大数据中的因果关系发现

蔡瑞初　郝志峰　著

科学出版社

大数据中的因果关系发现

蔡瑞初　郝志峰　著

本书得到广东省自然科学杰出青年基金项目“高维数据因果关系发现理论及其应用”资助(No.2014A030306004)

科学出版社

北京

内 容 简 介

因果关系严格区分了“因”变量和“果”变量，在揭示事物发生机制、指导干预行为等方面具有相关关系不可替代的重要作用。在大数据时代，如何探索海量、高维、观察性的数据背后的因果机制具有重要的商业价值和科学意义。观察数据的因果关系方向判断困难、高维数据的因果结构发现能力不足、现有算法适用场景有限等仍然严重阻碍着因果推断领域的发展，是当前因果关系研究的难点和热点。本书从因果关系与相关关系之间的区别与联系出发，从因果关系模型、因果关系发现方法、因果关系与机器学习关系等角度对上述进展进行探讨，并对作者研究团队关于上述问题的一些最新研究进展进行总结。

本书可作为高等院校计算机、生物信息学、社会科学等相关专业的本科生或研究生教材，也可供对因果关系感兴趣的研究人员和工程技术人员参考。

图书在版编目（CIP）数据

大数据中的因果关系发现 / 蔡瑞初，郝志峰著. —北京：科学出版社，2018.8

ISBN 978-7-03-058476-2

Ⅰ. ①大… Ⅱ. ①蔡… ②郝… Ⅲ. ①数据处理-研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 180498 号

责任编辑：王 哲 王迎春 / 责任校对：郭瑞芝

责任印制：张 伟 / 封面设计：迷底书装

科学出版社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

北京教图印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 8 月第 一 版 开本：720×1000 1/16

2018 年 8 月第一次印刷 印张：7 3/4 插页：1

字数：156 000

定价：56.00元

(如有印装质量问题，我社负责调换)

前　言

本书是一部面向因果关系研究领域的入门书籍，主要对因果关系的一些基础知识、最新进展进行概要介绍，以帮助读者尽快了解因果关系研究的相关方法及进展。因果关系是计算机、数学和哲学等多领域交叉的理论前沿学科，为了降低阅读门槛，本书试图减少形式化描述和数学证明，侧重模型隐含假设及算法设计思想的介绍。因此，本书适合因果关系研究领域的入门学者以及对本领域感兴趣的相关研究者阅读。

本书共6章，第1～2章为基础知识部分，对因果关系的基本概念及后续研究涉及的数学基础进行概要性介绍；第3～5章为因果关系发现算法介绍部分，分别对基于约束、基于因果函数和混合型方法这三类基于观察数据的因果关系发现算法进行介绍；第6章对前面几章未能覆盖且与因果发现较为密切的相关主题进行概要性介绍。

本书的出发点是为相关领域研究者提供较为简明、系统的基础介绍。作者在数据挖掘与信息检索实验室开设因果关系讨论班期间发现，虽然因果关系在国际上发展迅速，但是国内针对这一细分领域的系统性介绍则较为缺乏，从而促成了本书的写作。基于这个出发点，本书内容的收集及撰写均有所侧重，可能存在严谨性不够、重要工作未能覆盖等问题。因果关系领域博大精深，由于作者水平有限，书中不足之处在所难免，若蒙各位读者告知，不胜感激。

最后，本书对于因果体系的梳理来源于作者与卡内基·梅隆大学张坤老师合作的因果关系发现算法综述，许多具体的内容来源于数据挖掘与信息检索实验室因果关系讨论班的讨论内容。团队研究生陈薇、曾艳、谢峰、乔杰等为本书的出版做了大量的贡献，在此一并表示真诚的感谢。

目　录

第1章 导 论

寻求事情背后的因果机制一直是人类认识世界的基本方式。以打雷现象为例，在科学不发达的时代，人们尝试用“雷神发怒导致打雷”这种因果机制进行解释；随着科技的发展，人们逐步认识到正负电子放电是打雷的原因。在一定程度上，现代科学的发展也是伴随着人们对一件件具体事情的因果关系的探索而发展的。

在对因果机制的探索过程中，人们已逐步积累了大量关于因果关系的朴素观念，如佛教里面的“因果报应”。这些朴素的因果机制大致可归纳为“因”事件的发生必然导致了“果”事件的发生。上述朴素观念包括两个层面的含义：原因事件和结果事件之间联系的必然性及事件发生的时序性。其中，联系的必然性是引发最多哲学争议的。最典型的争议是，休谟认为我们无法用已经观察到的经验证明必然联系。最典型的是，即使现在观察到的正负电子放电事件都导致打雷，也不能保证下一次正负电子放电事件必然会导致打雷。休谟这一论断往往被很多人用来论证休谟是否认因果关系的存在性的。

实际上，休谟并非否认了因果关系的存在性，而是说我们无法用经验证明必然联系，应该用经常性连接取代无法验证的必然联系。休谟这一贡献使得因果关系从一种朴素抽象的概念，成为一种可以用经验、历史数据验证的自然规律，一定程度奠定了基于数据因果关系发现的哲学基础。

在上述哲学思想的指导下，如何从大量的经验(实验或观察数据)中发现特定事物之间的因果关系发现算法进入了很多学者的研究视野。尤其是随着实验的开展、观察数据的积累，越来越多的学者开始尝试如何从积累的数据中自动发现这些事件之间的因果关系。尤其是 20 世纪 80 年代以来，图灵奖得主 Judea 教授，卡内基·梅隆大学的 Spirtes 教授、Glymour 教授、Scheines 教授等做了很多奠基性的工作，使得因果关系进入了“可计算”的范畴，吸引了一大批统计学家、计算机科学家的关注。

1.1 因果关系的概念

因果关系属于哲学和计算机交叉领域，给出严格的因果关系定义是非常困难的。在计算机领域，我们仅给出人们在大部分场合所认可的因果关系概念。

定义 1.1　因果关系(causality)：因果关系是“因”事件和“果”事件之间客观存在的关系，其中“因”事件是导致“果”事件发生的原因。

在上述抽象的因果关系定义中，一般蕴含了以下两方面的特性：①因果关系的客观性，因果关系作为客观现象之间引起与被引起的关系，“因”事件导致“果”事件的发生这一机制，并不以人的主观意志为转移；②因果关系的时间先后性，原因必定在先，结果只能在后，二者的时间顺序不能颠倒。

上述两方面特性为基于数据的因果关系研究提供了重要基础。其中，因果关系的客观性保证了因果关系是真实存在而且是可计算的，使得因果关系进入了可计算的范畴；因果关系的时间先后性则为“因”事件和“果”事件的判别提供了重要依据。

除上述性质以外，因果关系中的“因”事件和“果”事件往往是多对多的关系，即一个“因”事件可能导致多个“果”事件的发生，而一个“果”事件的发生往往也由多个“因”事件导致。以图 1.1 中的因果机制为例，吸烟导致黄牙、肺癌这两个事件；肺癌的发生则由遗传、环境、吸烟等多个原因引发。在现实世界中，这种多对多的因果关系导致事件之间呈现复杂的因果作用网络，是导致因果关系发现成为一个复杂任务的原因之一。

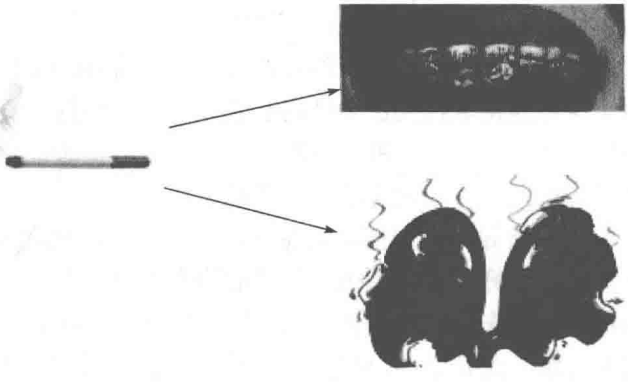

图 1.1　因果机制示意图

1.2　因果关系与相关关系

因果关系一定程度是在“因果关系-相关关系”的混淆、争议中发展的。仍然以上述例子(图 1.1)为例，吸烟、黄牙、肺癌具有较强的相关关系，然而只有吸烟才是肺癌的原因，也只有戒烟才能降低肺癌的发病概率，而把牙齿洗白则不能降低肺癌的发病概率。

定义 1.2　相关关系(association)：相关关系是指两个事件之间的关系，其中

一个事件总是与另一个事件共同发生。

具体来说，因果关系和相关关系的区别主要体现在以下几方面。

(1) 因果关系是有方向的，相关关系则是没有方向的。例如，从因果关系角度来看，“吸烟是黄牙的原因”这一论断成立，“黄牙是吸烟的原因”这一论断则不成立；从相关关系角度来看，“吸烟和黄牙具有较强的相关性”与“黄牙和吸烟具有较强的相关性”这两个论断是等价的。

(2) 因果关系往往导致“因”事件和“果”事件间呈现相关关系，而事件间呈现出的相关关系则不一定存在因果关系。仍以图 1.1 为例，吸烟是肺癌的原因，从而吸烟和肺癌存在较强的相关关系；图 1.1 中黄牙和肺癌也存在较强的相关关系，但是这种相关关系中并不蕴含着因果关系。因果关系导致的相关关系一定程度上也是因果关系发现的重要基础之一。

(3) 相关关系蕴含万物皆有联系，因果关系则是稀疏的。在图 1.1 的例子中，“吸烟”、“黄牙”、“肺癌”等因素存在较强的相关关系，而只有“吸烟” → “肺癌”等少数的因果关系。

(4) 因果关系可以比相关关系提供更加精确的干预建议。例如，对“吸烟”等直接原因的干预才能降低肺癌发生的概率。而对“黄牙”等相关事件的干预则不能降低肺癌的发生概率。

在因果关系和相关关系的区别与联系中，干预是一个非常重要的概念。在因果机制中，由于因果关系是有方向、客观存在的，所以可以通过“操纵”原因来获得预期的结果，使得人们对自然世界获得了较强的主观能动性，这也是因果关系与相关关系最为本质的区别。

定义 1.3 干预(intervention)：对数据中的某个变量 X 模拟物理干预，将它的原因对它的因果关系删除，用常数 $X = x$ 代替它们，同时保持模型的其余部分不变。

1.3 因果关系与机器学习

因果关系和现在基于相关的机器学习理论也有密切的联系，一定程度上因果关系正在使得机器学习理论更加扎实。因果关系中常用的产生式模型的学习方法如图 1.2 所示，其核心假设是数据背后存在一个稳定的联合分布式，如典型的样本独立同分布假设。而经典的机器学习任务则是基于观察到的数据推断某种模型。例如，依据朴素贝叶斯理论推断患者症状就是典型的分类问题。上述框架最大的问题在于若联合分布发生了变化，则上述框架无法迁移。例如，在男性样本上训练的肺癌预测模型无法适用于女性，因为样本的联合概率(如吸烟的概率)发生了变化。

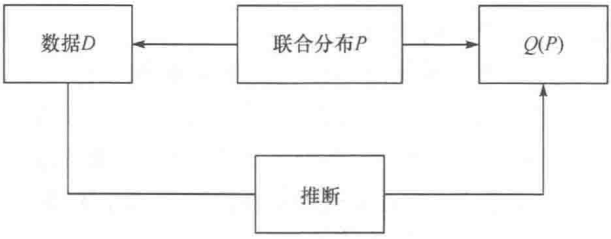

图 1.2　机器学习

而下述因果机制模型则较好地解决了这个问题，如图 1.3 所示。因为因果机制是稳定不变的，数据产生的联合概率是基于因果机制而产生的，所以因果关系模型可以非常容易地推广到不同的场景。仍以吸烟例子(图 1.1)为例，可以通过改变吸烟的概率分布而将模型进行迁移。在上述框架中，虽然 P(吸烟) 发生变化了，但是因果机制 P(肺癌|吸烟) 是不变的，因此模型可以较好地迁移到女性样本中。在上述场景中，其本质在于因果机制一定是比数据联合更加本质也是更加稳定的，是机制的内在本质的因素。所以最近卡内基·梅隆大学的张坤老师等也开始采用上述机制对很多迁移学习(多领域学习)问题进行研究[1]。该问题我们会在第 6 章进行详细介绍。

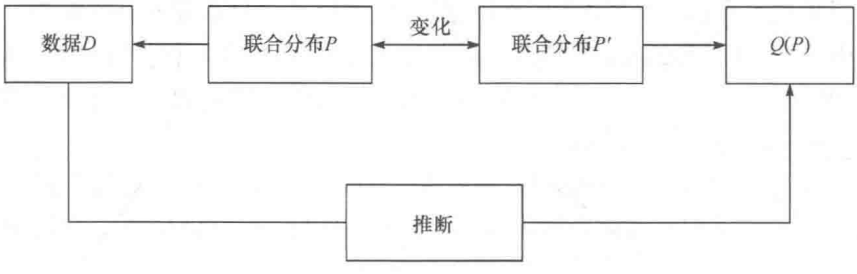

图 1.3　基于因果机制的机器学习迁移

1.4　基于实验与基于观察数据的因果关系发现

随机控制实验是发现因果关系的传统途径[2]。随机控制实验的基本方法是，将研究对象随机分组，对不同组实施不同的干预，在这种严格的条件下对照效果的不同。在研究对象数量足够的情况下，这种方法可以抵消已知和未知的混淆因素对各组的影响。此处的干预是指通过外部因素迫使某些变量的取值变化。例如，通过随机选择一个人群 A，让其吸烟；同时选择另外一个人群 B 让其不吸烟，通

过 A、B 两个人群的肺癌发生概率来研究吸烟与肺癌之间的因果关系。但是由于实验技术的局限性，绝大部分场合只能进行被动式观察，而无法进行主动式干预[3]。例如，对大规模的吸烟行为进行干预是不可能的，在临床等场合我们不能让患者使用未经验证的药物，也难以操控人的基因表达水平，即使是互联网广告等场合，随机实验仍然是代价巨大的。

从观察数据上进行有效的因果关系发现避免了以上限制，而且有可能给出从因到果的函数模型，因而具有重要的应用价值，是当前因果关系发现领域的研究热点[4, 5]。针对观察数据的不同特性，基于观察数据的因果关系发现方法可以分为基于时序观察数据的因果关系发现方法和基于非时序观察数据的因果关系发现方法(图 1.4)。虽然时序观察数据中时间维度蕴含了“因-果”方向的重要信息——“果”在时间上不能发生在“因”的前面，但是时序数据需要获取一个对象在不同时刻的观察值，对观察手段具有较高的要求。例如，现有的基因表达数据测量方法会破坏观察样本，使得我们无法获取该样本下一时刻的状态。更进一步，一般来说，基于时序数据来发现因果关系的结果对数据采集的频率等因素很敏感[6, 7]。所以基于非时序观察数据的因果关系发现具有更广的适用范围，也是当前因果关系发现领域的研究热点。

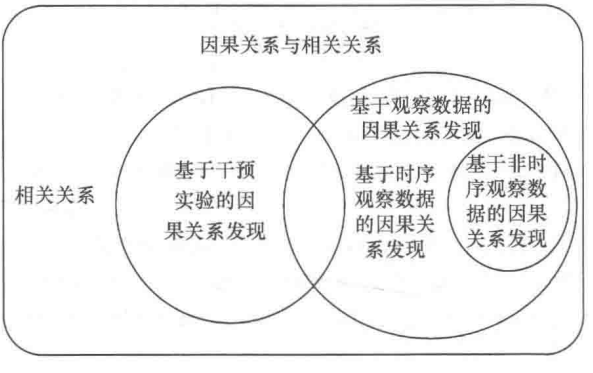

图 1.4 因果关系与相关关系

因此，本书主要对非时序观察数据上的因果关系发现方法进行介绍，并在第 6 章对时序数据和不完全数据的因果关系发现、因果关系与迁移学习进行简单的阐述。

当然，真正的因果关系还是要立足于理论模型的思考，如图 1.5 所示。如果推导技术没问题，那么这个“果”的合理性就直接依赖于给定的“因”假设了。这也是因果关系领域如此重视假设的原因。Simon 也指出，从数据中决定因果结

构的问题未被严格约束，感知到的因果结构依赖于我们为其设定的先验假设[8]。

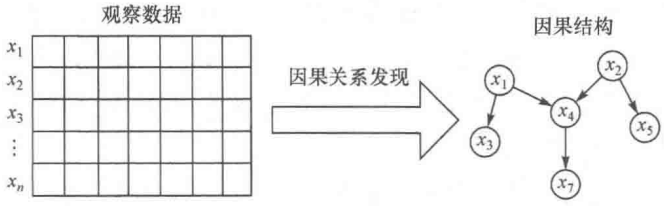

图 1.5　因果关系发现示意图

1.5 小　　结

因果关系严格区分了原因变量和结果变量，在揭示事物发生机制、指导干预行为等方面具有相关关系不能替代的重要作用。理解因果关系的核心在于正确区分因果关系和相关关系，并理解机器学习与因果关系之间的关联与区别。随着因果关系哲学基础的建立，基于观察数据的因果关系发现算法，已经成为数据科学中最有可能创造商业价值和进行科学发现的研究领域之一，正受到国际同行的广泛关注。

参 考 文 献

[1] Zhang K, Schölkopf B, Muandet K, et al. Domain adaptation under target and conditional shift// The 30th International Conference on Machine Learning, Atlanta, 2013.

[2] Pearl J. Causality: Models, Reasoning and Inference. Cambridge: Cambridge University Press, 2009.

[3] Cooper G, Yoo C. Causal discovery from a mixture of experimental and observational data// The 15th Conference on Uncertainty in Artificial Intelligence, Stockholm, 1999.

[4] Muchnik L, Aral S, Taylor S. Social influence bias: A randomized experiment. Science, 2013, 341(6146): 647-651.

[5] Aral S, Walker D. Creating social contagion through viral product design: A randomized trial of peer influence in networks. Management Science, 2011, 57(9): 1623-1639.

[6] Danks D, Plis S. Learning causal structure from undersampled time series// Proceedings of the NIPS Workshop on Causality, Nevada, 2013.

[7] Gong M, Zhang K, Schoelkopf B, et al. Discovering temporal causal relations from subsampled data// The 32nd International Conference on Machine Learning, Lille, 2015.

[8] Simon H. Spurious correlation: A causal interpretation. Journal of the American Statistical Association, 1954, 49(267): 467-479.

第 2 章　基础知识

因果关系发现是多领域的交叉学科。本章主要对本书中常用的贝叶斯网络、函数因果模型、独立性假设检验、回归分析等基础算法和模型作概要性介绍，以帮助读者在后续章节中对这些算法形成初步认识，具体算法的理论基础及细节等可以参考经典的著作。

2.1　贝叶斯网络

一般而言，贝叶斯网络的有向无环图(directed acyclic graphs, DAG)中的节点表示随机变量，它们可以是可观察到的变量或是隐藏变量等，连接两个节点的箭头代表这两个随机变量是非条件独立的；而两个节点间没有箭头相互连接在一起的情况就称为随机变量彼此间条件独立。节点之间的关系往往用条件概率分布来表示。一个典型的贝叶斯网络如图 2.1 所示。

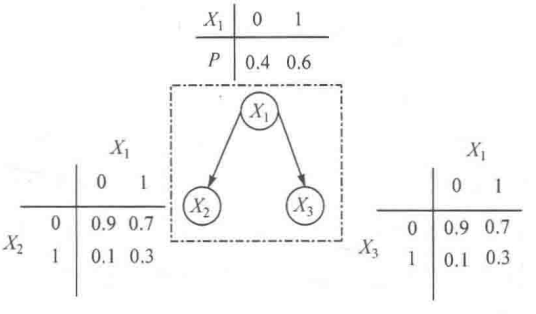

图 2.1　贝叶斯网络示意图(图中变量 X_1、X_2、X_3 都是二值变量 0 和 1)

定义 2.1　贝叶斯网络(Bayesian network)：贝叶斯网络是一种概率图型模型，其采用有向无环图和 n 组条件概率分布 $p(X_i \mid \mathrm{pa}_i)$ 共同表示一组变量 $X=\{X_1, X_2, \cdots, X_n\}$ 的联合概率密度分布，即 $p(X_1, X_2, \cdots, X_n)=\prod_{i=1}^{n} p(X_i \mid \mathrm{pa}_i)$。

通过上述定义，贝叶斯网络可以用一个有向图来表示一组变量之间的条件独立性关系，极大地降低了联合概率表示的复杂性。如图 2.1 所示，三个变量的联

合概率分布可以简化为 $p(X_1,X_2,X_3)=p(X_1)p(X_2\mid X_1)p(X_3\mid X_1)$。

定义 2.2　因果贝叶斯网络(causal Bayesian network)：因果贝叶斯网络[1]是在贝叶斯网络的基础上进一步规定两个节点之间的边表示两个节点之间存在因果关系，边的出度节点为原因节点(或原因变量)，边的箭头所指向的节点为结果节点(或结果变量)。

从贝叶斯网络到因果贝叶斯网络的扩展，其中蕴含了一个重要的前提，直接因果关系必然导致相关关系，这也是基于因果贝叶斯网络进行因果推断的重要前提。因为若两个变量间存在直接因果关系，却没有相关关系，则导致有向图中边的含义与贝叶斯网络中的条件独立相违背。一个简单的因果贝叶斯网络示意图如图 2.2 所示。

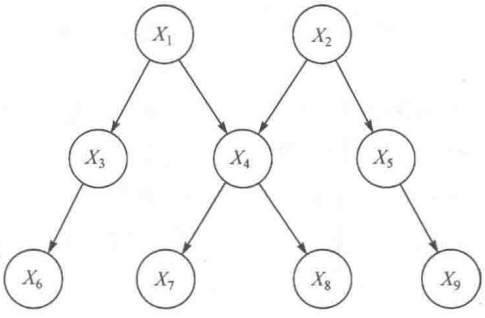

图 2.2　因果贝叶斯网络示意图

定义 2.3　父亲节点(parent node)：在因果关系网络图中，父亲节点是指一个节点的直接原因节点。每个节点的所有直接原因节点即为该节点的父亲节点集。

例如，图 2.2 中变量 X_8 的直接原因节点是 X_4，所以其父亲节点集为 $\{X_4\}$。

定义 2.4　祖先节点(ancestor node)：在因果关系网络图中，祖先节点是指一个节点的直接的或间接原因节点。每个节点的所有(直接的和间接的)原因节点即为该节点的祖先节点集。

例如，图 2.2 中变量 X_8 的直接原因节点为 X_4，间接原因节点为 X_1 和 X_2，所以其祖先节点集为 $\{X_1,X_2,X_4\}$。

定义 2.5　d-分离(d-separation)：d-分离准则可以在因果图中形式化描述变量间的独立性关系。d-分离准则的定义如下。

假设在无向图 G 中存在一个变量集 X，且 X_1 和 X_3 不在变量集中，α 表示 X_1 和 X_3 之间的一条通路，当路径 α 满足以下条件之一时，称变量集 X d-分离 X_1 和 X_3，即 α 阻断了 X_1 到 X_3 的所有通路：

(1) α 包含一种顺连 $X_1\rightarrow X_2\rightarrow X_3$ 或一种分连 $X_1\leftarrow X_2\rightarrow X_3$，且 $X_2\in X$；

(2) α 包含一种汇连 $X_1 \to X_2 \leftarrow X_3$，且 X_2 及其后代都不在 X 里。

例如，图 2.3 中变量 X_1 和 X_3 被 X_2 d-分离。

将一组变量的联合概率分布与有向无环图等价，依赖于一个非常重要的假设：忠诚性假设。其隐含意义是，变量之间不会出现额外的(条件)独立关系。在忠诚性假设下，模型不仅包含定义在变量或变量集上的结构方程，而且在实际情况下，真实的函数形式和系数的真实值没有额外隐含的约束。

定义 2.6　忠诚性假设(faithfulness assumption)：如果在给定变量集 X 的前提下[1]，变量 X_i 和 X_j 相互独立或条件独立，那么在由变量及其之间因果依赖关系组成的因果关系网络图 G 中，X_i 和 X_j 之间的所有路径被变量集 X 中合适的变量 d-分离(图 2.3)，则称所有随机变量的联合分布与图 G 是忠诚的。

定义 2.7　条件独立(conditional independence)：变量集 X_i 和变量集 X_j 在给定 X_k 下条件独立，即 $X_i \perp X_j \mid X_k$，当且仅当 $(X_i, X_j \mid X_k) = p(X_i \mid X_k)\ p(X_j \mid X_k)$。

上述条件独立的定义中，当 $X_k = \varnothing$ 时，条件独立实际上退化为边际独立，即 $(X_i, X_j) = p(X_i)p(X_j)$。

变量间的独立性不能直接表示变量间的因果机制。典型的如等价类问题。以图 2.4 为例，三个因果贝叶斯网络是等价的，因为它们之间蕴涵着同样的条件独立性关系。

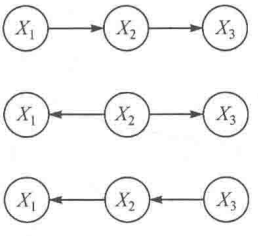

图 2.3　X_1 和 X_3 被 X_2 d-分离

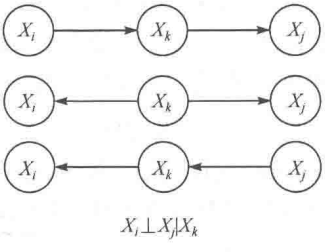

图 2.4　马尔可夫等价类示例图

在忠诚性假设下，d-分离和条件独立是等价的。这是贝叶斯网络结构学习、基于约束方法的因果结构学习的基本方法。

2.2　函数因果模型

由于在因果贝叶斯网络中，变量之间的关系只能通过条件概率描述，无法有效地刻画变量之间的复杂因果机制。而结构方程模型(structural equation model, SEM)[2]是建立和测试统计模型的统计技术，描述了变量的生成形式，通过结构

方程模型，我们能够从数据的生成过程描述变量间的因果关系。因此很多学者和专家尝试利用结构方程模型，从因果作用机制引发的数据分布特性等角度发现事物间的因果关系。以结构方程模型为基础，引入到因果图模型中，这类模型统称函数因果模型(functional causal model, FCM)[1]，也有学者称为结构因果模型(structural causal model, SCM)[3]，本书统一称为函数因果模型。

定义 2.8　函数因果模型：一个完整的函数因果模型表示的是图模型中 $\mathrm{pa}(x_i)\to x_i$，形式化为

$$x_i = f(\mathrm{pa}(x_i), n_i) \tag{2-1}$$

其中，$\mathrm{pa}(x_i)$ 表示变量 x_i 的父亲变量的集合，即此变量集通过某种函数直接作用变量 x_i；n_i 表示噪声变量，且彼此独立，父亲变量集合 $\mathrm{pa}(x_i)$ 和噪声变量 n_i 满足 $\mathrm{pa}(x_i)\perp n_i$，如图 2.5 所示。

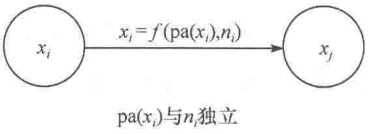

图 2.5　一种简单的函数因果模型

在函数因果模型中，集合 $\{\mathrm{pa}(x_i), i=1,2,\cdots,n\}$ 实际上已经包括了 DAG 中所有的结构信息，因此很多时候函数因果模型是找到此集合信息。那么如何来根据此模型发现因果关系呢？例如，对于两个变量 x_i 和 x_j，如果正确的因果方向为 $x_i\to x_j$，且联合概率 P_{x_i,x_j} 满足函数因果模型，那么反方向上是否也满足呢？事实上如果没有任何假设，我们总能够找到反方向上联合概率也满足函数因果模型(详细的证明见文献[3]4.1 节中定义 4.1)。

基于上述原因，我们常常认为通过观察数据是无法推断上述函数因果模型的，因此必须引入额外的合理假设。目前有代表性的模型包括线性非高斯无环模型(linear non-Gaussian acyclic model, LiNGAM)[4]、后非线性(post-nonlinear, PNL)[5]模型、非线性条件下的加性噪声模型(additive noise model, ANM)[6]，信息-几何因果推断(information-geometric causal inference, IGCI)[7, 8]模型等。其中的额外假设及其识别方式将会在第 4 章详细介绍。

2.3　独立性假设检验

条件独立的假设检验依赖于具体的数据类型及分布。常见的方法包括：离散数据的 G^2 检验方法、线性数据的偏相关检验方法和非线性数据的核条件独立性检验方法等。