软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(12):2733−2752 [doi: 10.13328/j.cnki.jos.004724]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

因果关系及其在社会媒体上的应用研究综述*

赵森栋, 刘 挺

(哈尔滨工业大学 计算机科学与技术学院 社会计算与信息检索研究中心,黑龙江 哈尔滨 150001)

通讯作者: 刘挺, E-mail: tliu@ir.hit.edu.cn

摘 要: 诸如物理学、行为学、社会学和生物学中许多研究的中心问题是对因果的阐述,即变量或事件之间直接作用关系的阐述.由于人们的日常行为和语言越来越多地映射到互联网上,或者根本就是互联网引起了大量新的行为和语言,致使社会媒体上存在大量的因果问题.与相关关系分析相比,社会媒体上的因果关系分析更加必要和迫切,首先,任何相关性的背后都隐藏着因果关系;其次,相关性分析得到的结论有时是不可靠的甚至是错误的;再次,基于相关性的方法无法用于管理、控制和干预变量或事件.论述了因果关系分析的必要性、重要性和社会媒体上存在的因果问题;综述了目前的因果分析与推断的基本理论、存在的问题和研究现状;通过比较现有因果关系分析的研究思路,预测未来的研究方向和因果分析理论及方法在社会媒体上的应用.

关键词: 因果关系;社会媒体;常识因果;贝叶斯网络;随机对照实验;准实验设计

中图法分类号: TP391

中文引用格式: 赵森栋,刘挺.因果关系及其在社会媒体上的应用研究综述.软件学报,2014,25(12):2733−2752. http://www.jos.org.cn/1000-9825/4724.htm

英文引用格式: Zhao SD, Liu T. Causality and its applications in social media: A survey. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2733−2752 (in Chinese). http://www.jos.org.cn/1000-9825/4724.htm

Causality and Its Applications in Social Media: A Survey

ZHAO Sen-Dong, LIU Ting

(Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Corresponding author: LIU Ting, E-mail: tliu@ir.hit.edu.cn

Abstract: The main objective of many studies in the physical, behavioral, social, and biological sciences is the elucidation of cause-effect relationships among variables or events. Many causality problems, occur when new words and behaviors are mapped from individuals to the Internet or are created by the Internetitself. Causality is hidden behind correlations; conclusion made by correlation analysis is likely to be unreliable or even wrong; and in absence of causality, methods based on correlation is unable to intervene, control and manage. Thus, causal analysis is necessary in social media. This paper first introduces the value, importance, and necessity of causality analysis, followed by causality problems existing in social media. Then, a brief overview of the recent research on causal inference is provided with analysis basic theory, problems and research status. Finally, comparisons among previous studies are made to suggest the future research directions and causality application in social media.

Key words: causality; social media; commonsense causality; Bayesian network; randomized controlled trial; quasi-experimental design

1 引 言

与相关关系相比,因果关系是对问题更本质的认识.诸如物理学、行为学、社会学和生物学中许多研究的

* 基金项目: 国家自然科学基金(61133012); 国家重点基础研究发展计划(973)(2014CB340503); 国家青年科学基金(61202277)
收稿时间: 2014-05-01; 修改时间: 2014-08-21; 定稿时间: 2014-09-30

中心问题是对因果的阐述,即对变量或事件之间直接作用关系的阐述[1].例如:一种新型药物在给定患者人群中疗效如何?一个新的法规可避免多大比例的犯罪?在一个特定事故中,个体死亡的原因是什么?这些都是因果问题,因为要回答这些问题都需要有数据生成过程的知识.这些问题的答案不能单独通过计算数据获得,也不能单独从控制观测数据的分布中获得.分析因果关系的黄金法则是实施随机对照实验,多数情况下,实施实验的代价很高,或者由于客观条件、伦理道德等因素的限制,使得随机对照实验根本不可行.

然而随着互联网和数据科学的发展,收集非实验的观测数据却要容易得多,尤其是在社会学领域.在以往对社会问题的研究中,要么采用设置对照实验的方法,要么采用分析观测数据的方法.第 2 种方法中的观测数据一般是通过调查问卷的方式获得.但是通过调查问卷获得观测数据的方法一般代价较高,获得数据量较小.社会媒体日益发展的今天,尤其是如 Facebook(www.facebook.com)、Twitter(www.twitter.com)、新浪微博(www.weibo.com)等在线社交网络(OSN)的大规模兴起,人们的日常行为和语言越来越多地映射到互联网上,或者根本就是互联网引发了大量新的行为和语言.这就使得社会问题研究所需数据的获取越来越容易,研究社会网络中人们语言和行为的需求随之也变得越来越迫切.

1.1 社会媒体上的因果问题

社会媒体上存在着大量的因果问题.例如:在微博和 Twitter 等社交媒体上,是什么原因导致一个用户去发布一条微博[2]?什么原因导致转发一条微博[3,4]?是什么原因导致一个用户去关注另一个用户[5]?是什么原因导致用户去购买一个特定的商品[6−9]?是什么原因导致信息在节点之间传播[10]?是什么原因导致一个用户去点击特定的广告[11]?同质性的原因到底是社会选择还是社会影响[12−14]?同伴影响对改变人们行为的作用到底有多大[15,16]?在 Wikipedia 上,文章的编辑数是否决定文章的质量[17]?在在线问答社区中,给用户颁发勋章导致用户的参与积极性增加还是减少[18]?高质量答案的出现是否导致其他用户贡献答案的积极性降低[19]?

上述这些都是对本质因果的探求,然而常识因果也同样大量存在于社会媒体中.大量的文本是社会媒体的重要组成部分,在文本中也蕴含着大量的常识因果.图 1 的例子中,左边的 3 条文本是 3 条真实用户的微博,右侧是可以从微博中学习到的常识因果.通过大量地学习这些常识因果,我们可以进行常识推理.如图可以知道:“因为淋雨所以感冒”,所以淋了雨之后就要预防感冒,要想不感冒就不要淋雨.由此可见,社会媒体中蕴含的大量常识因果是我们抽取信息、进行自动推理、实现自动问答系统、进行知识储备和自动理解语义等问题的宝贵财富.

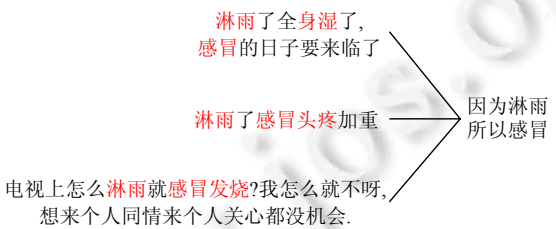


Fig.1 Common-Sense causality from Sina Weibo

图 1 从微博文本中抽取常识因果

总体来说,对因果问题的分析从层次上可以分为 3 层,即常识因果、浅层因果、深层因果.如图 2 所示,常识因果也就是经验因果,可以从每个人生成的因果逻辑文本中或专家的经验文本中直接抽取或进一步推理得到,简单直接但严谨性却无法保证.但是大量的互联网冗余数据能够给互联网文本中的常识因果的准确性提供一定的保证.浅层因果属于本质因果,是从观测的数据中使用统计分析方法和因果推断模型得到的观测变量之间的因果作用.这里涉及的本质因果是指通过因果推断模型和因果推断理论从观测样本数据中得到的样本数据中变量之间的直接作用关系.深层因果不是能够通过文本抽取或统计分析直接就能得到的因果关系,而是结合多个浅层因果或者在常识因果与浅层因果结合的基础上进行推理得到的因果知识,即在因果关系的基础上进一步推理得到的因果关系.

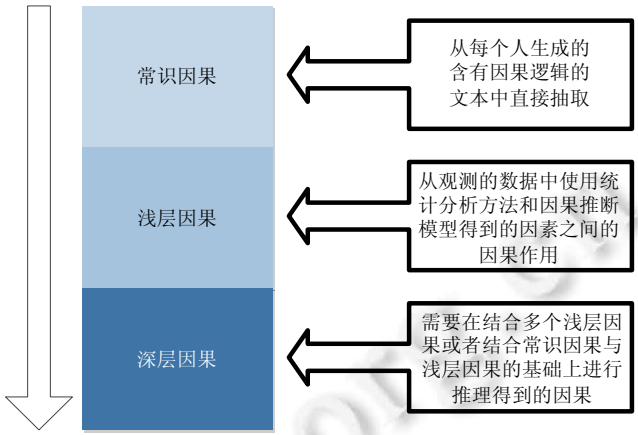


Fig.2 Hierarchical structure of causality

图 2 因果问题分析的层次

1.2 因果关系分析层次的哲学阐述

原因与结果是一对重要的哲学范畴.对事物间因果关系的探索自人类诞生以来就开始了,正如恩格斯所指出的:“由于人类的活动,就建立了因果观念的基础”.从概念来源上考究,史料考察表明:因果关系概念并不是由某一位先哲的个人思想所提出的,它也不是一种导出的规律或者派生的法则,而是人类在漫长的社会实践中逐步总结出来的一个关于事物联系和生灭变化的基本法则,并在历史和实践中长期受应用与检验进而不断完善,成为人们事实推理和认识未知的指南.

18 世纪英国经验哲学家休谟对因果关系的理解独树一帜,对其后的因果关系研究影响深远[20].休谟认为,对象间的关系表现为 3 种:原因与结果在空间上的接近关系;原因与结果在时间上的接续关系(前因后果);原因与结果的恒长结合.在这 3 种关系中,前两种关系不足以使我们断言任何两个对象之间的因果关系.但对于对象间是否恒长结合我们无法观测得到,所以休谟以其对因果关系决定论的批判,否定了因果关系的确定性,最终走向了唯心主义怀疑论的道路.继休谟之后,诸多哲学家关于因果关系提出了自己的论述.对康德来说,一个原因概念乃是一个变化概念,这个变化是遵循规律或法则而发生的,它是为规律所支配的事物的序列[21].德国哲学家黑格尔力图在思辨哲学的范围内突破因果关系的机械性质,把机械唯物主义的因果关系理解为形式上的因果关系,把辩证法上的因果关系理解为规定的因果关系[22].恩格斯唯物主义改造了黑格尔的合理思想,用实践观点论证了因果关系的客观性和普遍性,建立了辩证唯物主义的因果观[23].

在哲学领域,因果关系仍是一个不断发展和完善的概念,更是一个复杂的概念.因此,分层次、分步骤地认识因果关系就显得尤为必要.在本文中,作者把因果关系分为常识因果、浅层因果和深层因果:

- 常识因果对应哲学上的经验观点,即事物自有因果关联,知识的目的即在追求确定的因果关系[20].这种因果关系属于感性范畴,通常存在于经验性表达和人们的经验积累当中;
- 浅层因果对应哲学上的先验观点,这种观点的代表是康德哲学,是指事物(现象)本身无所谓因果,因果是认识主体本身具有的一种知性范畴,主体将此范畴普遍加于感性杂多之上,使之呈现为无例外的因果关系,也就是一般的科学世界图景[21];
- 深层因果体现辩证唯物主义的因果观,它综合经验性因果关系和先验性因果关系,在二者的基础上得到具有一定确定性的因果关系.深层因果关系属于理性范畴的概念.

1.3 因果分析的必要性

因果与相关是两个不同的重要概念,尽管在很多科学研究中因果比相关更重要,但是目前大多数统计方法仅专注于相关性研究.无因果关系的两个变量之间可能会表现出虚假的相关性;相反地,存在因果关系也可能表

现出虚假的独立性.很多例子可以说明虚假相关性,如张三和李四的手表上的时间具有很强的相关性,但是人为地改变张三的手表时间,不会引起李四的手表时间的变化.统计上的研究表明,小学生的阅读能力与鞋的尺寸有很强的相关性[24],但是很明显,它们没有因果关系,人为地改变鞋的尺寸,不会提高小学生的阅读能力.因果关系也可能表现出虚假的独立性.有统计数据表明,练太极拳的人和不练太极拳的人平均寿命相同或者更低.事实上,太极拳确实可以强身健体、延长寿命,但练太极拳的人往往是体弱多病的人,所以表现出虚假的独立性.

对问题的因果分析是十分必要的.因为相关性分析得到的结论有时是不可靠的,甚至是错误的.有一个体现这个问题的经典例子,叫做 Yule-Simpson Paradox.此悖论表明:变量 X 和 Y 边缘上正相关,但是给定另外一个变量 Z 后,在 Z 的每一个水平上,X 和 Y 都可能负相关.表 1 就是一个数值的例子[1].由表 1 可以看出:在整个人群中,吃药与康复之间存在正相关;然而当用性别对人群分层后发现,在男性和女性人群中,吃药与康复都是负相关.因此,当用相关性分析得到的结论来回答服用药物和安慰剂哪个会导致疾病的康复时,就变得非常困难.

Table 1 Yule-Simpson paradox

表 1 辛普森悖论

合并表	康复	未康复	康复率
吃药	20	20	50%
吃安慰剂	16	24	40%
在男性中情况	康复	未康复	康复率
吃药	18	12	60%
吃安慰剂	7	3	70%
在女性中情况	康复	未康复	康复率
吃药	2	8	20%
吃安慰剂	9	21	30%

其实,从初等数学中就可以证明:以上阐述的这个悖论没有什么新奇之处,$\frac{A}{B}>\frac{A'}{B'}$ 且 $\frac{C}{D}>\frac{C'}{D'}$ 是 $\frac{A+C}{B+D}>\frac{A'+C'}{B'+D'}$ 的既不充分也不必要条件.但是在统计上,这具有重要的意义,即变量之间的相关关系可以完全地被另外的一个第三变量扭曲.更严重的问题是,我们收集的数据可能存在局限性,忽略潜在的第三个变量可能改变已有的结论,而我们常常对此却一无所知.鉴于 Yule-Simpson 悖论的潜在可能,用相关性分析得到结论解释问题是靠不住的.因此对很多问题来说,因果分析是十分必要的.

1.4 因果分析的重要性

对于大部分统计机器学习方法来说,分析数据的目的是识别变量之间的统计相关性.用观测到的变量值去预测未观测变量的值时,这种相关性是非常有用的.如上文所述,我们可以利用小学生鞋子的尺寸预测他们的阅读能力.这种机器学习构造的相关性模型在很多领域都有广泛的应用,如自然语言处理、计算机视觉、信息抽取和信息检索等.在这些领域中,单使用变量之间的相关性对于满足系统要求来说就已经足够了.

但是,机器学习的方法却常常希望能够在给定的情景下提供决策支持.也就是说,如果已经设定想要的结果或者不想要的结果,该如何干预或者控制那些因素.例如在医学诊断中,大多数的医学专业人员不仅仅是想诊断疾病,更重要的是如何预防、控制和缓解疾病.他们想知道一个特定的医学干预手段会对病患的健康状况产生什么影响,所以医学专业人员就需要对这种影响进行建模来帮助他们设计有效的医学干预,这就需要因果模型而不单是统计相关模型.如图 3 所示,C 是 A 和 B 的共同原因,也就是 C 导致了 A 和 B.在没有其他混淆变量的前提下,A 与 B 统计相关是很显然的.但是如果想改变 B 的值,干预 A 是否有用?显然,要改变 B 的值,干预 A 是无作用的,只能干预 C 才能达到改变 B 的目的.

再者,几乎所有的相关关系背后都隐藏着深层的因果关系.如图 4 所示,任何两个变量 A 和 B 的相关性都是由 3 种情况的因果关系产生.如果 A 与 B 相关,那么或者是 A 导致 B,或者是 B 导致 A,或者是存在一个第三变量 C,既导致 A 又导致 B.如果不区分是 3 种情况的哪一种造成的 A 与 B 之间的相关性或者就在假定第 1 种情况的前提下去设计干预,很可能会导致干预的无效.

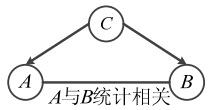


Fig.3 Causal and correlative relations among *A*, *B* and *C*

图 3 3 个变量 *A*,*B*,*C* 之间的因果关系与相关关系

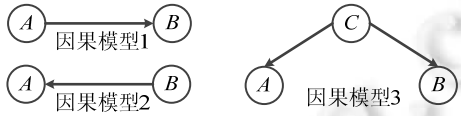


Fig.4 Three causality models which can generate correlation

图 4 3 种导致 *A* 和 *B* 统计相关性的因果模型

通过因果分析模型在观测数据上学习到特定领域的本质因果关系后,对我们理解问题来说可能还不够全面.如果能加入一些常识因果,如文本中抽取的因果或者专家经验等,对于理解问题、设计干预手段、评价干预效果会更加准确高效.图 5 给出了一个例子,体现常识因果的重要性以及结合常识因果与本质因果的重要性.我们可以利用常识因果抽取手段得到:淋雨→感冒;感冒→发烧.利用临床观测数据上的本质因果分析得到:姜糖水→治疗风寒感冒;抗生素→治疗病毒感冒.在综合利用上述得到的因果结论的前提下,如果新的病例描述为“今天雨下的特别大,出门又忘了带伞,结果就挨浇了,下午回到家就头疼发烧”,再结合感冒种类的知识,就可以推断出该病患得的是风寒感冒,且推断出治疗手段应该是“服用姜糖水”.

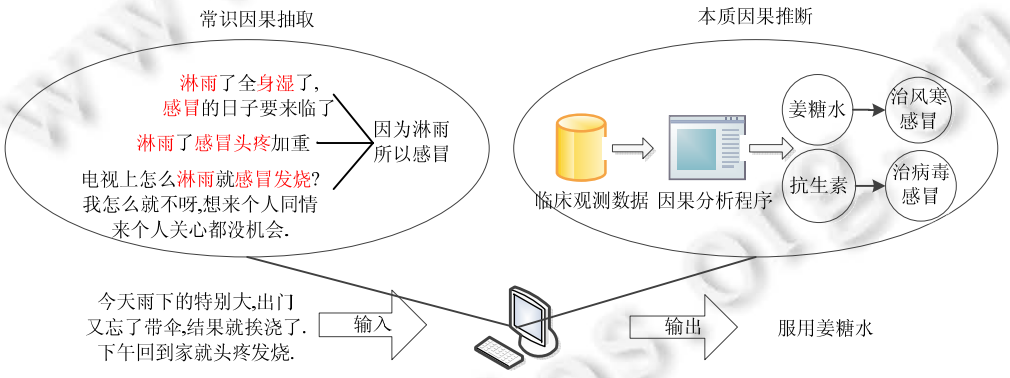


Fig.5 Design intervention means using common-sense and essential causality

图 5 利用常识因果和本质因果设计干预手段

2 研究现状分析

2.1 常识因果分析

常识因果,也可称为经验因果,是存在于自然语言文本中的、由每个社会个体根据自己的经验和知识生成的用来描述事物之间因果关系的文本.这是一种群体智慧体现的因果,因为虽然个体所表达的因果或许不是绝对正确,但是被大众所普遍认可的因果常识很可能是正确的.虽然正确性和严谨性不如本质因果,但是因为常识因果包含在自然语言文本中,所以获取这些因果比本质因果简单、直接,而且数据规模大.在文本中进行常识因果抽取就要用到自然语言的处理技术和方法,如词性标注、句法分析、短语抽取等.如图 6 所示,我们可以明显地看出句子中含有的因果对(Smoking $\xrightarrow{\text{导致}}$ poor blood flow),其中,触发词 cause 作为动词(VB),是 smoking 的父节点且关系为名词主语(nsubj);同时,cause 又是短语 poor blood flow 的父节点,关系为直接宾语

(dobj).对于这种含有关系触发词的表达因果的句子,原因和结果在句子中的词性和句法角色是有一定规律性的.基于这种认识,Ittoo 等人提出了一种基于词性、句法分析和因果关系模板的因果对抽取方法[25,26].在他们的工作中,首先使用 Wikipedia 上明确含有因果关系的句子抽取出一些表达因果的句子模板,然后用这些模板去抽取其他句子中的因果.Sorgente 等人利用人工构建的依存句法规则抽取出可能的因果对,并使用基于词性、语义和依存特征的贝叶斯分类器来过滤掉不是因果对的噪声词对[27].以上这些方法的核心都是基于词性和句法特征的规则来抽取因果词对.但是我们不得不承认,表达因果的句子千变万化,还有一大部分句子是没有因果关系触发词的.再加上社会媒体语言的随意性和中文语言表达的特点,使得有很大一部分表达因果关系的句子中并不含有因果关系触发词.但是上述研究工作中使用的方法,对于抽取这种句子中的因果关系对是非常困难的.

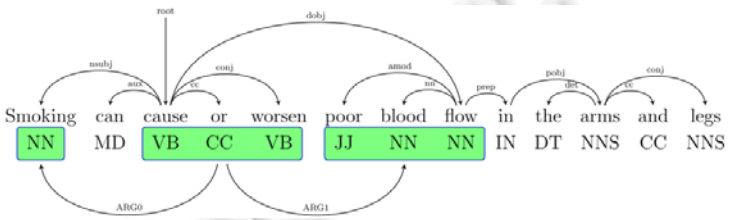


Fig.6 POS tagging and dependency parsing of sentence containing causality relation

图 6 含有因果关系句子的词性标注和依存句法分析

另外,还有一些学者基于常识因果知识的应用问题研究.Girju 在 2003 年使用基于关系触发词的规则方法抽取英文文本中的因果关系用于问答系统[28],又于 2010 年利用社会媒体的文本挖掘动词之间的因果关系并组成动词间的因果交互链,使用动词间的因果交互关系识别社会媒体上用户间的个人关系[29].例如,如果社会媒体上的文本是:Mary likes Paul for helping her sister,那么应该抽出 help(Paul, Mary's sister)导致 like(Mary, Paul). Gordon 等人就 Weblog 上的大量的人物故事,用 COPA(the choice of plausible alternatives: http://people.ict.usc.edu/~gordon/copa.html)(替代方案选择)问题来评价常识因果推理效果,发现在加入大量人物故事语料后,只是使用简单的 PMI(互信息)方法就比传统的统计和检索方法在 COPA 问题上表现要好[30].但是该研究基于一个潜在的假设,即共现所表现出的关联性就表示因果.显然,因果关系和共现关系的区别是非常大的,而且因果关系的要求也更加的严格.Kozareva 使用因果关系触发词抽取文本中的名词因果对,使用这种因果对来判断一个句子是否是描述因果逻辑的句子[31].Radinsky 等人利用因果关系词在大量的新闻语料中获取事件之间的因果关系,并把这些因果事件分类关联组成事件因果关系网络,使用这个网络预测未来事件[32].

上述这些研究在抽取因果关系时使用的方法本质上都是基于因果关系触发词的规则方法,这对于显式的存在关系触发词的文本无疑是合理的,但是对于没有明显关系触发词或者根本无关系触发词的隐式因果抽取却无能为力.这也是当前因果关系抽取的重要瓶颈.因此笔者认为:对于句子结构变化来说,词与词或者短语与短语之间语义上的逻辑关系更加稳定.例如"饿了,吃饭"、"困了,睡觉"、"受寒,感冒"这些词对,不用了解上下文我们就知道"因为饿了所以吃饭"、"因为困了所以睡觉"、"因为受寒所以感冒".人的大脑或许是在多种语境的训练下得到的这种词之间的逻辑关系而存储起来的.因此笔者认为:学习词或短语语义之间的逻辑关系(尤其是因果关系)进而表示成知识库的形式,与不断地优化因果抽取算法相比更有意义.

2.2 本质因果推断

本质因果是非常有价值的,所以到目前为止发展了一些从数据中自动挖掘这种因果的方法.目前,至少有 3 类挖掘本质因果的方法,它们在使用的数据类型和自动化程度上都存在差异:

第 1 类方法是随机对照实验方法.它要求分析者要对实验数据的产生过程有很深入的了解以及很高的控制能力;

第 2 类方法是准实验设计方法,它是一种在社会学研究中被大量使用的方法.这种方法是在观测数据中试图寻找能像随机实验方法一样满足因果推断条件的情形来进行观测数据上的因果推断[33−35];

第 3 类方法是联合模型方法,大致又分为图模型方法和虚拟事实模型方法.这种方法在一些假设的前提下自动对联合概率分布进行估计,从非实验数据或者观测数据中推断因果.

这 3 种方法面临共同的挑战:

首先,这些方法都需要识别一对变量之间是否存在统计相关性.推断统计相关性的原则和方法,即统计假设检验已经提出了几十年,不管对于人工的还是自动的算法来说,解决这个问题的困难都不大,但却仍旧存在一些挑战,比如它的对立面问题:统计独立性检验和完全的条件独立性检验,就是一个非常活跃的研究领域[36−39];

其次,这些方法必须要识别潜在因果的方向,即,哪个为因哪个为果.对于这个问题,往往通过考虑时序的方法来解决,即先发生的为因后发生的为果.但是在联合模型方法中也常常用一些其他的方法,下文详述;

最后,这些方法都必须避免其他混淆因素的影响,即,其他潜在的共同原因对变量之间因果关系的干扰.本节将详细论述不同类型的方法对于解决这些问题的尝试.

2.2.1 随机对照实验

当今对于挖掘有效的因果知识来说,可能最普遍的方法就是随机对照实验.在过去的 50 多年中,生物学、物理学和社会科学等的快速发展扩张,很大一部分有赖于如何设计实验并分析结果的知识.对于实验设计方法的发现、整理并传播,代表了过去一个世纪以来的人类智力成果.随机对照实验这种方法包含了两个非常关键的概念,即控制和随机化:

- 控制通常涉及研究人员有目的设置一些变量的替代值的能力,然后比较这些替代设计的效果.控制是实验这个概念的核心,并且有相当长的历史,最早可以追溯到 John Stuart Mill(1843)或许还可以追溯到更早一个世纪的时间[40].通过在实验中控制变量的方法,研究人员既可以通过保持变量不变屏蔽掉变量的效果,又可以通过系统改变变量值得到变量改变的效果.但是要做到这样,研究人员就必须知道特定变量的存在情况,并且能够改变和控制它们的值;
- 随机化涉及实验组随机分配对象的方法(例如随机分配医学实验中的病人),这样,研究人员就无法控制实验对象的特征也不能系统地影响被研究的变量.如果实施了随机化,这些不被控制的特征影响就会被均匀地分摊到足够大的组内.19 世纪 20 年代,Fisher 就概括了随机化的原则及其在实验设计中的应用[41],自此,随机化也成为了实验设计的重要内容.随机化的特别之处在于,它能移除那些对研究者们来说透明变量的影响.例如,只要把病人随机地分配到实验组中,研究者就不需要知道哪个具体的遗传因素可能会影响病人对某种特定药物的反应.研究实验环境现象的学者一般都会控制他们能够系统改变的变量或者能够保持不变的变量,并且对其他的大多数甚至全部变量进行随机化.通过这两种方法,就能够研究能被直接操控的变量效果,并且屏蔽掉几乎全部的其他潜在原因.

近几年,使用这种随机对照实验的方法,有大量的学者在社会媒体上做了有意义的研究和探索.Centola 等人邀请 1 540 名志愿者并随机地将他们一对一地分配到小世界特征的随机网络和高聚类的规则网络中,并观察研究两种不同网络下行为的传播规律来确定社会网络结构对与行为蔓延的因果作用[42].基于相同的原理,Centola 又基于性别、年龄及身体质量指数等特征,随机化地把在线社会网络中的用户分成两组,并让一组人员相互之间能发挥同质性作用,另一组完全没法发挥同质性作用,然后,通过分析用户的健康饮食日志在网络中传播的情况,分析人口组成的同质性对健康行为传播和革新采纳的因果作用[43].Lewis 等人也通过这种随机对照实验的方法研究在线社会网络上导致同质性的原因到底是社会选择还是社会影响[13].Aral 等人把 Facebook 上的 140 万朋友关系作为研究对象,使用随机对照实验的方法研究这些朋友关系所实施的同伴影响对同伴在某些产品使用上的因果作用[16],并且于 2013 年使用随机对照实验的方法研究社会媒体上的已有的用户投票结果对于后来者给出好评或坏评的因果作用[44].

2.2.2 准实验设计

准实验设计是社会科学领域中经常使用的因果推断方法,一般简称为 QEDs[35].这种方法试图利用能部分模拟对照实验环境的观测数据集来做因果识别[33,34].虽然 QEDs 无法总能具有像随机对照实验那样的内部合理

性,但是 QEDs 却增加了可分析数据的广度,尤其是对那些无法进行随机对照实验的情形问题中的因果推断,因此弥补了随机对照实验的某些不足.

在没有明确的控制和随机化的情形下,有些 QEDs 使用匹配的方法来确定对比数据实例对,以保证除了研究目标变量外的其他变量尽可能地相似,即非等值组设计.还有一些其他的 QEDs 研究相同数据实例上给定变量在特定事件前后随时间的变化,即断点回归方法.还有一些其他类型的 QEDs,包括 proxy pretest design[45], double pretest design[33],nonequivalent dependent variables design[33],pattern matching design[46]和 regression point displacement design[47].

准实验设计有一定的优越性:

首先,它在因果推断内部合理性方面超越统计控制方法,因为它可以控制全部变量,即使这些变量没有被识别、度量和建模;

其次,它在外部合理性上超越随机对照实验,因为准实验设计使用的是真实系统中的数据而不是人造实验环境下产生的数据.随机控制实验在因果推断上的有效性需要很高的代价,因此,随机对照实验有很高的内部有效性,但是需要牺牲外部有效性(即泛化到真实世界);相反,准实验的方法具有很高的外部有效性;

第三,QEDs 不需要额外地收集数据,反而可以把它们用于现有的数据集并推断出很强的因果结论;

最后,QEDs 不排斥其他的因果推断方法,它可以很好地辅助统计控制方法和随机实验方法.

当然 QEDs 也有一些局限性,比如:人工的 QEDs 只能用于有限的因果推断情形,例如双胞胎研究;由于 QEDs 只使用数据的子集来推断因果依赖,因此对与数据子集的代表性就要求很高.

准实验设计方法上非常著名的例子是双胞胎比较研究,这个研究已经延续数十年,其目的是探索某些疾病和情况的原因,比较同卵双胞胎集合和异卵双胞胎集合在某种疾病上的发病率.同卵双胞胎有相同的基因、共同的胎儿期环境和几乎相同的成长环境,异卵双胞胎也有相同的胎儿期环境和几乎相同的成长环境,但他们基因却不是完全相同而只是相似.这种典型的相似背景以及这两种类型的双胞胎在相似背景下又有特定的不同,为研究遗传因素在疾病上的作用提供了接近理想的环境.例如,为了识别某些已知的情况是由于遗传因素导致的,研究人员就可以在每种类型的成对双胞胎上确定相关性并且比较两种类型的相关性.如果差异大,说明这个特定情况很大程度上是由于遗传因素;反之,如果没什么不同,则说明这种情况是由于其他因素.

但是在传统的 QEDs 中,所有的步骤都是人工分析.这种方式耗时耗力,每次 QEDs 都要重复一遍所有的步骤.为了解决人工 QEDs 效率低下的缺陷,Jensen 等人提出了关系型数据上自动识别 QEDs 的方法 AIQ(自动准实验识别)[48,49].Oktay 等人使用这种自动的 QEDs 识别框架分析社会媒体上的因果分析,否定了人们对于问答社区中的认识"高质量答案的出现会导致用户继续贡献答案的积极性下降"[18].

2.2.3 图模型

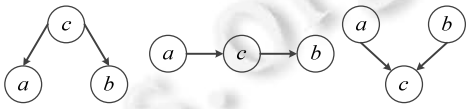
对于因果推断的图模型方法研究,是因果推断领域最活跃的研究方向之一.图模型的优越性在于直观,并且很容易地就能把因果推断和概率独立性理论联系起来.除了少部分学者研究线性有环模型上的因果推断[50−52],大部分图模型上的因果推断研究都是基于 DAG(有向无环图)的.对于 DAG,一般有两种观点认识它:一种是将 DAG 看成是表示条件独立性的模型;另一种观点则是将其看成是表示数据生成机制的模型.而因果推断中常常使用的 DAG 是将其看成数据生成机制的模型,一般称其为贝叶斯网络(或贝氏网络).贝叶斯网络中的节点代表随机变量,节点间的边代表变量之间的直接依赖关系(也可以看成因果关系),每个节点都附有一个概率分布,根节点 X 所附的是它的边缘概率分布 $P(X)$,而非根节点 X 所附的是条件概率分布 $P(X|\pi(X))$.

贝叶斯网络可以从定性和定量两个层面来理解:在定性层面,它用一个 DAG 描述了变量之间的依赖和独立关系;在定量层面,它用条件概率分布刻画了变量对其父节点的依赖关系.在语义上,贝叶斯网络是联合概率分布的分解的一种表示,它表征多个随机变量的联合生成的概率分布[53].更具体地,假设网络中的变量为 $X_1,\ldots,X_i$,那么把各个变量所附有的概率分布相乘就得到联合概率分布,即:

$$P(X_1,...,X_i)=\prod_{i=1}^{n}P(X_i\mid\pi(X_i)) \tag{1}$$

其中,$\pi(X_i)$表示变量 X_i 在贝叶斯网络中的直接父节点.

贝叶斯网络的构造方式有两种:一种是专家手工构建,另一种是通过数据分析获得.前者存在很大的缺陷:首先,人工构建贝叶斯网络,需要对这个贝叶斯网络所代表的问题本身有深刻的理解;其次,人工构建往往会遗漏掉一些变量.既然贝叶斯网络是描述数据生成机制的模型,那就假设所有存在因果关系的观测数据都是基于一个贝叶斯网络的,那么,如何从观测数据中学习出这个贝叶斯网络就成了一个非常重要的课题,即如何通过分析观测数据获得贝叶斯网络的结构和参数,其中,参数一般指贝叶斯网络中非根节点的条件概率表.然而,贝叶斯网络中的因果结构学习比贝叶斯网络结构学习要求更严格,因为表征因果结构的贝叶斯网络中每一条边都表征的是因果关系.大多数因果结构学习算法都有一个强假设:对所有变量 A,B 间的因果推断,可以观测所有潜在直接或间接的共同原因,即不存在图 10 所描述的情况.如此,因果图 $G(V,E)$上 $a\in V,b\in V$ 间的结构学习就变成了基于 D-分割理论的独立性检验问题:$p(a,b|c\in V-\{a,b\})$是否等于 $p(a|c)p(b|c)$.



(a) (b) (c)

Fig.7 Three situations of D-separation

图 7 D-分割的 3 种情况

D-分割理论是贝叶斯网络的基础,它是一种用来判断变量是否条件独立的图形化方法.对于一个 DAG,D-分割方法可以很快地判断出两个变量是否是条件独立的.D-分割一共有 3 种情况:

第 1 种情况是一个节点连接另外两个节点的箭头尾部,如图 7(a)所示.根据公式(1)和图 7(a)可知:如果 c 是可观测的变量,则 a 和 b 是给定 c 条件独立的;如果 c 不作为观察变量,则 a 和 b 不是给定 c 条件独立的;

第 2 种情况是一个节点分别连接另外两个节点的头部和尾部,如图 7(b)所示.由图可知:如果 c 是可观测变量,则可得 a 和 b 是给定 c 条件独立的;如果 c 不是可观测变量,则可得 a 和 b 不是给定 c 条件独立的;

第 3 种情况是有两个节点都共同的指向第 3 个节点,如图 7(c)所示.如果 c 作为观测变量,则 a 和 b 不是给定 c 条件独立的;如果 c 不作为观察变量,则可得 a 与 b 是独立的.

根据上述的 D-分割理论,Pearl 提出了 do 算子的概念[54].do 的意思可以理解成干预,没有干预的概念,很多时候没有办法谈因果关系.在 DAG 中,$do(X_i)=x_i'$ 表示如下操作:将 DAG 中指向 X_i 的所有有向边全部切断,且将 X_i 的取值固定为常数 x_i'.如此得到新的 DAG 的联合分布可以记为 $p(x_1,...,x_n\,|\,do(X_i)=x_i')$.可以证明,干预后的联合分布为

$$p(x_1,...,x_n\,|\,do(X_i)=x_i')=\frac{p(x_1,...,x_n)}{p(x_i\,|\,pa_i)}I(x_i=x_i') \tag{2}$$

请注意,$p(\cdot\,|\,do(X_i)=x_i')$ 和 $p(\cdot\,|\,X_i=x_i')$ 在很多情况下是不同的.如图 8 所示,在图 8(a)中,$p(B=b|A=a)$ 与 $p(B=b|do(A)=a)$相等.因为 A 是 B 的"原因",所以对于 B 来说在"条件"A 和"干预"A 下的分布相同.但在图 8(b)中,有 $p(B=b|A=a)$与 $p(B=b|do(A)=a)$不等.由于 A 是 B 的"结果"或者说是给定"条件"B 下的"结果",因此给定"结果"条件下"原因"的分布不等于"原因"自身的边缘分布.但人为地"干预"结果 A 并不影响原因 B 的分布,即 $p(B=b|do(A)=a)=p(B=b)$.

(a) (b)

Fig.8 Comparison between do operator and condition

图 8 do 操作和 condition

贝叶斯网络的结构学习问题引起大量学者对于因果结构学习算法的研究.早在1988年,Pearl就提出了图结构学习算法——IC 算法[55].随后,Spirtes 分别于 1990 年、1991 年和 2001 年提出了 SGS 算法[56]、PC 算法[56]和 FCI 算法[57].2010 年,Maier 等人基于 Heckerman 的有向无环概率实体关系模型提出了关系型数据上的 PC 算法,称为 RPC 算法[58].2012 年,Colombo 通过优化 FCI 算法提高其结构学习效率,提出了一种快速结构学习算法 RFCI[59].国内学者也做了一些贝叶斯网络上的理论研究和应用研究并取得了阶段性成果,如:在丢失数据上的贝叶斯网络结构学习研究[60]、用于数据挖掘的贝叶斯网络[61]、基于贝叶斯网络的不确定性知识的推理方法[61]、基于 Markov 毯分解的抽样近似推理算法[62]、基于 FMECA 的复杂装备故障预测贝叶斯网络建模[63]、基于混合方式的贝叶斯网络等价类学习算法[64]等.虽然从 1988 年开始到目前为止陆陆续续地提出了一些 DAG 上的因果结构学习算方法,但是从算法复杂度、排除隐变量的干扰作用和区分马尔科夫等价类这 3 个指标来看,并没有出现非常完美的算法.也就是说,没有同时在这 3 个指标上都表现优秀的算法出现.

2.2.4 虚拟事实模型

1974 年,哈佛大学统计系的 Rubin 提出了一种因果作用模型[65],此模型与 Lewis 的虚拟事实理论(counterfactual)[66]在理论上相似,所以统称为虚拟事实模型.该模型的核心就是引入了一个叫做虚拟结果的结果.比如,我们能同时观测到同一个体在接受处理和未接受处理的两个结果的话,我们就可以使用这两个结果的差异来评价处理对这个个体的因果作用.但是在一般情况下,个体在接受处理和不接受处理两种情况中只能选择一个,要么接受处理,要么不接受处理.例如,我们假设一家医疗单位要测试一种新药对于一种疾病的疗效.如果试吃药物的对象在吃完药后还能再回到和吃药前一模一样的状态,那么我们就可以设置这样的实验:让试药者试吃药物一段时间 T 后记录结果 $R1$,然后让试药者回到吃药前的状态不做任何治疗,时间 T 后记录结果 $R2$.那么分析 $R1$ 与 $R2$ 的差别,就是这种新药对于这种疾病在这个实验对象上的因果作用.显然,这种假设是不合理也是无法实现的,所以那个无法观测到的结果就叫做虚拟结果.基于虚拟事实模型进行观察性研究的因果推断时需要一些假定,而这些假定是无法用观测数据进行检验的.虽然虚拟事实模型的理论很完备,但是由于这些假设使得它在实用性上存在缺陷.虚拟事实模型的理论形式如下所述:

设 Z_i 表示个体 i 接受处理与否,处理取 1,对照取 0;Y_i 表示个体 i 的结果变量.另外,记$\{Y_i(1),Y_i(0)\}$表示个体 i 接受处理或者对照的虚拟结果(potential outcome),那么 $Y_i(1)-Y_i(0)$表示个体 i 接受治疗的个体因果作用.不幸的是,每个个体要么接受处理,要么接受对照$\{Y_i(1),Y_i(0)\}$中必然缺失一半,个体的因果作用是不可识别的.观测的结果是 $Y_i=Z_iY_i(1)+(1-Z_i)Y_i(0)$.但是在 Z 做随机化的前提下,我们可以识别总体的平均因果作用(ACE):

$$ACE(Z\to Y)=E\{Y_i(1)-Y_i(0)\} \tag{3}$$

这是因为:

$$ACE(Z\to Y)=E\{Y_i(1)\}-E\{Y_i(0)\}=E\{Y_i(1)|Z_i=1\}-E\{Y_i(0)|Z_i=0\}=E\{Y_i|Z_i=1\}-E\{Y_i|Z_i=0\} \tag{4}$$

最后一个等式表明,ACE 可以由观测的数据估计出来.其中,第 1 个等式用到了期望算子的线性性质;第 2 个式子用到了随机化,即 $Z\perp\{Y_i(0),Y_i(1)\}$,其中,$\perp$表示独立性.由此可见,随机化实验对于平均因果作用的识别起着至关重要的作用.

2.3 基于因果图的应用研究

由于社会媒体的出现也不过几年,是一个新兴事物,社会媒体上的本质因果分析工作还很有限.在因果分析的几种方法中,目前只有随机对照实验方法和准实验设计的方法用于OSM上因果分析,而且随机对照实验在研究数量和质量上占绝对优势.随机对照实验方法和准实验设计用于 OSM 上因果分析的具体研究工作已在第 2.2.1 节和第 2.2.2 节中详细论述.社会媒体上的常识因果分析的研究工作数量不多,也已经在第 2.1 节中详细论述.但是基于因果图的应用研究却有一定的数量.这些应用研究都是在贝叶斯网络这种表达因果的模型上的基于某些特定任务的研究工作.王飞跃等人指出,可能性而不是确定性是社会计算研究的主要特征[67],而贝叶斯网络正是建模不确定性的有效模型,因此在社会媒体的研究理论方法中,贝叶斯网络具有先天优势,比较有代表性的就是基于贝叶斯网络的推荐系统和基于贝叶斯网络的罕见事件预测.

2.3.1 基于贝叶斯网络的推荐系统

互联网的发展使得信息呈爆炸式的增长,这种增长反而使互联网用户的信息获取效率变低了.这种现象被称为信息过载.如今,有效解决信息过载的工具之一便是个性化推荐.个性化推荐的本质就是代替用户评价、过滤它从不知道的东西.这些东西包括好友、文本、书、电影、CD,甚至可以是饭店、音乐、绘画、美食等等.过去 20 年,个性化推荐算法得到长足的发展和进步,学术界和工业界的研究学者和实践者们从不同的领域出发,提出了各种模型和解决方案.总体上说,个性化推荐所涉及到的学科包括人工智能、机器学习、认知科学、信息抽取、数据挖掘、预测理论、近似理论,甚至是管理科学、市场营销和心理学.所使用的算法除了传统的协同过滤,还包括图模型、链接分析、回归分析、矩阵分解、奇异值分解以及机器学习领域各种分类和学习算法.然而,贝叶斯网络在个性化推荐问题上拥有独特的优势.贝叶斯学习理论将先验知识或背景知识与样本信息相结合、因果语义与概率语义相结合,是数据挖掘和不确定性知识表示的理想模型.贝叶斯网络有很多独特的优势,例如:贝叶斯学习能够方便地处理不完全数据(包括值隐变量和势隐变量);贝叶斯学习能够学习变量间的因果关系;贝叶斯网络具有推理功能,能够很好地处理过拟合问题;贝叶斯网络与贝叶斯统计相结合,能够充分利用领域知识和样本数据的信息.这些优势使它在推荐系统上表现出很强的优势,因为个性化推荐问题很多时候就是隐变量的学习问题.例如,Yang 等人使用基于贝叶斯网络的协同过滤算法提出了一个适用于社交网络的推荐系统[68],该系统在用户关系基础上建立贝叶斯网络模型,预测一个用户对某个电影的评分情况,根据评分给出推荐电影.基于类似的想法,Beutel 等人利用大量的用户对电影的评分数据,把用户 A 和 B 之间对同一电影评价的协同关系归约到一个含有隐变量的贝叶斯网络学习问题上.通过学习隐变量 r(表示用户 A 和 B 对某一电影的共同喜好关系)来获得协同推荐关系[69].从以上的两个工作可以看出:在解决个性化推荐问题的方法中,基于贝叶斯网络学习的方法是一种很重要的方法.

2.3.2 基于贝叶斯网络的罕见事件预测

罕见事件是指一种发生概率很低的事件,例如,公路交通事故、网络欺诈行为、网络入侵行为、信用卡诈骗行为、社会话题爆发等都属于罕见事件.罕见事件的预测是一个非常复杂的问题,它需要对问题本身的深刻理解和对问题中不确定性的建模,不像那些经常或大量发生的事件有很多相似的事件来训练预测模型或者起到协同过滤作用.因为罕见事件发生的概率很低,所以即使存在类似的事件,数量也很少,这对于传统的机器学习算法来说无疑是致命的.机器学习算法大都需要大量的训练数据,训练数据越大,预测效果越好.对于预测罕见事件来说,大量的相关关系或相关事件无疑是奢侈的.因此,对于罕见事件的预测就需要正确的因果知识和因果分析,并且充分利用可以用到的小样本数据.因为因果关系是预测和干预事件最有效的关系,所以贝叶斯网络就成为了解决这个问题的合理选择.另外,贝叶斯网络是非常强大的图模型,它在分析真实数据和寻找变量间关系方面有很大的优势.正是由于贝叶斯网络所包含的因果语义和概率语义,所以它被认为是融合背景知识和真实数据的理想表示模型.

尽管研究者们在罕见事件预测问题上尝试了很多种方法,例如多示例学习算法[70]、基于规则的方法[71]、基于逻辑回归的方法[72]、基于采样的方法[73]、代价敏感的学习[74−76]、Boosting 算法[77−79]、基于分割的算法[80]、Log-linear 模型[81]等,但是以上方法都没有抓住罕见事件的本质特性,即极大的稀疏性和不确定性.本文基于贝叶斯网络的天然优越性,贝叶斯网络模型是预测罕见事件的最合理的模型.基于贝叶斯网络的罕见事件预测也有一些研究成果出现:2010 年,Zhang 等人通过把搜索引擎用户的文档点击问题归约为贝叶斯网络隐变量学习问题,预测用户对搜索引擎返回结果文档的点击[82];2012 年,Shen 等人利用人工构造的贝叶斯网络来预测用户对广告的点击行为[83].在近几年基于因果推断的罕见事件分析也出现了少量的研究成果:Kleinberg 等人在大规模带有时序的数据上对罕见事件进行因果推断,通过建模系统函数模型和评估罕见事件与正常事件的不同,来推断罕见事件的结果[84].

3 问题与挑战

本质因果的推断在统计学中不是一个新问题,但是真正完全的因果推断算法(同时满足 3 个指标)和可用的

因果推断系统却尚未出现.同时,如何客观地评价分析得到的因果关系也仍旧没有统一的黄金法则.常识因果的抽取、分析和推理也还存在很多瓶颈问题.如下所述正是以上领域中存在一些具体问题和挑战.

3.1 因果对识别

因果对识别是因果分析的基本问题,既包含从观测数据上对本质因果关系对的识别分析,又包含在大量人造文本上对常识因果的识别.对于常识因果来说,对因果关系对的识别除了使用规则和模板的方法外,还没有发展出比较有效的方法.使用因果触发词的因果对识别在正规的新闻语料中效果较好,USC 的 Kozareva,Gordon 和 Roemmele 等人做过相关工作,准确率接近 90%[30,31,85].UIUC 的 Girju 等人想通过研究社会媒体上语言上的因果关系来分析社会媒体上的个人间的关系[29].Radinsky 等人通过抽取新闻预料中历史事件的因果关系来对未来发生事件进行预测[32].上述研究包括了词与词之间的因果对、词组短语之间的因果对、事件之间的因果对的抽取.他们使用的抽取方法都是使用基于因果触发词的规则方法,这种方法对于语法句法正规的预料来说效果不错,但是对于文字表达不规范的社会媒体文本效果非常不好.主要原因在于社会媒体文本的随意性,即使表达因果逻辑,也很少使用表示因果逻辑的触发词.基于这种情况,就需要一种更具普遍性的文本中的因果抽取方法.

本质因果对的识别更是因果对识别问题中的困难问题.所有因果关系结构学习算法和因果分析算法都依赖于因果对的准确识别.NIPS 2013 Workshop on Causality: Large-scale Experiment Design and Inference of Causal Mechanisms 专门设置了一个关于因果对识别的挑战竞赛[86].竞赛提供了包括变量之间因果关系和非因果关系的几百个真实数据样本集,这些数据样本集来自化学、气候、生态、经济、工程、流行病学、基因组学、医学、物理学和社会学等不同领域,要求参加竞赛的小组在这些样本集合中识别出哪些变量之间是因果关系.

在没有其他辅助信息(如时序信息等)的因果推断中,只给定 $P(X,Y)$,我们能够推断出 $X\to Y$ 还是 $Y\to X$?在没有任何第 3 个变量的情况下,这个推断是非常困难的.给定如图 9 的散点图,能否推断是 A 导致 B 还是 B 导致 A,或还是 A 与 B 之间根本就没有因果关系(图 9 中的散点是真实系统中存在因果关系的变量产生的,因此 A 与 B 之间存在因果关系).假定 A 与 B 没有共同的直接或者间接的共同原因的前提下,到底是 $A\to B$ 还是相反?如果基于概率论的知识我们无从知晓,因为在概率论中,关于 A 和 B 有太多的对称性:$p(B)p(A|B)=p(A)p(B|A)$.要想正确地识别 A 与 B 之间的因果方向,我们必须找到一个能够反映这种因果关系的函数,若 $A\to B$ 则 $B=f(\cdot)$且 $f(\cdot)$不可逆;反之亦然.但是,找到一个能很好地拟合数据又能反映因果关系的不可逆函数是非常困难的事情.

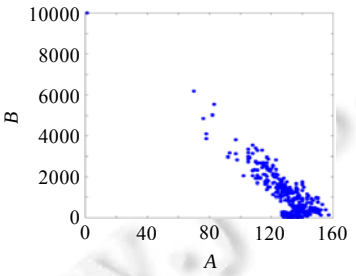


Fig.9 Scatterplot of A and B

图 9 两个变量 A 与 B 的散点图

在有上下文信息的因果对识别中,存在其他变量的因果对识别中,除 A 与 B 之外的其他变量对 A 与 B 之间的因果推断存在干扰.如果其他变量都是可观测的,那么我们可以运用 D-分割理论检验在给定其他所有变量的前提下 A 与 B 的条件独立性来判断 A 与 B 之间的因果关系,但是如果存在不可观测的隐变量作为 A 与 B 的直接共同原因或者间接共同原因,如图 10 所示,那么即使 A 与 B 之间本身不存在因果关系,也会得出 A 与 B 之间存在因果关系,这就是所谓的混淆变量对因果推断的干扰.如何解决这个问题,很多学者进行了深入研究[17,50,52,57,87,88],但是对于排除这种混淆因素对因果推断的干扰,还都既不彻底也不完全.

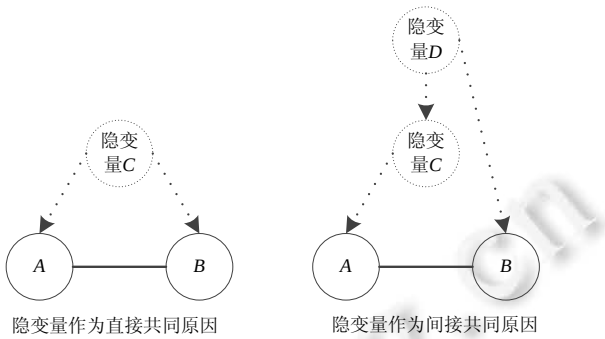


Fig.10 Causal inference between A and B in the case of hidden variables

图 10 存在隐变量情况下变量 A 与 B 之间的因果推断

3.2 因果分析模型的适用性

目前,几乎所有的观测数据上因果分析模型的能力都止步于浅层因果的分析,比如在维基百科上文章的编辑数量是文章质量的原因等.对于分析深层次的原因,现有的因果分析模型显得力不从心.究其原因主要有:(1) 因果分析对观测数据要求高;(2) 涉及的相关变量需要人为给出;(3) 因果分析需要很多前提假设,如faithfulness 和等价类;(4) 无法客观地评价因果分析结论的正确性.

首先,无论是用准实验设计的方法还是用图模型的方法在观测数据上进行因果推断,对数据都有很高的要求.准实验设计的方法需要在观测数据上能找到满足除控制变量外其他变量都相同或相似的数据对,而图模型的方法需要观测数据上潜在因果的变量间有观测到的强相关性、不存在其他隐性的混淆变量且存在时序或逻辑上的先后顺序;

其次,因果分析涉及的相关变量都要人为的给出,因为这是因果推断的前提所以相关的变量的覆盖度和准确性就有赖人力;

再者,因果分析特别是图模型方法的因果分析需要很多的前提假设,faithfulness 就是其中最重要的前提.在图模型中,D-分割蕴含着这样的条件独立性:如果 C 阻断了从 A 到 B 的所有路径,那么我们说 $A\perp B|C$;反之,要推出如果 $A\perp B|C$ 那么 C 阻断了从 A 到 B 的所有路径,就需要额外的假设.这个假设主要涉及条件概率分布的可识别性问题,一般就把概率独立性和因果条件关联起来所需要的涉及条件概率分布可识别性的假设叫做faithfulness.一个统计模型是可识别的,一般指在没有观测到的变量分布的改变的前提下,统计模型肯定没有变化.如果我们在改变部分统计模型后无法观测到观测变量分布的变化,我们称该部分统计模型不可识别.虽然我们在所因果推断的过程中已经对模型做了可识别性的假设,但是在因果结构学习中仍存在等价类问题,即不同的因果结构却对应到相同的概率分布上.假设存在 3 种因果结构:$A\rightarrow C\rightarrow B$;$A\leftarrow C\leftarrow B$;$A\leftarrow C\rightarrow B$.但是根据 D-分割理论,这 3 种因果结构却都对应到相同的条件独立性上,即,A 与 C 不独立,B 与 C 不独立,$A\perp B|C$.我们称这 3 个因果结构为等价类.等价类是因果结构学习中的干扰因果结构准确性的关键问题.

3.3 客观的评价策略

传统的不涉及因果的概率模型的一般评价方法都是在一个测试集上检验模型的准确率和召回率,但是如果用这种方法来评价因果模型,是无法评价因果语义的.现在有很多文章在人造数据上评价因果模型,前提是潜在的因果结构已经知晓,或者在已经知道因果结构的真实系统产生的真实数据上评价因果模型.还有一种方法就是在可干预的系统中直接评价学习到的因果模型,现在有一些少量的文章使用这种方法评价因果模型.对于社会媒体上的因果分析来说,只能使用真实的数据集来验证评价模型的正确性和因果结论的正确性.使用在已知因果结构的数据上训练出的因果模型,用于未知因果结构数据上的因果关系分析.对于未知因果结构的数据上学习到的因果结构目前还没有很好的方法以验证结论的正确性,只能依赖于因果分析过程的正确性或者基于因果结论干预的效果,间接评价学习到的因果结构的正确性.

3.4 社会媒体上的因果分析

相对于传统数据上的因果抽取与分析,社会媒体上的因果抽取和因果分析有其特殊性.这些特殊性源于社会媒体的大数据特性和变量类型的多样性.社会媒体是属于普罗大众的网络平台,其中承载了大量的文本数据、用户的关系数据、用户行为数据和信息传播数据等.普罗大众中的每个成员都可能是数据的制造者.社会媒体的上述特征决定了社会媒体上的数据庞杂.拿文本内容来说,社会媒体上的文本内容具有多样化、噪音多、篇幅短小、内容形式书写自由而且常常出现一些新的语言表达方式等特点.要处理这种不规范的文本,对现有的信息抽取相关技术,无论分词、词性标注还是句法分析,都是一个重大的挑战.现有的文本内容上的常识因果抽取主要是在分词、词性标注等自然语言处理技术基础上的,因此在这种不规范的社会媒体文本上抽取因果知识是极大的挑战.再者,社会媒体上变量类型的种类很多,其中包括网络结构、社会影响力、用户自身特征、文本内容等.因此,在分析社会媒体上的因果关系时就要综合考虑上述变量.然而,不同类型的变量如何融合、统一评价从而发现因果关系,是一个仍需继续讨论的问题.由于变量类型如此丰富,那么在探求一个特定结果的原因时可能会得到多个原因.例如,社会媒体上用户购买产品行为的原因就多种多样,有的是因为产品质量好,有的是因为受别人的影响,有的是由于产品广告做得好.但是在这些原因中,哪个是主要原因?每个原因对结果(即购买产品这种行为)的作用强度如何?这些都是需要进一步研究和待完善的问题.

4 未来研究方向

随着越来越多的学者对因果关系问题的关注,出现了一些新的理论和方法,但对于解决实际问题仍然有很大的局限性.另外,本文综合考虑社会媒体和几种因果分析方法模型的特点,预测因果分析在社会媒体上的研究热点主要还是社会媒体上因果知识的识别、抽取和基于因果图模型的应用研究.

4.1 理论方法

因为因果推断理论本身还有许多问题有待进一步的探讨,所以在未来若干年内,因果分析理论本身的完善和发展仍将是未来研究的热点.基于对因果分析理论本身存在的问题和挑战的分析,我们归纳了因果分析理论本身的几个未来研究的热点方向:融合准实验设计和因果图的因果分析方法、融合虚拟事实模型和因果图模型的因果分析理论、融合随机化实验和因果图模型的因果分析方法、融合常识因果和本质因果的深层因果分析.

4.1.1 融合准实验设计和因果图的因果分析方法

准实验设计的方法在因果分析上有其独特的优越性:与随机对照实验相比,它的外部合理性更强;与图模型上的因果分析相比,它的内部合理性更强.但是要想在观测的非实验数据上做因果分析推断,只能使用准实验设计方法和因果图这两种方法.再者,准实验设计的方法有很强的方法相容性,所以融合准实验设计和因果图的因果分析方法必然是因果分析领域未来的研究方向.而且在大数据的条件下,能够使准实验设计方法达到和随机对照实验一样的内部合理性.所以一种可能的结合方式是:使用准实验的设计分析每对变量之间的因果关系,然后使用因果图模型简化和检验变量间的因果关系,并用于因果推理.

4.1.2 融合随机对照实验和因果图模型的因果分析方法

随机对照实验作为因果分析的黄金法则,具有其他方法无法取代的内部合理性.因此要得到最可靠的因果结论,随机对照实验是优先考虑的方法.虽然随机对照实验在于因果对识别上具有无可比拟的优越性,但是在多变量间的因果分析问题上却存在严重的效率问题和可行性问题.因为如果只是使用随机对照实验的方法,就需要枚举把所有涉及的相关变量间的组合,且针对每个组合设计随机对照实验,显然这是不现实的.因此,融合图模型尤其是因果图模型的多变量因果分析,也就成了随机对照实验在分析多变量因果关系上的必由之路.

4.1.3 融合常识因果和本质因果的深层因果分析

对于这个问题,可以类比人类在认识事物本质上的过程:人们在认识一个新事物时,通常是在之前的知识积累的前提下总结新规律认识新问题.例如:假如没有牛顿的万有引力定律,科学家们就不可能发展出后来的那些关于宇宙的科学理论;假如没有达尔文的进化论,后来的那些医学上的遗传学发现也就无从谈起了.其实,万有

引力定律和进化论对于后来的发现来说就是常识因果,因为这些最初的本质因果理论都已经阐述在文字中,使人们可以直接地获取得到而不用再通过观测分析数据或者实验来探索发现它们.人类认识事物的规律是:最初的本质因果都会成为常识因果,而对新问题的认识都必须基于之前的常识因果.人工智能的发展轨迹一直都在模拟人的智能,企图使机器智能接近或者达到人类智能.那么作为人工智能的重要问题——因果推断来说,模拟人类认识问题的规律和过程也是理所当然的思路.所以,融合常识因果和本质因果的深层因果分析推断,将是未来因果分析推断理论研究的重要思路和方向.

4.2 社会媒体上的应用

大数据时代的背景,如何从海量数据中发现知识,寻找隐藏在数据中的模式、趋势和生成机制,揭示社会现象与预知社会发展规律,需要我们拥有更好的数据洞察力.随着社交网络、移动互联网和物联网的兴起,大数据会越变越大,网络科学和数据科学提供了新的科学发展观和方法论.大量的研究实践已经表明:基于相关性分析的一些数据挖掘手段和机器学习方法,对于社会媒体上产生的大数据分析起到了重要的作用.但是,这些基于相关性的方法对于深刻地理解大数据并从中揭示本质的社会规律和人类行为模式等问题却存在缺陷的.具体来说,基于相关性分析的方法预测能力强,但是解释和控制干预能力弱.也就是说,基于相关性分析的方法能够很好地预测未知和未来事物,但是对于解释已有事物和在分析已有事物的基础上反馈控制和干预上的可信度却非常低.恰好,因果分析和推断理论在这方面却有着天然的优势.

4.2.1 因果分析

到目前为止,社会媒体上因果推断的尝试只运用到了两种因果分析方法:随机对照实验和准实验设计.随机对照实验的方法从 2009 年开始的近几年内在社会媒体上产生了一批重大的理论成果,大都发表在《Nature》,《Science》和《PNAS》上,主要以 Christakis,Centola 和 Aral 等美国年轻学者为代表.Leskovec[13,89,90],Kleinberg[12],Jensen[18,48]和唐杰[4]等人则在观测数据上使用 QEDs 理论在社会媒体上发现一些社会媒体的本质规律.但是在使用因果图模型和虚拟事实模型用于社会媒体上的研究还属于空白.并且由于社会媒体语言文本的新特点,传统的常识因果分析方法对于解决社会媒体任务的效果不尽如人意.我们预测:社会媒体上的因果分析和推断研究必将成为一个值得研究方向,其中包括社会媒体上常识因果分析、浅层因果分析和深层因果分析.

4.2.2 基于因果知识的预测和推理

在 OSM 上的行为分析、社区挖掘、影响力分析、链接预测等多个领域的研究中,基于因果知识的预测和推理所利用的因果关系大都是从专家经验和背景知识中直接得到.一个典型的思路就如第 2.3 节中所介绍的那样,把预测问题或者推荐问题归约到一个对含有隐变量的贝叶斯网络学习问题上.这个含有隐变量的贝叶斯网络就是从专家经验和背景知识中直接人为构建的.这种解决预测问题或推荐问题的思路有一个非常苛刻的前提条件,即对待解决问题本身有非常深刻的理解.也就是说,这种方法的使用门槛很高,需要花费大量的时间来理解问题本身.而且这样获得的解决问题的模型(人工构造的贝叶斯网络)只能解决这一个问题,即迁移性和泛化能力非常差.目前也存在一些特定结构的、泛化能力很强的贝叶斯网络模型,如朴素贝叶斯网络、TAN(增广树贝叶斯网络)[91]、HNB(隐性朴素贝叶斯网络)[92]等模型,但是这些模型由于限定网络结构,所以对问题的适应性较差.如果能够根据训练样本数据自动学习出针对每个特定问题的贝叶斯网络,那么就可以同时解决迁移能力差和适应性差的缺点.这样,很多问题无需专家介入就能直接得到预测结果或者推荐结果.例如,疾病的自动诊断和自动干预治疗.这种自动学习的方法看起来是非常有吸引力的,但是,这种自动学习的方法在涉及样本数据的因果对判别、贝叶斯网络的结构学习和根据特定目标的参数学习这 3 个主要的问题上,贝叶斯网络结构学习问题是 NP 难问题[53],因果对识别也还存在很大的错误率.虽然现实如此不尽如人意,但是本文相信,针对不同特定问题的贝叶斯网络自动学习、推理和预测问题是非常值得研究的问题.

5 结 论

本文分析了因果分析与推断的基本理论、存在的问题、目前的研究现状和未来的研究方向,以及因果分析理论和方法在社会媒体上的应用.在社会媒体上的因果分析包括两个方面:常识因果分析和本质因果分析.常识

因果分析手段需要借助自然语言和信息抽取的技术和方法.本质因果分析的理论和方法主要集中在随机对照实验方法、准实验设计方法、因果图模型、虚拟事实模型这 4 种方法.这 4 种方法都存在各自的优缺点,实验数据的因果分析一般使用第 1 种方法,而观测数据的因果分析使用后 3 种方法.不管社会媒体上的常识因果分析还是本质因果分析都还处在理论和实践的起步阶段,存在大量机遇的同时也面临了很多挑战.总之,社会媒体上因果分析与推断是一个理论性与实践性都很强且是一个充满机遇与挑战的研究方向,所以必定会吸引更多的学者的关注和研究.

致谢 在此,我们向对本文的研究工作提供帮助的老师和同学表示感谢,尤其感谢赵妍妍、郭江、赵思成和北京大学计算机系的汪定给本文提出了很多宝贵的修改意见.

References:

[1] Pearl J. Causality: Models, Reasoning and Inference. Cambridge: Cambridge University Press, 2000.

[2] Xu ZH, Zhang Y, Wu Y, Yang Q. Modeling user posting behavior on social media. In: Proc. of the SIGIR 2012. New York: ACM Press, 2012. 545−554. [doi: 10.1145/2348283.2348358]

[3] Yang Z, Guo JY, Cai KK, Tang JZ, Li J, Zhang L, Su Z. Understanding retweeting behaviors in social networks. In: Proc. of the CIKM 2010. New York: ACM Press, 2010. 1633−1636. [doi: 10.1145/1871437.1871691]

[4] Zhang J, Liu B, Tang J, Chen T, Li JZ. Social influence locality for modeling retweeting behaviors. In: Proc. of the IJCAI 2013. Menlo Park: AAAI Press, 2013. 2761−2767.

[5] Hopcroft J, Lou TC, Tang J. Who will follow you back? Reciprocal relationship prediction. In: Proc. of the CIKM 2011. New York: ACM Press, 2011. 1137−1146. [doi: 10.1145/2063576.2063740]

[6] Even-Dar E, Kearns M, Suri S. A network formation game for bipartite exchange economies. In: Proc. of the SODA 2007. Philadelphia: SIAM, 2007. 697−706. [doi: 10.1145/1283383.128345]

[7] Guo SH, Wang MQ, Leskovec J. The role of social networks in online shopping: Information passing, price of trust, and consumer choice. In: Proc. of the 12th ACM Conf. on Electronic Commerce. New York: ACM Press, 2011. 157−166. [doi: 10.1145/1993574.1993598]

[8] Henkel J, Block J. Peer influence in network markets: A theoretical and empirical analysis. Journal of Evolutionary Economics, 2013,23(5):925−953. [doi: 10.1007/s00191-012-0302-4]

[9] Lin YH, Chen CY. Adolescents' impulse buying: Susceptibility to interpersonal influence and fear of negative evaluation. Social Behavior and Personality: An Int'l Journal, 2012,40(3):353−358.

[10] Ver Steeg G, Galstyan A. Information transfer in social media. In: Proc. of the WWW 2012. New York: ACM Press, 2012. 509−518. [doi: 10.1145/2187836.2187906]

[11] Bottou L, Peters J, Onero-Candela J, Charles DX, Chickering DM, Portugaly E, Ray D, Simard P, Snelson E. Counterfactual reasoning and learning systems: The example of computational advertising. Journal of Machine Learning Research, 2013,14(10):3207−3260.

[12] Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S. Feedback effects between similarity and social influence in online communities. In: Proc. of the KDD 2008. New York: ACM Press, 2008. 160−168. [doi: 10.1145/1401890.1401914]

[13] Lewis K, Gonzalez M, Kaufman J. Social selection and peer influence in an online social network. Proc. of the National Academy of Sciences, 2012,109(1):68−72. [doi: 10.1073/pnas.1109739109]

[14] La Fond T, Neville J. Randomization tests for distinguishing social influence and homophily effects. In: Proc. of the WWW 2010. New York: ACM Press, 2010. 601−610. [doi: 10.1145/1772690.1772752]

[15] Toulis P, Kao E. Estimation of causal peer influence effects. In: Proc. of The ICML 2013. New York: ACM Press, 2013. 1489−1497.

[16] Aral S, Walker D. Identifying social influence in networks using randomized experiments. IEEE Intelligent Systems, 2011,26(5):91−96. [doi: 10.1109/MIS.2011.89]

[17] Rattigan MJH, Maier M, Jensen D. Relational blocking for causal discovery. In: Proc. of the AAAI 2011. Menlo Park: AAAI Press, 2011. 145−151.

[18] Oktay H, Taylor BJ, Jensen DD. Causal discovery in social media using quasi-experimental designs. In: Proc. of the KDD Workshop on Social Media Analytics. New York: ACM Press, 2010. 1–9. [doi: 10.1145/1964858.1964859]

[19] Rattigan MJH, Jensen D. Leveraging *D*-separation for relational data sets. In: Proc. of the ICDM 2010. Washington: IEEE Computer Society, 2010. 989−994. [doi: 10.1109/ICDM.2010.142]

[20] 休谟,著;关文运,译.人类理解研究.香港:香港出版贸易公司,1972.

[21] 康德,著;邓晓芒,译.纯粹理性批判.北京:人民出版社,2011.

[22] 黑格尔.小逻辑.北京:光明日报出版社,2009.

[23] 马克思,恩格斯.著;中共中央马克思恩格斯列宁斯大林著作编译局,编译.马克思恩格斯全集(第 2 版).北京:人民出版社,2007.

[24] Freedman D, Pisani R, Purves R. Statistics. 4th ed., New York: WW Norton & Co., 2012.

[25] Ittoo A, Bouma G. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In: Muñoz R, Montoyo A, Métais E, eds. Proc. of the 16th Int'l Conf. on Applications of Natural Language to Information Systems. Berlin, Heidelberg: Springer-Verlag, 2011. 52−63. [doi: 10.1007/978-3-642-22327-3_6]

[26] Ittoo A, Bouma G. Minimally-Supervised extraction of domain-specific part—Whole relations using Wikipedia as knowledge-base. Data & Knowledge Engineering, 2013,85:57−79. [doi: 10.1016/j.datak.2012.06.004]

[27] Sorgente A, Vettigli G, Mele F. Automatic extraction of cause-effect relations in natural language text. In: Proc. of the 13th Conf. of the Italian Association for Artificial Intelligence (AI*IA 2013). 2013. 37−48.

[28] Girju R. Automatic detection of causal relations for question answering. In: Proc. of the ACL Workshop on Multilingual Summarization and Question Answering. Morristown: ACL, 2003. 76−83. [doi: 10.3115/1119312.1119322]

[29] Girju R. Toward social causality: An analysis of interpersonal relationships in online blogs and forums. In: Proc. of the ICWSM 2010. Menlo Park: AAAI Press, 2010. 66−73.

[30] Gordon AS, Bejan CA, Sagae K. Commonsense causal reasoning using millions of personal stories. In: Proc. of the AAAI 2011. Menlo Park: AAAI Press, 2011. 1180−1185.

[31] Kozareva Z. Cause-Effect relation learning. In: Proc. of the TextGraphs Workshop at ACL. Morristown: ACL, 2012. 39−43.

[32] Radinsky K, Davidovich S, Markovitch S. Learning causality for news events prediction. In: Proc. of the WWW 2012. New York: ACM Press, 2012. 909−918. [doi: 10.1145/2187836.2187958]

[33] Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. 2nd ed,. Wadsworth: Wadsworth Publishing, 2002.

[34] Campbell DT, Stanley JC, Gage NL. Experimental and Quasi-Experimental Designs for Research. Houghton Mifflin Boston, 1963.

[35] Thyer BA. Quasi-Experimental Research Designs. Oxford: Oxford University Press, 2012.

[36] Sriperumbudur BK, Gretton A, Fukumizu K, Schölkopf B, Lanckriet GR. Hilbert space embeddings and metrics on probability measures. The Journal of Machine Learning Research, 2010,99:1517−1561.

[37] Székely GJ, Rizzo ML. Brownian Distance Covariance. The Annals of Applied Statistics, 2009. 1236−1265.

[38] Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. The Journal of Machine Learning Research, 2012,13:723−773.

[39] Zhang K, Peters J, Janzing D, Schölkopf B. Kernel-Based conditional independence test and application in causal discovery. In: Proc. of the UAI 2011. Corvallis: AUAI Press, 2011. 804−813.

[40] Boring EG. The nature and history of experimental control. The American Journal of Psychology, 1954,67(4):573−589. [doi: 10.2307/1418483]

[41] Fisher SRA, Genetiker S, Fisher RA, Genetician S, Britain G, Fisher RA, Généticien S. Statistical Methods for Research Workers. Oliver and Boyd Edinburgh, 1970.

[42] Centola D. The spread of behavior in an online social network experiment. Science, 2010,329(5996):1194−1197. [doi: 10.1126/science.1185231]

[43] Centola D. An experimental study of homophily in the adoption of health behavior. Science, 2011,334(6060):1269−1272. [doi: 10.1126/science.1207055]

[44] Muchnik L, Aral S, Taylor SJ. Social influence bias: A randomized experiment. Science, 2013,341(6146):647−651. [doi: 10.1126/science.1240466]

[45] Cook TD, Campbell DT, Day A. Quasi-Experimentation: Design & Analysis Issues for Field Settings. Houghton Mifflin Boston, 1979.

[46] Knuth DE, Morris JJH, Pratt VR. Fast pattern matching in strings. SIAM Journal on Computing, 1977,6(2):323−350. [doi: 10.1137/0206024]

[47] Linden A, Trochim WM, Adams JL. Evaluating program effectiveness using the regression point displacement design. Evaluation & the Health Professions, 2006,29(4):407−423. [doi: 10.1177/0163278706293402]

[48] Jensen DD, Fast AS, Taylor BJ, Maier ME. Automatic identification of quasi-experimental designs for discovering causal knowledge. In: Proc. of the KDD 2008. New York: ACM Press, 2008. 372−380. [doi: 10.1145/1401890.1401938]

[49] Jensen DD. Beyond prediction: Directions for probabilistic and relational learning. In: Proc. of the ILP 2007. Berlin: Springer-Verlag, 2007. 4−21. [doi: 10.1007/978-3-540-78469-2_2]

[50] Hyttinen A, Eberhardt F, Hoyer PO. Learning linear cyclic causal models with latent variables. Journal of Machine Learning Research, 2012,13:3387−3439.

[51] Eberhardt F, Hoyer PO, Scheines R. Combining experiments to discover linear cyclic models with latent variables. In: Proc. of the AI Statistics 2010. 2010. 185−192.

[52] Hyttinen A, Eberhardt F, Hoyer PO. Causal discovery for linear cyclic models with latent variables. In: Proc. of the PGM 2010. 2010.

[53] Zhang LW, Guo HP. Introduction to Bayesian Networks. Beijing: Science Press, 2007 (in Chinese).

[54] Pearl J. Causal diagrams for empirical research. Biometrika, 1995,82(4):669−688. [doi: 10.1093/biomet/82.4.669]

[55] Geiger D, Verma TS, Pearl J. *d*-Separation: From theorems to algorithms. In: Proc. of the UAI'89. Amsterdam: North-Holland Publishing, 1989.

[56] Peter S, Clark G, Richard S. Causation, Prediction, and Search. Cambridge: The MIT Press, 2001.

[57] Spirtes P. An anytime algorithm for causal inference. In: Proc. of the AI Statistics 2001.

[58] Maier ME, Taylor BJ, Oktay H, Jensen D. Learning causal models of relational domains. In: Proc. of the AAAI 2010. Menlo Park: AAAI Press, 2010. 531−538.

[59] Colombo D, Maathuis MH, Kalisch M, Richardson TS. Learning high-dimensional directed acyclic graphs with latent and selection variables. The Annals of Statistics, 2012,40(1):294−321. [doi: 10.1214/11-AOS940]

[60] Wang SC, Yuan SM. Research on learning bayesian networks structure with missing data. Ruan Jian Xue Bao/Journal of Software, 2004,15(7):1042−1048 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/1042.htm

[61] M CD, Dai JB, Ye J. Bayesian network for data mining. Ruan Jian Xue Bao/Journal of Software, 2000,11(5):660−666 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/11/660.htm

[62] Wang H, Cao LY, Yao HL, Li JZ. A Sampling Approximate Inference Algorithm Based on Decomposition of Markov Blanket. Pattern Recognition and Artificial Intelligence, 2013(8):729−739 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-6059.2013.08.004]

[63] Cai ZQ, Sun SD, Si SB, Wang N. Modeling of failure prediction Bayesian network based on FMECA. Systems Engineering—Theory & Practice, 2013,33(1):187−193 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-6788.2013.01.022]

[64] Zhu MM, Liu SY, Yang YL. Structural learning Bayesian network equivalence classes based on a hybrid method. Acta Electronica Sinica, 2013,41(1):98−104 (in Chinese with English abstract).

[65] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 1974,66(5):688. [doi: 10.1037/h0037350]

[66] David K. Counterfactuals. Cambridge: Harvard University Press, 1973.

[67] Wang FY, Li XC, Mao WJ, Wang T. Social Computing Methods and Applications. Hangzhou: Zhejiang University Press, 2013 (in Chinese).

[68] Yang X, Guo Y, Liu Y. Bayesian-Inference-Based recommendation in online social networks. IEEE Trans. on Parallel and Distributed Systems, 2013,24(4):642−651. [doi: 10.1109/TPDS.2012.192]

[69] Beutel A, Murray K, Faloutsos C, Smola AJ. CoBaFi: Collaborative bayesian filtering. In: Proc. of the WWW 2014. New York: ACM Press, 2014. 97−108. [doi: 10.1145/2566486.2568040]

[70] Weiss GM, Hirsh H. Event prediction: Learning from ambiguous examples. In: Proc. of the Working Notes of the NIPS'98 Workshop on Learning from Ambiguous and Complex Examples. 1998.

[71] Vilalta R, Ma S. Predicting rare events in temporal domains. In: Proc. of the ICDM 2003. Washington: IEEE Computer Society, 2002. 474−481. [doi: 10.1109/ICDM.2002.1183991]

[72] King G, Zeng L. Logistic regression in rare events data. Political Analysis, 2001,9(2):137−163. [doi: 10.1093/oxfordjournals.pan.a004868]

[73] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. Mining data with rare events: A case study. In: Proc. of the ICTAI 2007. 2007. 132−139.

[74] Weiss GM. Mining with rarity: A unifying framework. ACM SIGKDD Explorations Newsletter, 2004,6(1):7−19.

[75] Zhao J, Li X, Dong Z. Online rare events detection. In: Zou Z, Li H, Yang Q, eds. Proc. of the PAKDD 2007. Springer Berlin Heidelberg, 2007. 1114−1121. [doi: 10.1007/978-3-540-71701-0_126]

[76] Sun Y, Kamel MS, Wang Y. Boosting for learning multiple classes with imbalanced class distribution. In: Proc. of the ICDM 2006. Washington: IEEE Computer Society, 2006. 592−602. [doi: 10.1109/ICDM.2006.29]

[77] Joshi MV, Kumar V, Agarwal RC. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In: Proc. of the ICDM 2001. Washington: IEEE Computer Society, 2001. 257−264.

[78] Joshi MV, Agarwal RC, Kumar V. Predicting rare classes: Can boosting make any weak learner strong? In: Proc. of the KDD 2002. New York: ACM Press, 2002. 297−306. [doi: 10.1145/775047.775092]

[79] Chawla N, Lazarevic A, Hall L, Bowyer K. SMOTEBoost: Improving prediction of the minority class in boosting. In: Lavrač N, Gamberger D, Todorovski L, *et al*., eds. Proc. of the PKDD 2003. Berlin, Heidelberg: Springer-Verlag, 2003. 107−119. [doi: 10.1007/978-3-540-39804-2_12]

[80] Wu J, Xiong H, Wu P, Chen J. Local decomposition for rare class analysis. In: Proc. of the KDD 2007. New York: ACM Press, 2007. 814−823. [doi: 10.1145/1281192.1281279]

[81] Agarwal D, Agrawal R, Khanna R, Kota N. Estimating rates of rare events with multiple hierarchies through scalable log-linear models. In: Proc. of the KDD 2010. New York: ACM Press, 2010. 213−222. [doi: 10.1145/1835804.1835834]

[82] Zhang Y, Chen W, Wang D, Yang Q. User-Click modeling for understanding and predicting search-behavior. In: Proc. of the KDD 2011. New York: ACM Press, 2011. 1388−1396. [doi: 10.1145/2020408.2020613]

[83] Shen S, Hu B, Chen W, Yang Q. Personalized click model through collaborative filtering. In: Proc. of the WSDM 2012. New York: ACM Press, 2012. 323−332. [doi: 10.1145/2124295.2124336]

[84] Kleinberg S. Causal inference with rare events in large-scale time-series data. In: Proc. of the IJCAI 2013. Menlo Park: AAAI Press, 2013. 1444−1450.

[85] Roemmele M, Bejan CA, Gordon AS. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In: Proc. of the AAAI Spring Symp.: Logical Formalizations of Commonsense Reasoning. 2011.

[86] Guyon I. Causality Challenge #3: Cause-effect pairs. 2013.

[87] Maier M, Marazopoulou K, Arbour D, Jensen D. A sound and complete algorithm for learning causal models from relational data. In: Proc. of the UAI 2013. Oregon: AUAI Press, 2013.

[88] Allman ES, Rhodes JA, Stanghellini E, Valtorta M. Identifiability of binary directed graphical models with hidden variables. In: Proc. of the UAI 2013. Oregon: AUAI Press, 2013.

[89] Guo S, Wang M, Leskovec J. The role of social networks in online shopping: information passing, price of trust, and consumer choice. In: Proc. of the 12th ACM Conf. on Electronic Commerce. New York: ACM Press, 2011. 157−166. [doi: 10.1145/1993574.1993598]

[90] Anderson A, Huttenlocher D, Kleinberg J, Leskovec J. Effects of user similarity in social media. In: Proc. of the WSDM 2012. New York: ACM Press, 2012. 703−712. [doi: 10.1145/2124295.2124378]

[91] Friedman N, Koller D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Machine Learning, 2003,50(1-2):95−125. [doi: 10.1023/A:1020249912095]

[92] Jiang L, Zhang H, Cai Z. A novel Bayes model: hidden naive Bayes. IEEE Trans. on Knowledge and Data Engineering, 2009, 21(10):1361−1371. [doi: 10.1109/TKDE.2008.234]

附中文参考文献:

[53] 张连文,郭海鹏.贝叶斯网络引论.北京:科学出版社,2007.

[60] 王双成,苑森淼.具有丢失数据的贝叶斯网络结构学习研究.软件学报,2004,15(7):1042−1048. http://www.jos.org.cn/1000-9825/15/1042.htm

[61] 慕春棣,戴剑彬,叶俊.用于数据挖掘的贝叶斯网络.软件学报,2000,11(5):660−666. http://www.jos.org.cn/1000-9825/11/660.htm

[62] 王浩,曹龙雨,姚宏亮,李俊照.基于 Markov 毯分解的抽样近似推理算法.模式识别与人工智能,2013,(8):729−739. [doi: 10.3969/j.issn.1003-6059.2013.08.004]

[63] 蔡志强,孙树栋,司书宾,王宁.基于 FMECA 的复杂装备故障预测贝叶斯网络建模.系统工程理论与实践,2013,33(1):187−193. [doi: 10.3969/j.issn.1000-6788.2013.01.022]

[64] 朱明敏,刘三阳,杨有龙.基于混合方式的贝叶斯网络等价类学习算法.电子学报,2013,41(1):98−104.

[67] 王飞跃,李晓晨,毛文吉,王涛.社会计算的基本方法与应用.杭州:浙江大学出版社,2013.

赵森栋(1987－),男,山东临沂人,博士生,主要研究领域为社会计算,自然语言处理,因果发现.
E-mail: sdzhao@ir.hit.edu.cn

刘挺(1972－),男,博士,教授,博士生导师,CCF 理事,主要研究领域为社会计算,信息检索,自然语言处理.
E-mail: tliu@ir.hit.edu.cn