果性与相关性采取综合和协调的研究方法，通过两者的相互补充和整合，不断推动理论创新和实践进步，更好地利用大数据的力量推动中国的数字化转型和社会进步。

参考文献

[1] 维克托·迈尔-舍恩伯格, 肯尼思·库克耶. 大数据时代[M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013.

[2] 冉奎. 因果关系研究的条件论[D]: [硕士学位论文]. 武汉: 华中科技大学, 2013.

[3] 罗素. 西方哲学史(上卷) [M]. 何兆武, 李约瑟, 译. 上海: 商务印书馆, 1963: 98-99.

[4] 宋文婷, 贺天平. 局限与整合: 大数据下的因果和相关[J]. 系统科学学报, 2021, 29(1): 42-46.

[5] 宋文婷. 科学哲学视域下的大数据问题研究[D]: [博士学位论文]. 太原: 山西大学, 2021.

[6] 哈金. 驯服偶然[M]. 刘钢, 译. 上海: 商务印书馆, 2015: 274-275.

[7] 徐道一. 从因果性走向相关性的科学思维变革趋势——读《大数据时代》一书的一点启示[J]. 办公自动化, 2014(11): 34-36.

[8] Shuaiyin, M., Yuming, H., Yang, L., *et al*. (2023) Big Data-Driven Correlation Analysis Based on Clustering for Energy-Intensive Manufacturing Industries. *Applied Energy*, **349**, Article 121608. https://doi.org/10.1016/j.apenergy.2023.121608

[9] Lyria, B.M. and Chan, J. (2018) Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability. *An International Journal of Research and Policy*, **28**, 806-822. https://doi.org/10.1080/10439463.2016.1253695

[10] 王士博, 潘嘉曈. 论大数据证据的关联性[J]. 山东警察学院学报, 2023, 35(3): 102-112.

[11] 周涛. 为数据而生: 大数据创新实践[J]. 中国商界, 2016(6): 117.

[12] 刘东亮, 闫玥蓉. 大数据分析中的相关性和因果关系[J]. 国家检察官学院学报, 2023, 31(2): 23-41.

[13] 张梦娟. 大数据时代因果关系和相关关系之辩[J]. 经济师, 2024(2): 225-226.