УДК 004.855:519.216

О.С. БАЛАБАНОВ[*]

ВІДТВОРЕННЯ КАУЗАЛЬНИХ МЕРЕЖ НА ОСНОВІ АНАЛІЗУ МАРКОВСЬКИХ ВЛАСТИВОСТЕЙ

[*]Інститут програмних систем НАН України, Київ, Україна

Анотація. *Охарактеризовано новий підхід до виведення каузальних моделей з емпіричних даних, який спирається на виявлення фактів умовної незалежності. Підхід, базований на незалежності, забезпечує розробку асимптотично-коректних методів виведення каузальних мереж, у той час як регресійна методологія непридатна для цього. Базованим на незалежності методам притаманна дворівнева декомпозиція задачі, що сприяє зниженню розмірності потрібних статистик та обчислювальних витрат. Для підвищення ефективності метод доцільно озброїти набором резолюцій, які забезпечують усікання простору пошуку сепараторів та фокусування верифікації зв'язків. Пропоновані засоби ґрунтуються на необхідних вимогах до члена локально-мінімального d-сепаратора. Ефективність розроблених методів продемонстровано на прикладах. Викладено принципи контролю ефективності методів і адекватності моделі.*

Ключові слова: *каузальні мережі, умовна незалежність, локально-мінімальний d-сепаратор, каузальний вплив, верифікація та орієнтація ребра, ідентифікація структурних параметрів, оцінка адекватності моделі.*

Аннотация. *Охарактеризован новый подход к выводу каузальных моделей из эмпирических данных, который опирается на выявление фактов условной независимости. Подход, основанный на независимости, обеспечивает разработку асимптотически-корректных методов вывода каузальных сетей, в то время как регрессионная методология непригодна для этого. Основанным на независимости методам присуща двухуровневая декомпозиция задачи, что способствует снижению размерности необходимых статистик и вычислительных расходов. Для повышения эффективности в метод целесообразно ввести набор резолюций, которые обеспечивают усечение пространства поиска сепараторов и фокусировку верификации связей. Предложенные средства основаны на необходимых требованиях к члену локально-минимального d-сепаратора. Эффективность разработанных методов продемонстрирована на примерах. Изложены принципы контроля эффективности методов и адекватности модели.*

Ключевые слова: *каузальные сети, условная независимость, локально-минимальный d-сепаратор, каузальное влияние, верификация и ориентация ребра, идентификация структурных параметров, оценка адекватности модели.*

Abstract. *We characterize an independence-based approach to causal model inference from data. In contrast to regression, methods of this approach are aimed to asymptotically correctly recover a generative model. The merit of independence-based methods is inherent decomposition of model inference. This results in reducing dimensionality of statistics used as well as problem hardness. Aiming to enhance efficiency of methods we devise a few resolutions which facilitate contraction a space of search for separator and reducing a hardness of edge verification. The resolutions are grounded on necessary requirements on a member of locally-minimal d-separator. Efficiency of methods developed is demonstrated via few examples. Principles for verification of method effectiveness and model adequacy are presented.*

Keywords: *causal networks, conditional independence, locally-minimal d-separator, causal influence, edge verification and orientation, identification of structural parameters, evaluation of model adequacy.*

1. Вступ

Одна з центральних задач аналізу даних та моделювання – виведення каузальних моделей, придатних для аналізу рішень та планування дій у досліджуваній предметній галузі. Такому призначенню найкраще відповідають каузальні мережі – моделі ймовірнісних залежностей, структуровані згідно з марковськими властивостями [1, 2]. Впродовж останніх 20-ти років у провідних країнах інтенсивно розвиваються методи відтворення каузальних мереж

зі статистичних даних. Ми розглядаємо тут найважчу постановку задачі, коли структура моделі не відома апріорі (й активні експерименти недоступні). Невідомим може бути навіть темпоральний порядок змінних. У такій проблемній ситуації стають некоректними традиційні методи, наприклад, регресійні (коли виводиться неадекватна модель, яка «викривлює» картину каузальних зв'язків). У статті розглядається систематичний підхід до вирішення проблеми реконструкції моделі, який забезпечує асимптотично-коректний розв'язок. Модель відтворюється за фрагментами, через знайдення локальних марковських патернів й статистичних свідчень про зв'язки. (Темпоральний порядок змінних з'являється як логічний наслідок із структури моделі.) Невизначена проблемна ситуація робить точне рішення недосяжним. Виведена модель буде, по-перше, нечіткою («розмитою» у статистично-ймовірнісному сенсі) і, по-друге, ідентифікована як клас еквівалентності моделей (з невизначеними напрямками деяких зв'язків). Модель може містити безпосередні зв'язки (ребра) чотирьох типів. Ребро (дуга) вигляду $X \to Y$ відображає каузальний вплив X на Y. Ребро $U \leftrightarrow W$ позначає існування прихованої змінної (причини), що впливає рівночасно (паралельно) на U та W. Ребро $V \circ\!\!\to Z$ відображає два можливих варіанти: каузальний вплив або існування прихованої змінної (спільної причини). Ребро $Q \circ\!\!-\!\!\circ R$ означає, що каузальний характер цього зв'язку зовсім невизначений.

Процес виведення моделі стикається з обчислювальними проблемами, бо кількість можливих структур моделі є факторіально (експоненційно) великою. Кількість варіантів порядку змінних – також експоненційна. Для підвищення ефективності реконструкції моделі запропоновано систематичний підхід до пошуку сепараторів, базований на використанні імплікацій марковських властивостей каузальних мереж. Теоретичний ґрунт новацій – поняття локально-мінімального d-сепаратора (ЛоМС) та необхідні вимоги до членів ЛоМС. Запропоновані засоби дозволяють адаптивно оптимізувати і звужувати пошук складних мінімальних сепараторів, виходячи із знання вже знайдених «сусідніх» простих сепараторів та патернів залежностей. Відсікаються цілі сектори простору пошуку адекватної моделі.

2. Незадовільність застосування регресії

Монографії з аналізу даних традиційно застерігають від каузальної інтерпретації кореляції та фактів залежності. В [3] проаналізовано трактування каузальності регресійними методами в економетриці. Відомо, що регресійний аналіз – недосконалий метод реконструкції структур залежностей [2–4]. Дійсно, за невідомого темпорального порядку змінних незрозуміло, як обирати цільову змінну та предиктори. Результат регресії не дає каузальної інформації. Регресійний аналіз не гарантує навіть коректної ідентифікації сукупності безпосередніх статистичних зв'язків між змінними. Зрозуміло, можна перебрати всі варіанти впорядкування змінних і виконати багато регресійних задач, а потім серед багатьох виведених моделей вибрати мінімальну (найпростішу). Одначе кількість варіантів порядку змінних вже при двох десятках змінних стає астрономічно великою. Модель, виведена через регресію, «викривлюється» також внаслідок наявності прихованих змінних.

Для ілюстрації розглянемо простий приклад. Нехай система структуральних рівнянь (генеративна модель) має вигляд

$$X := a \cdot Z + b \cdot H + \varepsilon_X, \quad Y := c \cdot H + \varepsilon_Y$$

(структура моделі зображена на рис. 1). Припустимо, змінна H – прихована (не включена в дані). Нехай темпоральний порядок змінних – відомий, і змінна Y стоїть у порядку після (пізніше) X та Z. Тоді буде виконана регресія змінної Y на змінні X, Z. Стандартна процедура множинної регресії визнає X та Z значущими предикторами (врахування

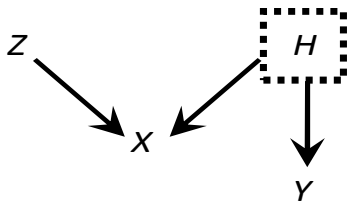

Рис. 1. Генеративна модель із прихованою змінною

змінної X «втягує» Z). А це створює ілюзію, що X та Z мають каузальний вплив на Y. Але насправді жодна з цих змінних зовсім не має каузального впливу на Y. Змінна Z навіть не асоційована з Y (між ними нульова кореляція). Зазначимо, що аналітики дійсно припустилися подібного неадекватного висновку в реальному дослідженні [5].

3. Нові методи відтворення каузальних моделей

Модель виводиться у формі каузальної мережі (точніше, її модифікації, що відображає невизначеність орієнтацій ребер) [2, 6, 7, 9]. Структура моделі визначається ациклонним орієнтованим графом (АОГ). АОГ-модель залежностей описується як $(G,\boldsymbol{\Theta})$, де G – АОГ, а $\boldsymbol{\Theta}$ – сукупність локально заданих параметрів у формі умовних розподілень імовірностей (або функції щільності) $p(X \mid \mathbf{B}(X))$, де $\mathbf{B}(X)$ – множина всіх батьків вершини (змінної) X. (Батько відповідає безпосередній «причині».)

Ефективний інструментарій розроблено для певних різновидів АОГ-моделей і саме з ними працюють найчастіше. До таких різновидів АОГ-моделей належать баєсові мережі (БМ) та гаусові мережі (ГМ). БМ визначені на категорних (дискретних) змінних; ГМ – моделі з неперервними змінними, нормальними дистурбаціями та лінійними залежностями. Для БМ параметри задані безпосередньо як компоненти $p(X \mid \mathbf{B}(X))$, які зазвичай подають як таблиці. (Тому баєсові мережі іноді називають напівпараметричними моделями.) Загальна форма опису параметрів виглядає як $p(X \mid \mathbf{B}(X), \vartheta)$, де $\vartheta \in \boldsymbol{\Theta}$ – підмножина параметрів, прив'язаних до $(X, \mathbf{B}(X))$.

Нехай $\mathbf{U}$ – множина всіх змінних моделі, $\mathbf{A}$ – множина всіх дуг орграфа G, а $|\mathbf{A}|$ – їх кількість. У теоретичній постановці задача формулюється так: задано розподілення ймовірностей $p^*(\mathbf{U})$; знайти таку $(G,\boldsymbol{\Theta})$, що $|\mathbf{A}| \mapsto \min$ за вимоги $p(\mathbf{U} \mid G,\boldsymbol{\Theta}) = p^*(\mathbf{U})$. Проте, оскільки на практиці задається вибірковий розподіл $\tilde{p}(\mathbf{U})$, який випадково відхиляється від генеративного (теоретичного) $p^*(\mathbf{U})$, вказана постановка задачі неприйнятна. (Вона веде до того, що модель «підганяється» до «гамору» у даних.) Реалістичніша постановка: задано розподіл $\tilde{p}(\mathbf{U})$ (практично, вибірковий); знайти модель $(G,\boldsymbol{\Theta})$ за максимумом обраного критерію. Таким критерієм може бути BIC; він об'єднує правдоподібність моделі й «штраф» за складність. Вказана задача – важка через величезну кількість структур моделі. До того ж розв'язання часто потребує великоформатних статистик.

Нами обрано інакший, базований на незалежності (або «constraint-based»), підхід до розв'язання задачі [2, 6, 7]. Структура АОГ-моделі характеризується марковськими властивостями, які накладають на розподіл $p(\mathbf{U} \mid G,\boldsymbol{\Theta})$ обмеження (типу рівність), інваріантні до параметризації моделі. Марковські властивості орієнтованих мереж залежностей формалізовано у графовому апараті за допомогою критерію d-сепарації [1, 2, 7, 8]. Предикат $\mathrm{Ds}(X;\mathbf{S};Y)$ означає, що вершини X та Y є d-сепаровані (d-незалежні), і множина $\mathbf{S}$ називається d-сепаратором для пари (X,Y). Умовну незалежність змінних X та Y позначимо як $\mathrm{Ind}(X,Y \mid \mathbf{S})$.

Каузальна марковська умова (постулат) встановлює, що в АОГ-моделі з кожної d-сепарації випливає відповідна умовна незалежність:

$$\forall X,Y,\mathbf{S}\ (X,Y \notin \mathbf{S}): [\mathrm{Ds}(X;\mathbf{S};Y) \Rightarrow \mathrm{Ind}(X,Y \mid \mathbf{S})].$$

Для обґрунтування методів, базованих на незалежності, потрібна обернена імплікація. Вона сформульована як припущення, яке в основному виконується в моделях (за виключенням особливих випадків) і в асимптотично-великих вибірках даних [2, 7].

Припущення Каузальної неоманливості. В кожному (точному) розподіленні ймовірностей змінних, генерованому з АОГ-моделі, для всіх змінних чинна імплікація вигляду

$$\forall X, Y : [\exists \mathbf{S}\ (X, Y \notin \mathbf{S}) : \mathrm{Ind}(X, Y \mid \mathbf{S})] \Rightarrow \mathrm{Ds}(X; \mathbf{S}; Y).$$

Завдяки цій властивості можна виводити модель на основі виявлених умовних незалежностей. Набір $\mathbf{S}$, що забезпечує незалежність, називають сепаратором. Процес реконструкції моделі базується на пошуку сепараторів; такі методи звуться сепараційними.

Постановка задачі трансформується у завдання, розгорнуте за етапами:

1. Відтворити сукупність ребер, тобто верифікувати ребро для кожної пари змінних, відшукуючи сепаратори і трактуючи умовну незалежність як факт d-сепарації (відсутність ребра).

2. Ідентифікувати напрямки ребер (орієнтувати), спираючись на аналіз сусідніх зв'язків [1, 2, 7, 9].

3. Обчислити параметри моделі $p(X \mid \mathbf{B}(X))$, виходячи з $\tilde{p}(\mathbf{U})$ та знайдених $\mathbf{B}(X)$.

Таким чином, здійснено концептуальну декомпозицію задачі, причому головна декомпозиція захована всередині першого (найбільш складного) етапу. Рішення щодо існування кожного зв'язку моделі можна приймати автономно. Це означає, що замість перебору цілих моделей (чи навіть фрагментів-«родин») виконується перебір сепараторів для пар змінних. Тим самим досягається зниження розмірності потрібних статистик. Ребра отримують статус каузальних тільки внаслідок виявлення патерна «*Y*-конфігурації» [1, 2, 7, 9]. Не всі ребра вичерпно орієнтуються, тому результат етапу «2» репрезентовано у формі «повного частково орієнтованого ациклічного графа» (CPDAG).

Найпершими серед базованих на незалежності (сепараційних) алгоритмів виведення моделі були IC та PC [1, 2]. З метою прискорення в алгоритм PC закладено кілька принципів, серед яких є такий: сепаратор для пари (X, Y) шукається як підмножина вершин (змінних), потенційно-суміжних до X або Y. В ході виконання етапу «1» ребра видаляються, отже, дерево перебору звужується. Але в умовах каузальної недостатності (тобто за наявності прихованих спільних причин двох змінних) алгоритм PC може втратити сепаратор. Тому для умов каузальної недостатності розроблено інший алгоритм – FCI [2, 6], який включає додаткові спеціальні етапи пошуку сепараторів та орієнтації ребер.

Огляд методів виведення каузальних моделей можна знайти в [2, 10–12]. Останніми роками серед методів, базованих на незалежності, виокремилося ціле відгалуження, характерне тим, що замість сепараторів (для пар змінних) виявляються «марковські бланкети». Мабуть, першим таким алгоритмом був GS-алгоритм [13]. Довкола кожної змінної формується «марковський бланкет»; верифікація системи зв'язків здійснюється у два підетапи – «розростання» та «усікання» (мінімізація) бланкетів. До алгоритмів цього відгалуження належать, зокрема, GLL [14], MBFS [15], TC [16]. Використання «марковських бланкетів», з одного боку, сприяє зменшенню кількості тестів, але з іншого – призводить до ускладнення формату тестів незалежності. Внаслідок цього зростає важкість обчислення статистик, необхідних для виконання тестів, що відображається на обчислювальній важкості алгоритму виведення. Крім того, ускладнення статистик тягне загострення проблеми ненадійності (погіршення адекватності моделі).

Та особливість підходу, що рішення щодо різних ребер можна приймати автономно, не означає, що рішення треба шукати ізольовано. Останнє призвело б до багатократного дублювання роботи. Алгоритм PC [2, 17] економить кількість тестів, зменшуючи набори

потенційно-суміжних вершин (змінних) для всієї моделі, обходячи пари змінних по спіралі. Припустимо, ребро $(X - Y)$ існує. Для того, щоб алгоритм переконався у цьому факті (в ході розв'язання задачі на етапі «1»), він має виконати $\sum_{i=0}^{n-2}\binom{n-2}{i}$ перевірок незалежності для (X, Y), де n у гіршому випадку дорівнює $|\mathbf{U}|$, а у кращому випадку дорівнює кількості вершин, суміжних до X або Y. Зі зростанням насиченості моделі зв'язками переваги підходу у швидкості поступово втрачаються. Тож треба знаходити подальші резерви оптимізації пошуку. Резерви знайдено у підвищенні ефективності пошуку сепараторів за рахунок використання «інформаційного обміну» між різними гілками пошуку сепараторів. Ключем до рішення стало залучення необхідних вимог до локально-мінімальних сепараторів та систематичне використання «глибоких» властивостей сукупності фактів d-сепарації в каузальних мережах [7, 18, 19].

Визначення. Локально-мінімальним сепаратором для пари (X, Y) називається такий сепаратор $\mathbf{S}$, що в результаті видалення будь-якого його елемента $Z \in \mathbf{S}$ множина $\mathbf{S}/\{Z\}$ не буде сепаратором для (X, Y).

Достатньо зосередити пошук на локально-мінімальних сепараторах. Необхідні вимоги до члена локально-мінімального сепаратора імплікують резолюції (правила) відсіювання змінних зі списку кандидатів до складу сепараторів. Ці резолюції дають засоби відсікати цілі сектори простору пробних (потенційних) сепараторів, фокусувати верифікацію ребер і скорочувати тривалість реконструкції моделі.

4. Адаптивне звуження простору пошуку

Формально було доведено низку резолюцій (правил) для локально-мінімальних d-сепараторів та їх членів [7, 8, 18, 19]. Розроблено цілий комплект правил оптимізації пошуку сепараторів. Вони були втілені в алгоритмах виведення моделі та випробувані [7, 20, 21]. За результатами експериментів ефективними й практично найбільш важливими показали себе кілька правил, поданих нижче. Найкориснішим є правило «відсторонення».

Правило «відсторонення» кандидатів у сепаратор (‘placing aside’): якщо в орграфі G вершина X d-сепарує Z та Y, то вершина Z не є членом жодного локально-мінімального сепаратора для пари (X, Y).

Правило обов'язковості потенційного стрижня. До складу кожного не порожнього локально-мінімального d-сепаратора для пари вершин (X, Y) входить щонайменше одна вершина Z, така, що чинні факти $\neg\mathrm{Ds}(Z;\{\};X)$, $\neg\mathrm{Ds}(Z;\{\};Y)$, $\neg\mathrm{Ds}(Z;\{Y\};X)$, $\neg\mathrm{Ds}(Z;\{X\};Y)$.

Також корисним показало себе правило «замкнених стрижнів», яке формулюється більш громіздко. Це правило можна вивести з кількох положень, таких, як пропозиція 4 (про ізольовані спільні близькі) з [19], принцип композиції ненадлишкового сепаратора з [8] або твердження 2.2 з [7]. Сенс правила «замкнених стрижнів»: для заданої пари (X, Y) маємо список кандидатів у стрижні сепаратора та список інших (нестрижневих, «залучених») кандидатів у члени сепаратора для (X, Y). Якщо серед членів другого списку немає жодного, який безумовно залежить від якогось члена першого списку, то треба видалити увесь другий список (нестрижневих кандидатів).

Якщо модель (за своєю структурою) належить до підкласу, названого лісами [17, 22], то достатньо озброїти стандартний алгоритм лише двома правилами – «відсторонення» кандидатів у сепаратор та правилом стрижня – для того, щоб алгоритм вже після тестів першого рангу розпізнавав вичерпання можливостей реконструкції моделі [7, 17].

Зрозуміло, що для виведення моделі з даних застосовуються емпіричні (статистичні) «зліпки» (counterparts) вказаних правил. Замість фактів d-сепарації використовуються результати статистичних тестів. Алгоритми виведення моделі, озброєні такими засобами, залишаються асимптотично-коректними, і водночас помітно переважають відомі алгоритми за швидкістю [7, 17, 20, 21]. Корисність розроблених резолюцій пояснюється тим, що вони спираються на прості відношення і сепаратори, а застосовуються для пошуку складних сепараторів та для з'ясування, що шуканого сепаратора не існує.

Сенс звуження пошуку сепараторів у ході реконструкції моделі полягає у заміні необов'язкових тестів логічним аналізом (зіставленням) результатів виконаних тестів з дотичними змінними. Найбільший виграш очікується в моделях, де тестування потребує важких обчислень і де кожний тест вимагає нового сканування вибірки даних. Таку ситуацію маємо в нелінійних моделях, баєсових мережах та моделях зі змінними різних типів. Натомість виведення ГМ – обчислювально найбільш просте, бо спочатку обчислюється матриця коваріацій, з якої потім можна отримати статистики для будь-якого тесту.

5. Ілюстративні приклади

З метою наочності розглянемо просту ГМ (хоча для лінійних моделей ефект наших новацій – найменший). Нехай ГМ описується такою системою структуральних рівнянь.

$$X_1 := 0{,}75 \cdot X_5 + \varepsilon_1, \quad X_2 := 0{,}65 \cdot X_5 + \varepsilon_2,$$

$$X_3 := 0{,}8 \cdot X_6 + \varepsilon_3, \quad X_4 := 0{,}4 \cdot X_3 + 0{,}55 \cdot X_6 + \varepsilon_4,$$

$$X_5 := 0{,}8 \cdot X_7 + \varepsilon_5, \quad X_6 := 0{,}9 \cdot X_7 + \varepsilon_6,$$

$$X_7 := 0{,}5 \cdot X_8 + 0{,}6 \cdot X_9 + 0{,}45 \cdot X_{10} + \varepsilon_7,$$

$$X_8 := \varepsilon_8, \quad X_9 := 0{,}7 \cdot X_{10} + \varepsilon_9, \quad X_{10} := \varepsilon_{10},$$

де $\varepsilon_i \sim N(0,1)$, $\varepsilon_i \perp \varepsilon_j$ $(\varepsilon_i \neq \varepsilon_j)$.

У формулах замість рівності свідомо використано знак присвоєння; тим самим підкреслено, що рівняння є структуральними. Коефіцієнти виражають каузальний вплив. Якщо перенести члени вліво/вправо через знак «:=», то рівняння перестане бути структуральним.

Структура моделі показана на рис. 2 *а*. З цієї моделі було генеровано вибірку даних обсягом 1000 записів. Дані було оброблено і отримано матрицю парних коваріацій (див. таблицю). Цю матрицю коваріацій було подано на вхід алгоритму Razor-1.3. Ніяких апріорних знань про модель алгоритм не отримав (втім, зрозуміло, що подача на вхід лише матриці коваріацій автоматично означає прийняття гіпотези, що модель належить до класу ГМ).

Таблиця 1. Матриця коваріацій, обчислена з даних

2,46									
1,33	2,07								
0,998	0,847	2,75							
1,05	0,916	2,31	3,15						
1,99	1,72	1,40	1,52	2,72					
1,25	1,07	2,19	2,38	1,78	2,73				
1,47	1,25	1,58	1,73	2,02	2,04	2,36			
0,340	0,252	0,382	0,423	0,442	0,507	0,577	1,03		
0,782	0,674	0,809	0,854	1,08	1,01	1,22	0,054	1,49	
0,508	0,448	0,543	0,568	0,713	0,681	0,855	0,060	0,671	0,951

У результаті було виведено структуру моделі, показану на рис. 2*б*. Коректно відтворено клас еквівалентності генеративної моделі. Залишилися невизначеними (невідомими) спрямування деяких зв'язків. Орієнтацію ребер $X_9 \circ\!\!-\!\!\circ X_{10}$ та $X_3 \circ\!\!-\!\!\circ X_4$ неможливо ідентифікувати на базі фактів умовної незалежності (в такому оточенні). Частково орієнтовані ребра (субкаузальні зв'язки): $X_8 \circ\!\!\rightarrow X_7$, $X_9 \circ\!\!\rightarrow X_7$ та $X_{10} \circ\!\!\rightarrow X_7$.

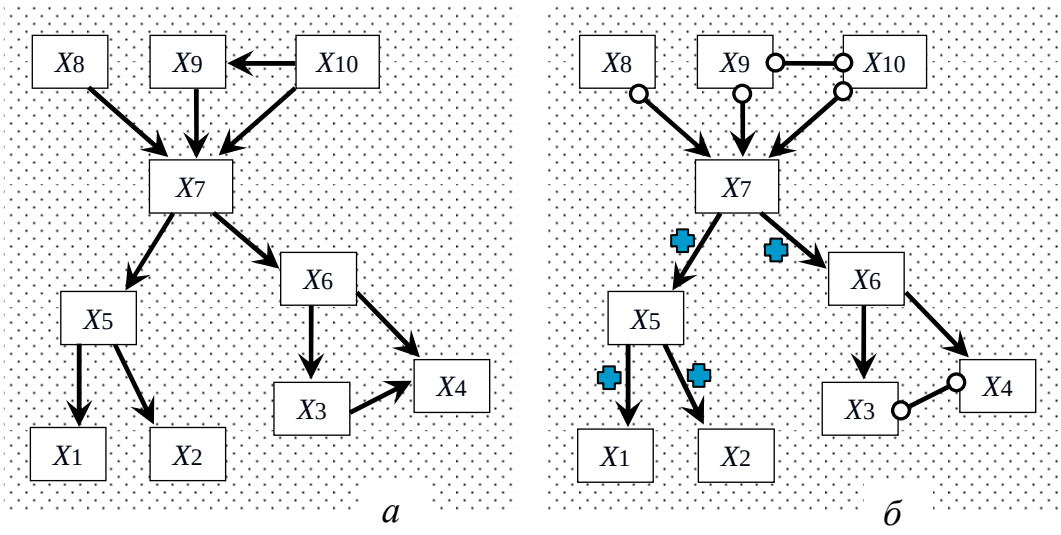

Рис. 2: *а* – генеративна модель; *б* – виведена модель

Для зв'язків з невизначеним напрямком неясно, як ідентифікувати структурні параметри (коефіцієнти), хіба що зробити відповідне припущення. Але все ж таки модель дає змогу однозначно отримати «майже структурні» параметри. Тобто можна оцінити «пряму» (безпосередню) кореляцію суміжних змінних. Наприклад, для ребра $X_3 \circ\!\!-\!\!\circ X_4$ «пряма» кореляція обчислюється як частинна кореляція з умовою на X_6, а для ребра $X_9 \circ\!\!-\!\!\circ X_{10}$ – як безумовна кореляція. З іншого боку, навіть для повністю визначеного каузального ребра не завжди можливо однозначно оцінити структурний параметр. На рис. 2*б* знаком «+» позначені зв'язки, які мають однозначні оцінки структурних параметрів. Для каузальних зв'язків $X_6 \rightarrow X_3$ та $X_6 \rightarrow X_4$ структурні параметри неможливо однозначно оцінити через невизначеність орієнтації ребра $X_3 \circ\!\!-\!\!\circ X_4$. (Є два можливих варіанти.) Натомість для субкаузального зв'язку $X_8 \circ\!\!\rightarrow X_7$ існує «майже структурний» параметр, який оцінюється як безумовна кореляція.

Застосування правил звуження простору пошуку сепараторів для цієї моделі не дає прискорення порівняно з алгоритмом PC (тут прискорення й не потрібне; тривалість виведення – менша за 1 секунду). Але виграш полягає в іншому: виведення цієї моделі алгоритмом Razor-1.3 закінчилося на циклі тестів першого рангу; натомість PC дійшов до тестів четвертого рангу. (Хоча в лінійних моделях статистики для всіх тестів обчислюються з парних кореляцій, підвищення рангу тесту тягне зростання похибки і ризику помилок.)

Нехай модель виводиться в умовах каузальної недостатності. Тоді, якщо додатково задати темпоральний порядок змінних, то всі неорієнтовані ребра перейдуть у статус субкаузальних. Але це не додасть жодного каузального ребра і не збільшить кількість ідентифікованих структурних параметрів.

Для контрастності опишемо також приклад виведення БМ, який був найважчим серед виконаних експериментів. Модель (іменована “BB55”) має 30 тризначних змінних та

120 ребер. Позиції ребер і значення параметрів були обрані випадковим механізмом [23]. Для виведення такої складної моделі потрібна велика вибірка даних. У першому експерименті з цією моделлю використана вибірка даних 20000 записів. Робота алгоритму Razor-1.2 тривала 76 хвилин. Для алгоритму PC тривалість досягла 137 хвилин. У ході виведення алгоритм Razor-1.2 виконав 16950 тестів, а алгоритм PC – 27650 тестів. Зокрема, Razor-1.2 виконав 5 тестів дев'ятого рангу, а PC – 110 таких тестів. Суттєвого прискорення Razor-1.2 досяг за рахунок правил звуження простору пошуку сепараторів. У ході виведення моделі правило «відсторонення» кандидатів у сепаратор продуктивно спрацювало 132 рази. Правило «замкнених стрижнів» продуктивно спрацювало 43 рази.

Алгоритм Razor-1.2 пропустив 45 автентичних ребер моделі, а алгоритм PC – 54 ребра. Велика кількість таких помилок зумовлена тим, що використаний механізм генерації параметрів породжує «хаотичний» характер залежностей. Найбільш небезпечним типом помилки у виведенні моделі є реверсування ребра, тобто виведення ребра, спрямованого у протилежному напрямку порівняно з генеративною моделлю. Помилка типу пропущене ребро може бути компенсована іншими зв'язками. Натомість реверсування ребра призводить до принципової неадекватності у застосуванні виведеної моделі та хибних висновків. У ході виведення моделі “BB55” алгоритм PC зробив реверсування одного ребра. Алгоритм Razor-1.2 не зробив жодної такої помилки в експериментах з двома десятками моделей (включно з цією).

В другому експерименті з цією моделлю використана вибірка даних 50000 записів. Тривалість виведення моделі з такої вибірки досягла 746 хвилин для алгоритму Razor-1.2 та 912 хвилин для алгоритму PC. Це більше 15 годин. Кількість пропущених автентичних ребер моделі склала 29 ребер для алгоритму Razor-1.2 та 40 ребер для алгоритму PC. Краща точність алгоритму Razor-1.2 пояснюється тим, що звуження простору пошуку сепараторів відсікає цілі ареали простору з підвищеним ризиком помилок тестів.

Наведемо також приклад виведення моделі «помірної» складності. Генеративна модель (“BB31”) являє собою БМ, яка має 30 тризначних змінних та 90 ребер. Сукупність ребер цієї моделі відображена на рис. 3. Виведення моделі “BB31” з вибірки даних (20000 записів) дало такі результати. Алгоритм Razor-1.2 витратив 12 хвилин, алгоритм PC – 19 хвилин. Кількість виконаних тестів була 4771 та 7188 відповідно. В ході виведення моделі алгоритм Razor-1.2 продуктивно застосував правило «відсторонення» 106 разів, а правило «замкнених стрижнів» – 81 раз. Алгоритм Razor-1.2 пропустив 16 автентичних ребер моделі, а алгоритм PC – 29 ребер. Якщо брати до уваги тільки «каузальні» ребра генеративної моделі (тобто ті ребра, що теоретично мусять бути ідентифіковані повністю), то алгоритм Razor-1.2 вірно ідентифікував 6 таких ребер, а алгоритм PC – 3 ребра. Реверсування каузальних ребер не сталося.

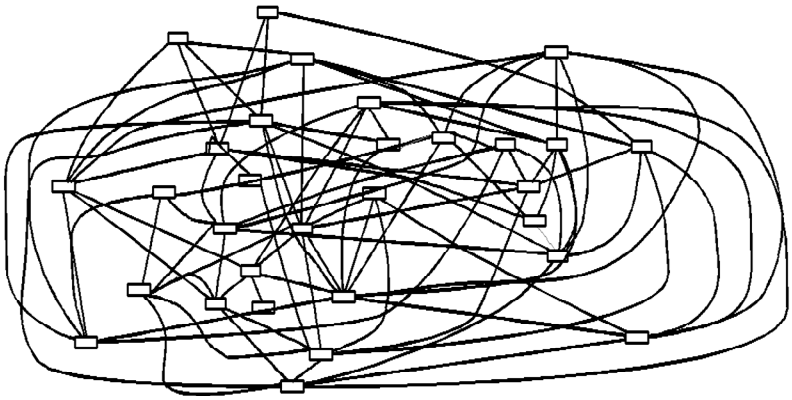

Рис. 3. Структура моделі “BB31” (“кістяк”)

Велика кількість ребер з (частково) невизначеними орієнтаціями робить неможливим оцінку багатьох параметрів баєсової мережі. Нагадаємо, що на відміну від ГМ, у БМ ребра не мають своїх «окремих» параметрів (коли є кілька батьків).

6. Верифікація методів та алгоритмів реконструкції моделі

Розглянуті методи виведення моделі з даних в першу чергу призначені для проблемних ситуацій, коли модель у дійсності невідома. Але для того, щоб моделі, виведені в таких ситуаціях, заслуговували на довіру, метод виведення має переконливо демонструвати коректність та ефективність. Здатність методу відтворювати адекватні моделі можна було б підтвердити, застосувавши виведені моделі на практиці й дочекавшись наслідків. Але такого підтвердження не завжди можна дочекатися (чи навіть приступити до впровадження).

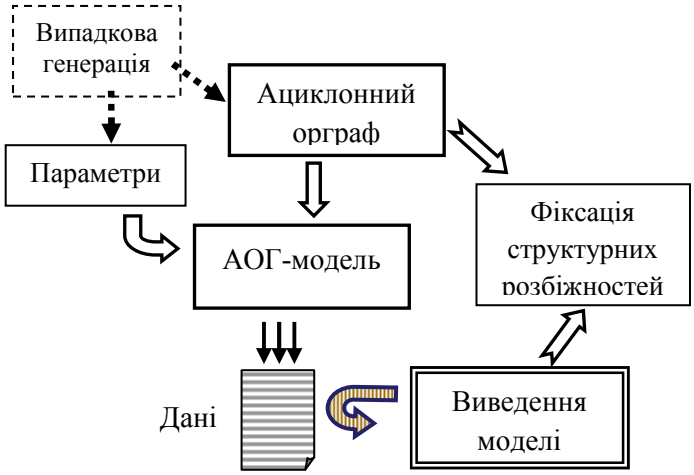

Рис. 4. Схема експериментів

Реалістичний шлях випробування розроблених алгоритмів пролягає через експерименти з можливістю порівняти виведену модель з генеративною (справжньою). Генеративна модель має бути відома досліднику (аналітику), але невідома методу. Для «жорсткого» випробування методу модель треба виводити виключно із статистичних даних (без будь-яких апріорних знань про структуру). Схема експериментів показана на рис. 4. Для «чистоти» експериментів потрібно виконати своєрідну «квазі-рандомізацію», тобто і структуру, і параметри моделі генерувати випадково [7, 23].

Оцінити адекватність моделі можна через порівняння прогнозів наслідків управління (втручання), які дають виведена та автентична моделі. Але в ситуації, коли конкретні вимоги до моделі не задано, компактний спосіб оцінки полягає у виявленні та підрахунку структурних відхилень (помилок). Зустрічаються структурні помилки різних типів; особливу увагу треба надавати каузальним зв'язкам. Наочну оцінку адекватності виведеної моделі можна отримати за допомогою індексу каузальної продуктивності [7], який вимірює коректність та повноту відтворення каузальних зв'язків. Мабуть, ґрунтовнішим критерієм адекватності є структурно-інтервенційна дистанція [24] (хоча той критерій стикається з труднощами через невизначеність орієнтацій ребер).

7. Висновки

В роботі запропоновано підсилення методу відтворення каузальних мереж з даних засобами усікання простору пошуку, які ґрунтуються на закономірностях марковських властивостей. Згідно з результатами, таке підсилення не тільки значно прискорює відтворення моделі, але в більшості випадків також зменшує кількість помилок. Останнє пояснюється тим, що відсікаються сектори простору з підвищеним ризиком помилок.

Виведення каузальних моделей з емпіричних даних призначене для пізнавальних задач і підтримки управління об'єктами та процесами у погано досліджених галузях і середовищах. Така ситуація типова для відкритих середовищ, де взаємодіють багато факторів, які вивчаються різними дисциплінами (все «заплутане»), так що неможливо побудувати модель на теоретичних підставах. У багатьох ситуаціях також неприйнятно (ризиковано) цілком покладатися на експертів. Водночас можуть бути неприпустимими (рандомізовані) експерименти на об'єкті моделювання (з огляду на етичні чи економічні міркування або через тимчасову недосяжність об'єкта).

Для реконструкції адекватної моделі (особливо у випадках складних та нелінійних

форм залежностей) необхідно мати великі вибірки даних. Кожний елемент даних має бути виміряний точно і відображати «миттєвий» стан процесу. Втім збільшення обсягу даних не розв'язує проблему принципової неідентифікабельності еквівалентних моделей (невизначеність напрямків орієнтацій зв'язків у «заплутаних» моделях). У структурах, перенасичених зв'язками, неможливо навіть розпочати процес орієнтацій ребер. Для виходу з такого тупика потрібно «перезавантажити» (оновити) завдання, включивши в номенклатуру даних додаткові змінні (сподіваючись, що деякі з них зіграють роль інструментальних). Також корисно збільшити частоту вимірювання тих самих характеристик. Звичайно, якщо є достовірні апріорні знання, їх треба використати. Це дозволить прискорити виведення та уточнити модель. Нагадуємо, що невизначеність у виведеній моделі об'єктивно зумовлена і застерігає аналітика від необґрунтованих висновків.

У наш час дослідження в цій галузі ведуться в кількох напрямах. Паралельно з методами реконструкції моделі з даних дослідження охоплюють: теоретичне узагальнення класу каузальних моделей; техніку тестування умовної незалежності у складних (загальних) випадках; методи застосування каузальних моделей, в першу чергу – прогнозування каузального ефекту управління.

СПИСОК ЛІТЕРАТУРИ

1. Pearl J. Causality: models, reasoning, and inference / Pearl J. – Cambridge: Cambridge Univ. Press, 2000. – 526 p.
2. Spirtes P. Causation, prediction and search / Spirtes P., Glymour C., Scheines R. – New York: MIT Press, 2001. – 543 p.
3. Chen B. Regression and causation: a critical examination of six econometrics textbooks / B. Chen, J. Pearl // Real-World Economics Review. – 2013. – Issue 65. – P. 2 – 20.
4. Bollen K.A. Eight myths about causality and structural equation models / K.A. Bollen, J. Pearl // Methods in Social Epidemiology / J.M. Oakes, J.S. Kaufman (eds.). – John Wiley & Sons, Jossey-Bass, 2006. – P. 301 – 329.
5. Fienberg S. E. Expert statistical testimony and epidemiological evidence: the toxic effects of lead exposure on children / S.E. Fienberg, C. Glymour, R. Scheines // Journal of Econometrics. – 2003. – Vol. 113. – P. 33 – 48.
6. Learning high-dimensional directed acyclic graphs with latent and selection variables / D. Colombo, M.H. Maathuis, M. Kalisch [et al.] // Annals of Statistics. – 2012. – Vol. 40, N 1. – P. 294 – 321.
7. Балабанов О.С. Каузальні мережі: аналіз, синтез та виведення з статистичних даних: дис. … доктора фіз.-мат. наук: спец. 01.05.01 / Балабанов Олександр Степанович. – К.: Інститут кібернетики імені В.М. Глушкова НАНУ, 2014. – 305 с.
8. Балабанов А.С. Логика минимальной сепарации в каузальных сетях / А.С. Балабанов // Кибернетика и системный анализ. – 2013. – № 2. – С. 36 – 47.
9. Балабанов О.С. Від коваріацій до каузальності. Відкриття структур залежностей в даних // Системні дослідження та інформаційні технології. – 2011. – № 4. – С. 104 – 118.
10. Kalisch M. Causal structure learning and inference: a selective review / M. Kalisch, P. Bühlmann // Quality Technology & Quantitative Management. – 2014. – Vol. 11, N 1. – P. 3 – 21.
11. Fu S. Markov blanket based feature selection: a review of past decade / S. Fu, M.C. Desmarais // Proc. of the World Congress on Engineering (WCE-2010). – London, UK: Intern. Association of Engineers: Newswood Limited, 2010. – Vol. 1, June 30 – July 2. – P. 321 – 328.
12. Koski T. J. T. A review of Bayesian networks and structure learning / T.J.T. Koski, J.M. Noble // Annales Societatis Mathematicae Polonae. Series 3: Mathematica Applicanda. – 2012. – Vol. 40, N 1. – P. 53 – 103.
13. Margaritis D. Bayesian network induction via local neighbourhoods / D. Margaritis, S. Thrun // Advances in Neural Information Processing Systems. – 1999. – Vol. 12. – P. 505 – 511.
14. Local causal and Markov blanket induction for causal discovery and feature selection for classification / C.F. Aliferis, A. Statnikov, I. Tsamardinos [et al.] // J. Machine Learn. Res. – 2010. – Vol. 11. – P. 171 – 234.
15. Ramsey J.D. A PC-style Markov blanket search for high dimensional dataset / J.D. Ramsey // Techni-

cal Report. – 2006. – N 177. – Department of Philosophy, Carnegie Mellon University. – Pittsburgh, PA, 2006. – 13 p.
16. Pellet J.P. Using Markov blankets for causal structure learning / J.P. Pellet, A. Elisseeff // J. Machine Learn. Res. – 2008. – Vol. 9. – P. 1295 – 1342.
17. Балабанов О.С. Прискорення алгоритмів відтворення баєсових мереж. Адаптація до структур без циклів / О.С. Балабанов // Проблеми програмування. – 2011. – № 1. – С. 63 – 69.
18. Балабанов А.С. Минимальные сепараторы в структурах зависимостей. Свойства и идентификация // Кибернетика и системный анализ. – 2008. – № 6. – С. 17 – 32.
19. Балабанов А.С. Формирование минимальных d-сепараторов в системе зависимостей / А.С. Балабанов // Кибернетика и системный анализ. – 2009. – № 5. – С. 38 – 50.
20. Быстрый алгоритм вывода структур байесовых сетей из данных / А.С. Балабанов, А.С. Гапеев, А.М. Гупал [и др.] // Проблемы управления и информатики. – 2011. – № 5. – С. 73 – 80.
21. Balabanov O.S. On perspectives of causal networks reconstruction by independence-based methods / O.S. Balabanov // Proc. of the 4th Intern. Conf. on Inductive Modelling (ICIM'2013), (Kyiv, September 16 – 20 2013). – Kyiv, 2013. – P. 139 – 142.
22. Балабанов О.С. Системи ймовірнісних залежностей: графові та статистичні властивості / О.С. Балабанов // Математичні машини та системи. – 2009. – № 3. – С. 80 – 97.
23. Балабанов О.С. Базовані на незалежності методи індукції каузальних мереж і сепарація в орграфах / О.С. Балабанов // Матеріали VI Всеукр. наук.-практ. конф. "Інформатика та системні науки" (ІСН-15), (Полтава, 19–21 березня 2015 р.). – Полтава, 2015. – С. 12 – 16.
24. Peters J. Structural intervention distance for evaluating causal graphs / J. Peters, P. Bühlmann // Neural Computation. – 2015. – Vol. 27, N 3. – P. 771 – 799.

Стаття надійшла до редакції 24.11.2015