УДК 004.855:519.216

О.С. Балабанов

ВІДКРИТТЯ ЗНАНЬ У ДАНИХ ТА КАУЗАЛЬНІ МОДЕЛІ В АНАЛІТИЧНИХ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЯХ

Оглянуто методологію індуктивного виведення каузальних моделей. Аргументовано, що каузальні мережі, відтворені з даних спостережень (без апріорних знань), адекватно відображають структури зв'язків та впливів у середовищі і придатні для прогнозування наслідків керування. Окреслено передумови та вимоги до статистичних даних і процесу їх збору для успішного виведення адекватної каузальної мережі. Розглянуто підхід до виведення каузальних мереж, базований на незалежності. Підхід підтримує розробку швидких та асимптотично-коректних методів, які здатні працювати в умовах прихованих факторів. Аргументовано, що модель, виведена з даних, зазвичай має деякі зв'язки з невизначеною спрямованістю. Така невизначеність об'єктивно зумовлена й дозволяє зберігати адекватність моделі. Показано засоби підвищення ефективності виведення моделі за рахунок озброєння алгоритмів набором резолюцій, які забезпечують усікання простору пошуку сепараторів (фокусуючи процес верифікації зв'язків). Пропонована модернізація методів ґрунтується на систематичному застосуванні концепції локально-мінімального сепаратора та марковських властивостей моделей. Ефективність нових алгоритмів «Razor» продемонстрована контрольними експериментами та предметним прикладом. Роз'яснюється відмінність каузального прогнозу (що оцінює наслідки планованого втручання) від традиційного «пасивного» прогнозу. Показано можливості оцінювати каузальний ефект на основі неповно ідентифікованої моделі.

Ключові слова: каузальна мережа, виведення моделі з даних, марковські властивості, умовна незалежність, структура залежностей, каузальний ефект, орієнтація дуг, d-сепарація.

Каузальні моделі – відповідь на потреби аналізу та прогнозу ефектів рішень

В багатьох організаціях та відомствах доволі часто збір статистичних даних та їх аналіз слабо пов'язані з підготовкою й вибором планів та рішень. Доволі типовою є ситуація, коли організація (орган управління) має у своєму розпорядженні великі зібрання даних, але ці дані дуже вибірково та обмежено залучаються до предметних досліджень, підготовки планів та прогнозування наслідків пропонованих управлінських рішень. Вибір і обґрунтування рішень робиться на основі експертних суджень і оцінок, адекватність яких важко контролювати. А коли аналітики все ж вдаються до побудови математичних або інформаційних моделей, то в основу цих моделей знов лягають експертні міркування та суб'єктивні уявлення. Необхідно позбуватися суб'єктивізму й консерватизму в механізмах підготовки й обґрунтування важливих рішень. Те, що інформаційні технології глибокого аналізу даних та математичного моделювання відіграють незначну роль в аналітичній роботі штабів та органів управління, певною мірою можна пояснити тим, що традиційні, давно відомі методи та моделі спираються на ідеалізовані припущення, неадекватно відображають зовнішні (приховані) фактори, а також не забезпечені ефективними процедурами контролю адекватності.

У провідних країнах стало нормальною практикою управління використовувати моделі, виведені науковими методами на основі даних спостережень за об'єктом моделювання. Користувачеві потрібна модель об'єкта, яка допомагає зрозуміти реальні процеси та взаємозалежності між різними субпроцесами й характеристиками. В практичних проблемних ситуаціях фундаментальні науки не дають потрібної моделі через те, що предметна галузь лежить на перетині різних дисциплін і характеризується взаємодією великої кількості різнорідних факторів. Зазвичай адекватна модель є невідома, а знання про об'єкт існують як сукупність розрізнених відомостей та уявлень вузьких спеціалістів, а також усталених (упереджених) переконань практиків. Таку «скирту інформації» важко узгодити, ве-

рифікувати та звести у робочу модель. Отже, шукана адекватна модель приречена бути емпіричною (за витоками) та феноменологічною й конгломеративною (за рівнем репрезентації). Актуальна задача – ідентифікація моделі «об'єктивними» методами на основі зібраних даних спостережень (рис. 1).

Аналіз, осмислення, прогнозу і пояснення треба виконувати в термінах показників та індексів, які можна виміряти на реальному об'єкті. Названі види пізнавальної діяльності потребують єдиної мови відображення процесів через реальні характеристики в їх природному об'єктивному зв'язку та взаємодії. Найбільш універсальною і зрозумілою мовою відображення зв'язків, взаємодій та впливів є причино-наслідкові відношення (в їх сучасному розумінні). Отже, спільною основою розв'язання всіх названих задач має бути каузальна модель об'єкта в середовищі. Виявлення каузальних відносин на основі даних спирається на статистичні залежності. Але перехід від статистичних залежностей до каузальних зв'язків – критичний і фундаментальний крок, що потребує ґрунтовної методологічної аргументації.

Для підтримки управління об'єктом (процесом) необхідна саме каузальна модель (а не просто модель залежностей). Відмінність каузального зв'язку від статистичної залежності можна проілюструвати на наступному прикладі. Уявіть, що аналіз даних про населення району (чи міста) показав, що відсоток захворювань на грип – значно вищий серед тих, хто за кілька днів перед тим придбав в аптеці та вживав анти-грипозні препарати. Наївний «аналітик» має підстави для висновку, що вживання препаратів є серед причин захворювання. Насправді до покупки анти-грипозних препаратів людину спонукало погіршення самопочуття, а також знання, що вона схильна до захворювання (генетично або через умови праці). Саме названі фактори є причинами як захворювання, так і покупки препаратів. Для з'ясування істини потрібно відтворити каузальну модель. (Повернемося до цього прикладу, коли розглядатимемо каузальний ефект).

Принципову відмінність каузальних та некаузальних моделей можна пояснити через різницю задач «пасивного» прогнозу та «активного» (каузального) прогнозу [1, 2]. Задача першого типу може формулюватися так: «яким правдоподібно було значення характеристики Y об'єкта у тих випадках, коли характеристики X та Z мали значення x, z. (Власне кажучи, це не прогноз, а реконструкція стану, тобто заповнення пропущених значень атрибутів, виходячи із значень інших атрибутів об'єкта). Задача каузального прогнозу формулюється у формі: «якою правдоподібно має бути значення характеристики Y об'єкта, якщо ми надамо характеристикам X та Z значення x, z відповідно. Це є прогноз наслідків (ефекту) втручання в об'єкт.

Адекватність моделі означає пра-

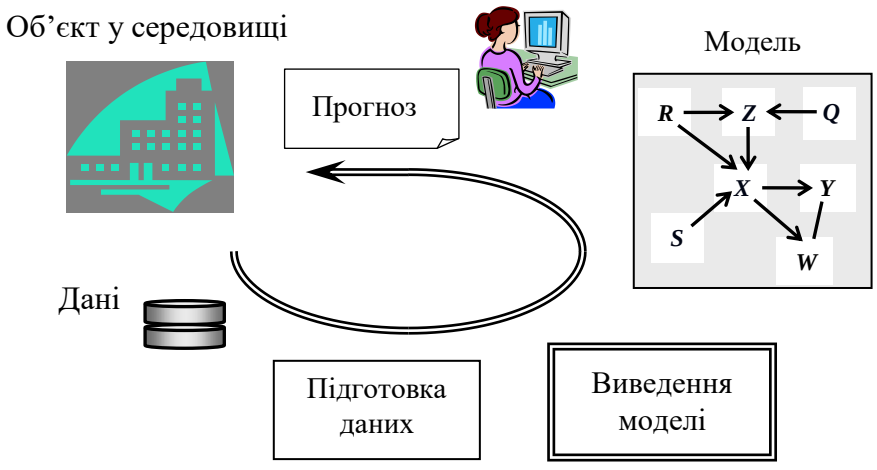


Рис. 1. Цикл інформаційних технологій з індуктивним моделюванням

вильне відображення системи впливів та зв'язків у середовищі (світі), точний опис причинно-наслідкових відношень у заданому середовищі (за умов «відкритого світу»). Традиційні методи часто не здатні знайти адекватну модель і задовольнити вищевказані потреби. Наприклад, регресійний аналіз не озброєний систематичними засобами контролю, які б забезпечили відтворення виключно автентичних зв'язків та каузальних відношень. (Регресія рідше відтворює асоціації безвідносно до їх характеру, в тому числі й «фальшиві асоціації»). Інші відомі методи теж або не забезпечують системності підходу, або надто спеціалізовані, або ґрунтуються на нереалістичних припущеннях, або неефективні у реалізації. Багато методів потребують апріорних знань, яких може й не бути. Принциповий крок від «просто статистичних» моделей до каузальних моделей було зроблено завдяки новому типу каузальних моделей і новим ефективним методам їх виведення [1–3].

Як відомо, марковський процес адекватно відображає поведінку окремого реального процесу. Якщо спробувати поширити подібне відображення на систему впливів та взаємодій між багатьма процесами й змінними (показниками), то постане потреба описувати багатовимірні марковські властивості. Знадобиться перейти від простої послідовності до графової (мережевої) структури. Але модель має описувати не епізодичний фрагмент конкретних подій, а типову узагальнену закономірну поведінку процесів. Таким чином, з'ясовуються засадничі принципи шуканих моделей. Перелічимо ці принципи: структуру зв'язків – граф (вершини відповідають змінним); ребра (дуги) графа – орієнтовані для відображення напряму впливу; принцип «зборки» моделі – умовна незалежність змінних; кількісний опис залежностей – локальний та ймовірнісний.

Кількісний аспект каузальної мережі описується фрагментами у формі умовних розподілів ймовірностей $Y \sim p(Y \mid x, z,..)$, де $x, z,...$ – значення змінних $X, Z,...$, які безпосередньо впливають на Y. Тобто компоненти розподілів ймовірностей залежної змінної є функціями безпосередніх причин $p(y_i) = f_i(x, z,..)$. В певних класах моделей локальні описи мають окрему детерміністичну частину та адитивний випадковий гамір: $y := f(x, z,..) + \varepsilon_Y$. В цьому структуральному рівнянні свідомо застосовано знак присвоєння, а не звичайна алгебраїчна рівність, для передання семантики впливу. (В спеціальній літературі використовується звичайний знак рівності, проте треба пам'ятати вказану відмінність. Не можна переносити якийсь член на інший бік структурального рівняння.) Отже, маємо систему локальних описів, інтегрованих в єдину модель без «швів». Ймовірнісний характер моделі дозволяє зберігати адекватність, незважаючи на те, що багато релевантних факторів залишилося поза аналізом і що були відхилення від схеми вимірювання даних.

Відтворення структур зв'язків та закономірностей, які неявно відбиті в багатовимірних масивах статистичних даних – одна з центральних проблем глибокого аналізу даних та відкриття знань в базах даних [4, 5]. Каузальне моделювання – передовий фронт досліджень у цій галузі. Марковський характер каузальних мереж сприяє тому, що ці моделі можна виводити індуктивно (рис. 1), тобто відтворювати на основі обробки емпіричних даних спостережень (зокрема, за відсутності апріорних знань). За відповідних передумов виведена модель буде адекватно відображати причинно-наслідкові відношення в об'єкті й середовищі. Той факт, що задачу ідентифікації каузальної моделі (за яку навіть не брався традиційний статистичний аналіз) відносять до напрямків галузі відкриття знань у даних, пояснюється наступним.

Рандомізовані експерименти на об'єкті моделювання можуть бути неприпустимими або недоступними з огляду на етичні міркування, економічні чинники, або через тимчасову недосяжність об'єкта. Наприклад, не можна проводити експерименти з населенням та економікою країни. Водночас для того, щоб «докопатися» до каузальних відносин, аналі-

зуючи дані «пасивних» спостережень, тих спостережень має бути багато, і вони мають охоплювати досить широкий спектр характеристик. Сучасна методологія відтворення каузальних мереж на основі емпіричних даних об'єднує потужні можливості технологій відкриття знань в даних з сучасними статистичними методами, і завдяки цьому дозволяє в одному циклі обробки здійснити те, що раніше розділялося на експлоративний аналіз даних та конфірмативний аналіз даних (включаючи перевірку гіпотез).

Виведення каузальних моделей з емпіричних даних призначене для пізнавальних задач і підтримки керування об'єктами та процесами у недостатньо досліджених галузях і середовищах [2, 4–8]. Каузальні мережі є багатоцільовими (на відміну, скажімо, від регресійних моделей). Вони дозволяють давати відповіді на запити у будь-якому форматі, тобто можна задавати різні цільові змінні, різні сполучення умов та ще задавати керування обраними змінними. У підсумку, клас ймовірнісних орієнтованих моделей залежностей та притаманні їм методи дозволяють втілити як комп'ютерну технологію закінчений цикл робіт (рис. 1) за схемою {вимірювання, спостереження} → дані → модель → {аналіз рішень, прогноз}.

Відомо декілька класів та різновидів каузальних мереж, які відрізняються типом зв'язків, структурними обмеженнями та формами параметризації локальних залежностей. Більшість аналітиків працює з моделями на основі ациклонних орієнтованих графів (АОГ); в цих графах заборонені структури вигляду $X \to Z \to \cdots \to X$ (тобто циклони). Клас «ординарних» АОГ-моделей (оАОГ-моделей) утворюється з використанням виключно одно-орієнтованих дуг $X \to Y$. Серед оАОГ-моделей найбільш відомі байєсові та гауссові мережі [1–7]. Байєсові мережі побудовані на дискретних змінних, а залежності описуються у формі таблиць умовних розподілів $p(Y \mid x, z, ..)$. Гауссові мережі побудовані на лінійних залежностях та нормально-розподілених змінних. Також гауссові мережі звуться системами лінійних структуральних рівнянь. Приклад одного рівняння:

$$y := a \cdot x + b \cdot z + ... + \varepsilon_Y, \ (\varepsilon_Y \sim N(m_Y, \sigma_Y^2)).$$

Узагальнені класи моделей структуруються графами, які додатково містять біорієнтовані дуги $X \leftrightarrow Y$. Така дуга відображає вплив прихованої змінної. До моделей з біорієнтованими дугами належать нерекурсивні каузальні мережі, моделі на основі анцестральних («предкових») графів та каузальні діаграми Дж. Перла [1, 2].

Передумови та запорука успішного виведення та застосування каузальних моделей

Поштовхом для розробки нових методів виведення каузальних моделей з даних стало розповсюдження технологій збору емпіричних даних, накопичення великих масивів даних та відкриття доступу до них через Інтернет. Оскільки мета полягає у виведенні каузальних зв'язків з даних пасивних спостережень, висуваються жорсткі вимоги до обсягів залучених даних. Для реконструкції адекватної моделі необхідно мати великі вибірки даних (особливо у випадках складних та нелінійних форм залежностей).

Для того, щоб виведена модель відображала певні причинно-наслідкові зв'язки, необхідно, щоб дані, подані на вхід методу, містили причини та наслідки. Методам виведення каузальних мереж необхідні дані з характеристиками відповідного рівня, організовані у певних форматах. Підготовлені дані становлять статистичну вибірку, тобто складаються з багатьох «випадків», кожний з яких містить фіксований набір характеристик, виміряних за єдиною схемою. Елементами запису виступають значення вимірянних величин у відповідні («характерні») моменти (інтервали) часу (але астрономічно різні). Дані мають відображати статистику поведінки об'єкта впродовж багаторазового проходження об'єктом типового циклу функціонування (з варіацією умов і факторів, частина яких може залишатися поза спостереженнями). Вибірка

даних передбачає повторюваність механізмів поведінки, і ця повторюваність може бути просторовою чи темпоральною. Кожний «випадок» може відповідати або окремому екземпляру популяції (індивіду), або окремій транзакції чи періоду (циклу) життя. В одних БД записи даних відносяться до різних індивідів, членів популяції, екземплярів однотипних об'єктів. В інших БД всі записи характеризують один і той самий реальний об'єкт, але в різні періоди життя, в різні цикли функціонування. (Такі об'єкти функціонують з багаторазовим перезапуском.) В останньому разі формування вибірки може потребувати розбиття («нарізку») серії вимірювань за періодами. В процесі підготовки даних потрібного змісту і формату неприпустимі «викривлення» (втручання у значення), усереднення або підміна вимірювання довільною інтерпретацією. Дані треба вимірювати якомога точніше і не допускати додавання якихось величин. Кожний елемент даних має бути виміряний точно і відображати «миттєвий» стан процесу.

Можлива проблемна ситуація, коли на структуру моделі апріорі не накладено обмежень й нічого не відомо про цю структуру. Зазвичай застережене єдине обмеження – в структурі немає орієнтованих циклів (циклонів). Відсутність циклонів можна вважати вимогою коректного збору даних. Тобто в процесі генерації одного запису даних кожна змінна X вимірюється досить швидко, так що «сигнал» від X, поширюючись до інших змінних, не встигає «оббігти коло» й вплинути на X в цьому запису.

Звичайно, якщо є достовірні апріорні знання, їх треба використати. Це дозволить прискорити виведення та уточнити модель. Типова форма апріорних знань – темпоральний порядок змінних моделі.

Оскільки оперувати занадто великою номенклатурою даних важко й недоцільно, доведеться обмежитися прийнятним набором взаємозв'язаних релевантних характеристик. Деякі змінні можуть бути недоступні для вимірювання. Практично обраний формат даних майже завжди буде неповно описувати предмет. (Тобто це буде «вікно», «кадр», вирізаний з реальності). Багато факторів залишаться поза межами обраного «вікна». Відтак, аналітик приречений працювати з моделлю у «відкритому світі». Але це не буде перепоною для виведення адекватних каузальних мереж, зокрема, завдяки ймовірнісному характеру опису. Прихованість деяких факторів впливу не заважає тестувати марковські властивості серед наявних змінних.

Далі розглядається один з перспективних підходів до виведення каузальних мереж з даних, базований на незалежності. Процес виведення за цим підходом розгортається через три послідовні фази (рис. 2).

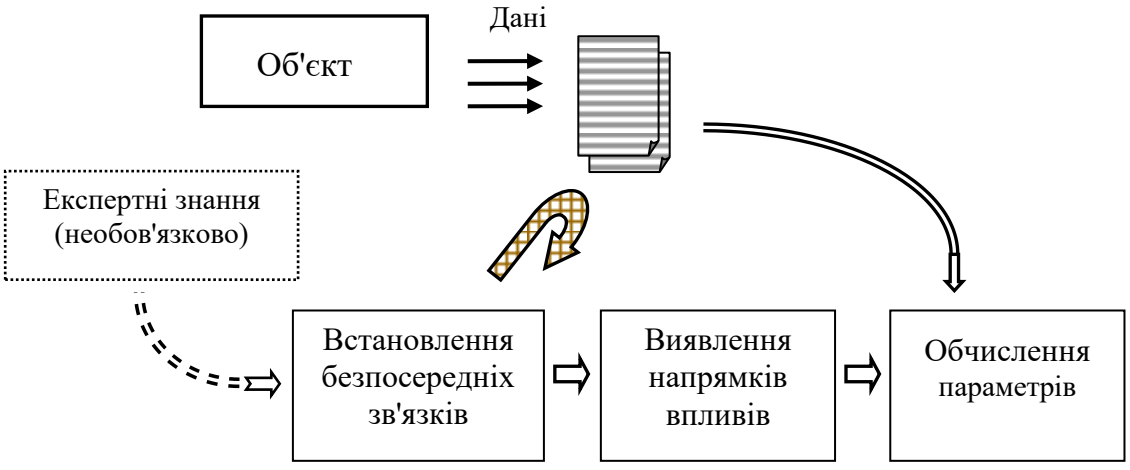


Рис. 2. Схема виведення каузальної мережі з даних

Припустимо, генеративна модель, тобто модель, яка вичерпно адекватно і однозначно відображає розгортання процесів та формування змінних об'єкта, складається виключно із звичайних односпрямованих дуг вигляду $A \to B$. Тобто нехай «повна» (уявна) модель є однозначною і точною; але вона невідома. В результаті виведення буде отримана інша модель, яка зазвичай не буде вичерпно однозначною. Це випливає з того, що різні варіанти спрямованості (орієнтації) деяких дуг не змінюють марковських властивостей моделі. Додаткова неоднозначність породжується тим, що у форматі даних не репрезентовано деякі важливі змінні. Навіть якщо прийняти постулат, що в моделі не існує жодної прихованої змінної, яка впливає рівночасно на дві (чи більше) спостережувані змінні, виведена структура має залучати дуги двох типів – неорієнтовані та каузальні. Якщо ж існування таких прихованих змінних не виключено, то для збереження адекватності виведеної моделі можуть знадобитися дуги принаймні чотирьох типів. Для відображення безпосередніх зв'язків модель залучає дуги наступних типів. Дуга вигляду $X \to Y$ відображає каузальний вплив X на Y. Дуга $U \leftrightarrow W$ позначає існування прихованої (латентної) змінної, що впливає рівночасно («паралельно») на U та W. Дуга $V \circ\!\!\to Z$ резервує два можливих варіанти: каузальний вплив або існування прихованої змінної («посередника»). Дуга $Q \circ\!\!-\!\!\circ R$ означає, що спрямованість цього зв'язку зовсім не визначена. (Деякі методи виведення моделі видають спеціальні позначки для заборони сполучень орієнтації деяких сусідніх ребер [2]).

Отже, коли в результаті виведення (серед іншого) отримали дугу вигляду $Z \circ\!\!\to Q$, прихована змінна між Z та Q можлива. Коли на виході отримуємо біорієнтовану дугу $Z \leftrightarrow Q$, це треба розуміти, що метод однозначно виявив приховану змінну. Біорієнтовані дуги (вигляду $Z \leftrightarrow Q$) ідентифікуються відомими методами тільки на основі відповідного спеціального сполучення марковських властивостей, який виникає внаслідок дії прихованої змінної у певному оточенні. Тому таку приховану змінну між Z та Q називають латентною і відрізняють від змінної-«посередника» [2, 9]. Розплутати систему залежностей буде особливо важко, коли пропущено «вузлові» змінні.

Невизначеність у проблемній ситуації робить однозначне рішення недосяжним. Виведена модель буде, по-перше, нечіткою («розмитою») у статистично-ймовірнісному сенсі, а, по-друге, на виході отримаємо лише клас еквівалентності моделей (де не визначені напрямки деяких зв'язків). Невизначеність у виведеній моделі об'єктивно зумовлена і застерігає аналітика від необґрунтованих висновків. Відзначимо, що вимушена невизначеність структури моделі не знаходить адекватної репрезентації у традиційному регресійному аналізі. До речі, клас каузальних мереж на основі оАОГ можна назвати системою регресійних моделей, поданих в інтегрованому вигляді. До цього треба додати, що в процесі виведення каузальної мережі метод знаходить коректну постановку (формат) задач регресії.

Виведення каузальної моделі з даних у форматі часових рядів має свої особливості (висока автокореляція, не виділено «випадки», невідома глибина лагу залежностей тощо). Тому процес виведення потребує додаткової підготовки та спеціальних процедур. Умовна незалежність використовувалась у виведенні моделей з часовий рядів ще у працях нобелевського лауреата К. Грейнджера. Зрозуміло, для того, щоб розплутати систему тісних взаємодій в багатовимірних рядах, потрібна досить висока частота вимірювань. Сучасний апарат каузальних мереж підносить можливості моделювання в економетриці на вищий рівень.

Коли (через невизначеність) результати не задовольняють аналітика, постає необхідність «перезавантажити» (оновити) завдання, включивши в номенклатуру даних додаткові змінні (сподіваючись, що деякі з них гратимуть роль «вузлових» або прокаузальних). Вирішення деяких проблем може потребува-

ти даних, виміряних з більшою частотою (але це – повернення до етапу збору даних).

Відтворення структури каузальної мережі з даних

Найбільш поширені два підходи до виведення моделей з даних: 1) оснований на незалежності («constraint-based», «сепараційний»); 2) «оптимізаційний», або апроксимаційний. Оптимізаційний підхід полягає у максимізації критерію якості моделі в процесі підбору структури моделі. Оснований на незалежності підхід базується на виявленні паттернів, які свідчать про відсутність дуг у структурі моделі (факти умовної незалежності). Цей підхід своєю ідеологією забезпечує декомпозицію задачі. По-перше, замість оперування цілою моделлю (або її великими фрагментами, «родинами») метод на кожному кроці розглядає «вирізку» з моделі, достатню для вирішення питання про існування відповідної дуги. По-друге, процес виведення моделі розпадається на три фази, де перебірний характер має тільки перша фаза (рис. 2).

Хоча два вказані підходи виглядають дуже відмінними, в їх фундаменті закладено єдиний принцип. Мета обох – вивести найпростішу модель, узгоджену з даними (це – сучасне розуміння принципу «Лезо Оккама»). У першому підході це «лезо» закладено у критерій якості моделі, який містить штраф за складність. У другому підході «лезо» проявляється в намаганні видалити якомога більше зв'язків, шукаючи для цього свідчення у формі умовних незалежностей. (Відзначимо, що оптимізаційний підхід потерпає від існування прихованих змінних). Найбільш відомі алгоритми сепараційного підходу – 'PC' та 'FCI'. Перший працює в класі оАОГ-моделей, другий дозволяє латентні змінні і використовує дуги чотирьох вказаних типів.

Отже, протягом першої фази для кожної пари змінних вирішується питання, чи існує між ними безпосередній зв'язок (дуга). Оскільки про орієнтацію дуги в цій фазі не йдеться, часто замість «дуга» кажуть ребро і позначають $A — B$. Теоретичним підґрунтям виведення моделі є ізоморфізм структури моделі та її марковських властивостей. Всі марковські властивості строго верифікуються суто графовим критерієм d-сепарації [1, 2, 7, 10–12]. Втім, для ідентифікації ребер моделі достатньо залучити простий наслідок з d-сепарації. (Він чинний не для всіх класів моделей). Наслідок такий: для пари вершин X,Y d-сепаратор існує тоді й тільки тоді, коли між X та Y немає дуги. Але виведення моделі з даних потребує емпіричної версії цього принципу. Якщо змінні X та Y безпосередньо не зв'язані (немає дуги), то існує такий набір змінних, що застосування його як умови робить змінні X та Y умовно незалежними. (Зрозуміло, що самі змінні X та Y не можуть входити до умови, коли тестується їх умовна незалежність).

Для обґрунтування коректності виведення моделі з даних треба «транслювати» властивості d-сепарації (з графової термінології) в емпіричну форму. Тобто необхідне припущення каузальної неоманливості, яке в загальній формі можна сформулювати наступним чином.

В розподілі ймовірностей змінних, генерованому з АОГ-моделі, для кожної пари змінних X,Y умовна незалежність (з умовою $\mathbf{S}$) чинна тільки тоді, коли $\mathbf{S}$ d-сепарує X та Y в графі моделі.

У модельному розподілі ймовірностей це припущення виконується за виключенням особливих випадків. У вибірковому розподілі ймовірностей воно виконується асимптотично. Проте методи обраного підходу потерпають навіть при наближенні до порушення припущення каузальної неоманливості, що стається доволі часто. (Неможливо відрізнити слабку залежність від прояву вибіркового ухилу).

Зазначимо, що в практиці не потрібно тотального виконання сформульованого вище припущення. Достатньо, щоб воно виконувалося в секторі пошуку сепаратора в процесі виведення. Тобто достатньо не наштовхнутися на «обманну» незалежність в процесі виконання першої фази виведення. З цієї точки зору розроб-

лені засоби звуження секторів пошуку сепараторів (див. далі) дуже корисні.

Перша фаза виведення полягає у пошуку сепараторів. Коли змінних багато і на структуру залежностей не накладено обмежень (апріорі нічого не відомо про цю структуру), пошук сепараторів стає комбінаторно важкою задачею [2, 6–8, 13]. Перебірний характер пошуку сепараторів не створює обчислювальних проблем тільки коли залежності лінійні. (Частинні кореляції швидко обчислюються з матриці парних кореляцій). Але лінійність моделі не усуває загальну проблему – ненадійність рішень щодо дуг через ризик обманних результатів тестування умовної незалежності. Обчислювальна складність зростає для нелінійних залежностей й особливо для залежностей невідомої форми. Тоді виконання кожного тесту потребує нового сканування вибірки даних (для обчислення статистики).

Важливі евристики для скорочення перебору в ході пошуку сепараторів було втілено в алгоритмі PC. Перша евристика: сепаратори підбираються і випробовуються в порядку зростання їх розміру і в циклічному обході пар змінних. Друга евристика: сепаратор для пари X,Y підбирається серед множин вершин, які вважаються (гіпотетично) суміжними відповідно до X та Y на поточний момент виведення моделі.

Було знайдено нові можливості оптимізації першої фази виведення за рахунок подальшого фокусування пошуку сепараторів. Сутність новацій зводиться до того, що знайдення одних сепараторів дає підказку для пошуку сепараторів для інших (сусідніх) пар змінних. Такі засоби й процедури прискорення пошуку сепараторів стали доступними через осягнення двох ідей. Перша – для кожної пари змінних достатньо знайти один «простий» сепаратор. Друга ідея – прості сепаратори для «сусідніх» пар змінних якось пов'язані один з одним, і «перетин» форматів сепарації треба використати разом з фактами залежності.

Теоретичним підґрунтям шуканої техніки стала концепція локально-мінімального d-сепаратора в каузальній мережі [10–12, 15, 16] та необхідні вимоги до кожного члена локально-мінімального d-сепаратора. Сепаратор **S** для пари вершин X, Y зветься локально-мінімальним, якщо після вилучення з **S** будь-якого його члена (елемента) Z «залишковий» набір $\mathbf{S}\setminus\{Z\}$ не буде сепаратором для X, Y.

Виходячи з властивостей структур залежностей та з необхідних вимог до члена локально-мінімального сепаратора, було виведено набір правил (резолюцій) мінімальної сепарації [7, 11–13, 15, 16]. Мабуть, найпростішим з цих правил є наступне.

Правило «відсторонення» кандидата 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.

Виведено також правила, які встановлюють вимоги до сепаратора у цілому. До складу кожного локально-мінімального d-сепаратора для пари вершин X, Y (якщо він не порожній) обов’язково входить щонайменше одна вершина, яка лежить на деякому ланцюгу між X та Y. Кожна така вершина Z задовольняє набір вимог: Z безумовно залежна від X та Y; Z не відсторонюється від X, Y. Вершина Z, що задовольняє цим вимогам, зветься потенційним стрижнем сепаратора для X, Y. Правило обов’язковості потенційного стрижня змушує включати в кожний пробний сепаратор принаймні один потенційний стрижень.

Зазначимо, що це правило (рівно як й інші) залишається коректним в ситуації, коли не існує жодного сепаратора для X, Y. Більш того, в таких ситуаціях правила ще корисніше.

Усі виведені правила (резолюції) мінімальної сепарації згідно їх ролі (характеру дії) можна розподілити на чотири наступні групи («родини»).

1. Родина правил суміжності (або встановлення ребра).

2. Родина правил заборони ребра (або не-суміжності).

3. Родина правил фільтрації (відкидання) кандидатів у сепаратор.

4. Родина правил стрижня і правил необхідного кандидата.

Правила суміжності негайно встановлюють ребра графу моделі. Правила-заборони ребра негайно видаляють ребра з моделі (тим самим завершуючи відгалуження пошуку). Правила фільтрації (відкидання) кандидатів видаляють певні вершини (змінні) зі списку можливих членів сепаратора для відповідної пари. Нарешті, правила необхідного кандидата фіксують обов'язкового члена сепаратора (який ще не знайдений, але можливий).

Емпіричні версії описаних правил – сепараційні резолюції, застосовують факти умовної незалежності (замість d-сепарації). Коректність емпіричних сепараційних резолюцій обґрунтовується припущенням, аналогічним до вищевказаного припущення каузальної неоманливості. Сепараційні резолюції, імплантовані в алгоритм виведення структури моделі, прискорюють виведення і навіть можуть підвищити надійність.

Для випадку, коли є апріорі відомий темпоральний порядок змінних, необхідно зробити певні корекції правил локально-мінімальної сепарації. В ситуації заданого темпорального порядку ефективність правил локально-мінімальної сепарації дещо знижується, але не нівелюється. Певні правила, згідно оцінки, забезпечують прискорення пошуку сепараторів приблизно на 67 % або 50 % відносно ситуацій з невідомим темпоральним порядком.

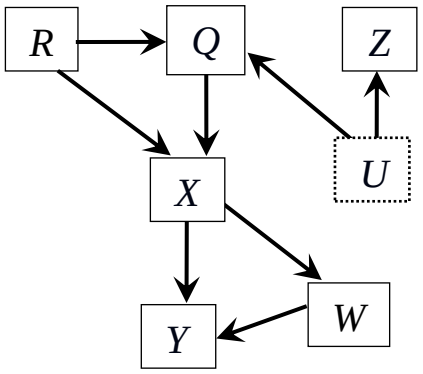
Для пояснення принципів розпізнавання спрямованості дуг (рис. 2) знадобляться деякі поняття.

Колізор (collider) – це шлях вигляду $X \to Y \leftarrow Z$ (або $X \circ\!\!\to Y \leftarrow\!\!\circ Z$). Цей колізор зветься нешунтованим, якщо відсутня дуга $X - Z$. В разі нешунтованого колізора змінні X та Z або є безумовно незалежні, або їх залежність опосередковується деякими «третіми» змінними і може бути заблокована (зруйнована). *Ланцюг* – це шлях, на якому немає жодного колізора.

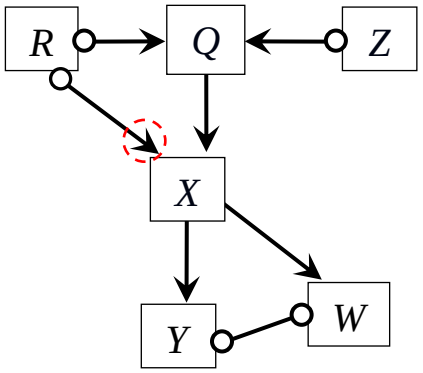
Процес орієнтації дуг стартує з застосування колізорного правила [2]. Це правило знаходить пари сусідніх дуг, які претендують стати нешунтованими колізорами. Якщо є відповідна умовна незалежність, ідентифікується колізор, тобто ставляться «вістря» дуг колізора. Потім виконується пост-колізорне правило орієнтації, яке встановлює «хвости» дуг, а далі – ідентифікує каузальні дуги [2, 9]. По-суті, для ідентифікації каузальної дуги потрібно знайти щонайменше одну прокаузальну змінну та одну квазіпрокаузальну змінну. (В [9] такі змінні неточно названі інструментальними).

На рис. 3, а показано приклад генеративної моделі, а на рис. 3, б – результат виведення після фази орієнтації дуг. Процес орієнтації дуг розпочинається з розпізнання нешунтованого колізора $R \circ\!\!\to Q \leftarrow\!\!\circ Z$. Далі працює пост-колізорне правило орієнтації, яке встановлює інші «вістря», а також «хвости» дуг. У наведеній структурі для ідентифікації дуг $Q \to X$ та $X \to Y$ як каузальних роль прокаузальної змінної зіграла Z, а квазіпрокаузальної – змінна R.

Відзначимо, що орієнтація дуги $R \circ\!\!\to X$ (рис. 3, б) не підтримується стандартними процедурами і потребує додаткового обґрунтування та пояснення. Визначення «стрілки» цієї дуги біля змінної X (на рис. 3, б обведено штриховим кільцем) аргументується наступним чином. Аналіз даних показав безумовну незалежність змінних R та Z (статистично незначущу залежність). Якщо прийняти варіант орієнтації $R \leftarrow X$, то виникає шлях $R \leftarrow X \leftarrow Q \leftarrow\!\!\circ Z$, який мусить забезпечувати залежність змінних R та Z. Тому для узгодження з даними треба прийняти саме стрілку до X на дузі $R \circ\!\!\to X$. Проте можна не погоджуватися з цією аргументацією, і не відкидати варіант $R \leftarrow X \leftarrow Q \leftarrow\!\!\circ Z$. Тоді той факт, що залежність між R та Z – незначуща, можна пояснити тим, що ця залежність є «дистанційна», тобто передається через шлях (ланцюг) з трьох послідовних дуг, і тому вона слабка. (Насправді той шлях



а



б

Рис. 3. Каузальна мережа:
а – генеративна модель;
б – виведена модель

утворений з чотирьох дуг – див. рис. 3, а). Втім, в аргументації такого штибу краще спиратися на силу (величину, тісноту) залежності, а не на довжину ланцюга зв'язку. Сила залежності – об'єктивна; довжина ланцюгу – похідна від формату завдання. Втім, необхідно мати на увазі, що в багатьох ситуаціях силу впливу некоректно оцінювати залежністю у форматі «одна від одної» [14]. В подібних ситуаціях внести ясність (і підвищити надійність ідентифікації кінцівок типу «вістря») може допомогти техніка провокованої залежності [14].

Треба звернути увагу, що невизначеність орієнтацій дуг у моделі не пояснюється невідомим темпоральним порядком змінних. В ситуаціях, де припускається існування прихованих змінних, навіть якщо точно й повністю задати темпоральний порядок змінних, це не усуває проблему невизначеності спрямованості деяких дуг. Наприклад, якщо відомо, що у темпоральному порядку змінна W стоїть раніше Y, то дуга між ними негайно уточнюється до вигляду $W \circ\!\!\rightarrow Y$ (без використання колізорного правила). Але кінцівка дуги, дотична до W, залишається невизначеною. Тоді, можна лише сказати, що Y не є причиною для W, і що всі шляхи залежності між W та Y закінчуються дугою $\rightarrow Y$.

Пропоновані засоби оптимізації пошуку сепараторів було реалізовано в алгоритмах серії "Razor". Алгоритми Razor є асимптотично-коректними [7]. Робота розроблених алгоритмів була випробувана на широкому наборі структур низької, середньої та помірно високої складності. Методологія оцінювання ефективності методів виведення каузальних моделей з даних викладена в [6, 7]. Результати випробувань показали перевагу алгоритмів Razor над базовим аналогом PC за швидкодією (кількістю тестів) і за адекватністю відтворення каузальних зв'язків [6–8]. Перевага у швидкодії – очікувана (звуження секторів пошуку сепараторів тягне зменшення кількості тестів). А перевагу у точності (адекватності) можна пояснити наступним. Фокусування пошуку сепараторів відсікає ареали високого ризику помилок тестів. Тобто зазвичай відсікаються частини простору пошуку, де ризик помилкового прийняття незалежності – високий, а ризик втрати сепаратора – малий.

Наведемо деякі результати випробування алгоритму версії Razor-1.2 на структурах байєсових мереж помірної складності. Генеративні моделі мали по 30 змінних, а кількість дуг варіювала від 60 до 120. Використані вибірки даних розміром 20000 записів. Показники кількості тестів, виконаних у ході виведення деяких моделей, показано на рис. 4.

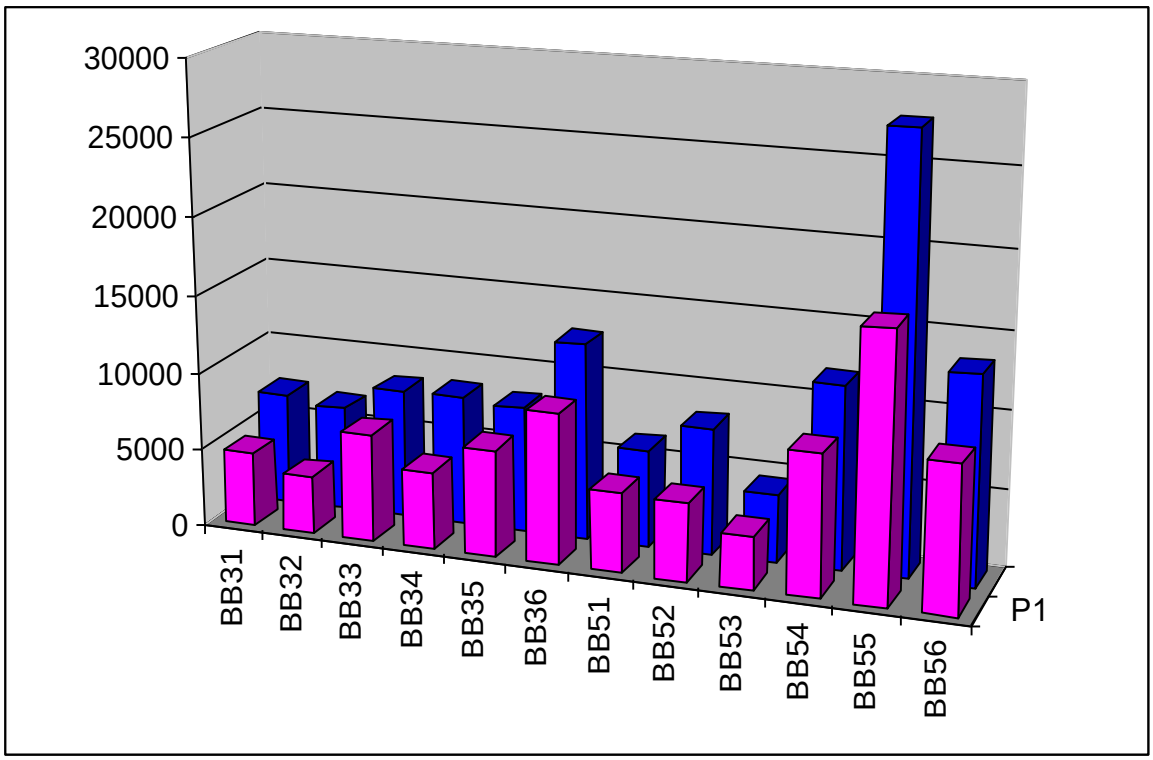


Рис. 4. Порівняння Razor-1.2 та PC за кількістю тестів. Моделі з 30-ма дискретними змінними; кількість ребер – 90 та 120

Найскладнішою виявилася модель, іменована “BB55”. Для виведення її структури алгоритму PC знадобилося 27650 тестів, а алгоритму Razor-1.2 – 16950 тестів. У ході роботи алгоритма Razor-1.2 правило «відсторонення» кандидата у сепаратор продуктивно спрацювало 132 рази. В середньому, Razor-1.2 працює у півтори рази швидше (за часом й за кількістю тестів). Важливо, що алгоритми серії Razor зменшують кількість тестів високого рангу.

Міру адекватності, забезпечувану алгоритмом, можна оцінити акуратністю алгоритму, тобто кількістю структурних помилок. Можна виділити різні типи помилок. Візьмемо тільки три з них: втрата ребра (в генеративній моделі є ребро, а у виведеній – немає); зайве ребро (навпаки); реверс ребра (орієнтація у зворотному напрямку). Реверс ребра є дуже небажаним, оскільки така структурна помилка не компенсується підбором параметрів. Результати експериментів підсумовані у таблиці. Кількість помилок дана на одну модель. Групи моделей виділено згідно кількості дуг. Алгоритм Razor-1.2, на відміну від аналога, не припустив жодного реверсу ребра.

Адекватність результатів роботи алгоритму в цілому можна оцінити «інтегральним» показником. Каузальна продуктивність [7] визначена як пропорція кількості правильно відтворених каузальних дуг відносно суми кількості помилкових дуг та кількості автентичних каузальних дуг у генеративній моделі. Ефективність алгоритмів за каузальною продуктивністю показано на рис. 5. У підсумку, алгоритм Razor-1.2 продемонстрував втричі вищу каузальну продуктивність, ніж аналог. Отже, розроблений алгоритм перевершив аналог за обома основними показниками. Ці алгоритми та втілені у них принципи є внеском у розвиток методів, базованих на незалежності.

Таблиця

Група моделей	PC			Razor-1.2		
	Втрата ребра	Зайве ребро	Реверс ребра	Втрата ребра	Зайве ребро	Реверс ребра
60	5,8	0,33	**0,33**	4	2,3	**0**
75	16,5	0,5	**0**	13	2,7	**0**
90	24,5	0,67	**0**	17,8	0,8	**0**
120	56,8	1,0	**1,17**	47,7	2,5	**0**

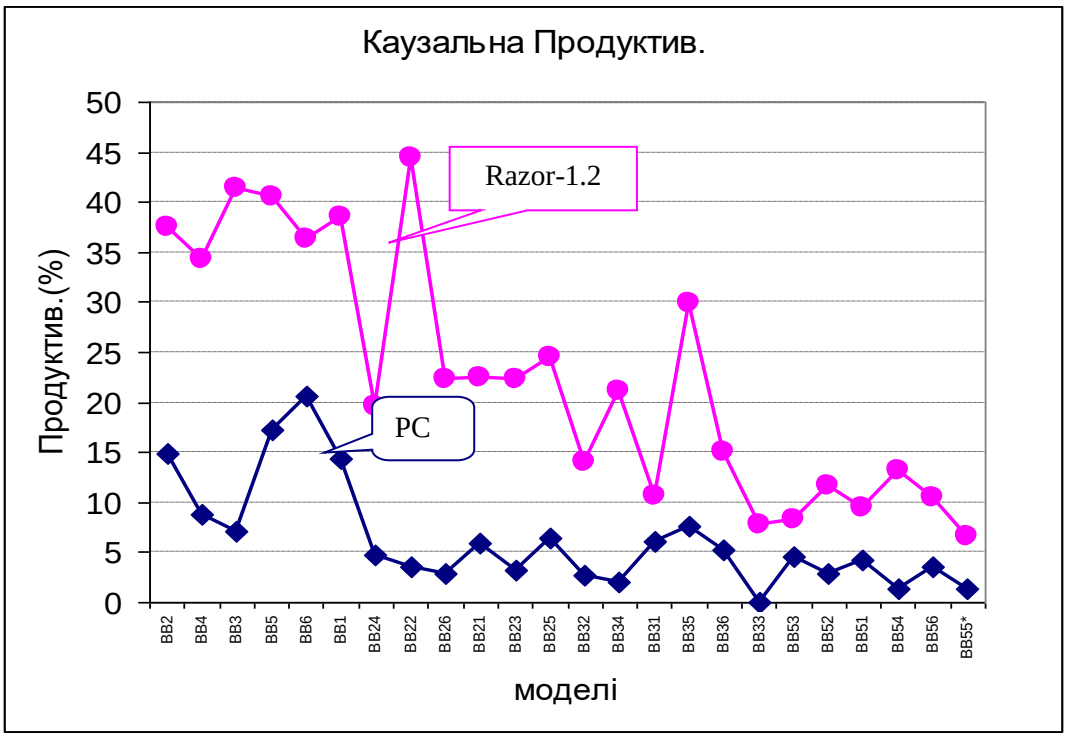


Рис. 5. Каузальна продуктивність алгоритмів; моделі з 30-ма дискретними змінними, впорядковані за тривалістю виведення

Оцінка параметрів та прогнозування каузального ефекту

У байєсових мережах (та в неадитивних моделях) роль параметрів виконують компоненти розподілу ймовірностей $p(Y \mid x, z, ..)$. В такому разі неясно, як виокремити вплив однієї змінної (окремого «батька»). Параметри характеризують усю «родину» цілком. Така ж ситуація виникає, коли форма залежності невідома й аналітик неспроможний ідентифікувати її.

Якщо значення залежної змінної формується як сума впливів її причин (батьків), то «родина» описується як адитивна модель. В адитивних моделях кожній дузі приписується свій параметр. Найпоширенішим класом адитивних моделей є лінійні.

Зрозуміло, якщо не вдається визначити орієнтацію дуги, то стає неможливим однозначно оцінити параметри відповідно-

го фрагмента моделі (список «батьків» невідомий). У структурах, перенасичених зв'язками, неможливо навіть розпочати процес орієнтацій ребер. В разі латентної змінної задача оцінки відповідних параметрів моделі стає ще більш проблематичною.

У випадку лінійних моделей з'являються додаткові можливості ідентифікації параметрів [1].

Коли і структура, і параметри ідентифіковані, задачу прогнозу втручання на певну змінну можна розв'язати однозначно. Будемо розглядати так званий «тотальний» каузальний ефект. Відрізняють ще так званий «прямий» (безпосередній) каузальний ефект, який описує вплив виключно через вказану дугу. Для оцінки каузального ефекту, який справляє одна змінна на іншу, спеціалісти розробили техніку перерахунку ймовірностей в каузальній мережі [1, 2].

Повертаючись до ілюстративного прикладу з захворюванням на грип, пояснімо відмінність прогнозу каузального ефекту від «пасивного прогнозу». Нехай у тому самому районі (місті) з наближенням несприятливого періоду в організаціях та фірмах проведено наступну профілактику серед працівників. На зібраннях перед працівниками виступили представники епідеміологічної служби, а потім було розповсюджено оплачені анти-грипозні препарати. В даній ситуації втручання (керування) не є ідеальним, бо працівники не змушені вживати препарати, і кожна людина враховує свої особисті обставини. І все ж таки внаслідок зазначених заходів буде частково розірвано (чи послаблено) зв'язок між застосуванням анти-грипозних препаратів де-факто та звичкою це робити. Тому за таким сценарієм, напевно, відсоток захворювань буде вже нижче серед тих, хто вживав препарати (а не вище). Отже, для заданої пари змінних каузальний ефект та кореляція (до втручання) мають протилежні знаки.

Технічно, оцінка каузального ефекту змінної X на змінну Y відрізняється від «пасивного прогнозу» тим, що перед перерахунком ймовірностей треба видалити з моделі дуги, вістря яких дотичне до X, і які лежать на безколізорному шляху до Y в обхід дуги $X \to Y$. Замість перерахунку ймовірностей на моделі можна виконати регресію, але до набору коваріат треба включити змінні, які блокують вказані безколізорні шляхи до Y в обхід дуги $X \to Y$ [1, 2].

Іноді каузальний ефект втручання можна оцінити навіть там, де залишається невизначеність моделі. Буває, що попри невизначену орієнтацію дуги та неповну ідентифікацію параметрів можна спрогнозувати каузальний («тотальний») ефект на основі даних. Наприклад, нехай, маючи модель рис. 3, б, потрібно визначити, як відіб'ється керування змінною X на змінній Y (ефект Y від маніпуляції на X). В цій моделі орієнтації дуги $Y \circ\!\!-\!\!\circ W$ невідома. Але достатньо знати, що W не є батьком для X. Несуттєво, чи тотальний ефект формується виключно дугою $X \to Y$, чи в цей ефект робить внесок ланцюг через W. Оцінка $p(Y \mid x)$, обчислена з даних, дає адекватне значення ефекту для обох випадків. (Істинна орієнтація дуги $Y \circ\!\!-\!\!\circ W$ «автоматично» коректно врахована в сумісному розподілі ймовірностей змінних X, Y). Сказане залишається правильним й у тому випадку, якщо асоціація між Y та W виникла як результат дії спільної прихованої причини.

Натомість для обчислення «прямого» (безпосереднього) каузального ефекту змінної X на змінну Y необхідно знати орієнтацію ребра $Y \circ\!\!-\!\!\circ W$. Все вищесказане рівною мірою чинне також для прогнозування каузального ефекту X на W.

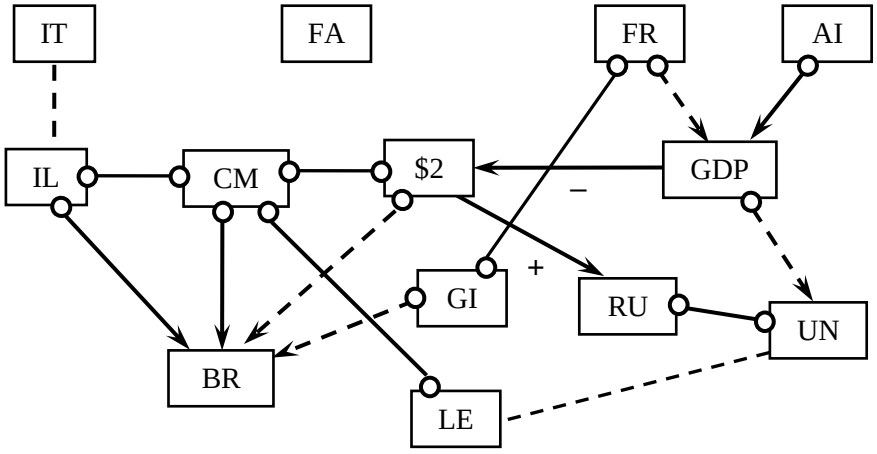
Припустимо, що модель лінійна. Розглянемо задачу обчислення структурного коефіцієнту для каузального зв'язку $Q \to X$ та каузального ефекту, який справляє Q на X. На шляху через R маємо невизначені кінцівки дуг, дотичні до змінної R. Але це не повинно зупинити аналітика. В ході виконання першої фази виведення (якщо відкинути можливість надзвичайних обставин) алгоритм знайде сепаратор $\{Q,R\}$ для пари змінних Z, X. Це означає, що варіант колізора

$X \leftrightarrow R \leftrightarrow Q$ треба виключити, цей шлях є ланцюгом. Так, каузальний зв'язок $Q \to X$ є конфаундований (сплутаний). Тому для оцінки структурного коефіцієнту та каузального ефекту для цієї дуги необхідно блокувати ланцюг через R. (Треба виконувати регресію X на пару Q, R). Отже, попри те, що модель залишає три можливі варіанти орієнтації дуг, дотичних до R, каузальний ефект Q на X (і відповідний параметр) однозначно ідентифікується. Але якщо модель не є адитивною, то постають питання, як оцінити структурні параметри для каузального зв'язку $Q \to X$ (бо невідомий другий «батько» для змінної X), і як розуміти каузальний ефект Q на X.

Приклад виведення моделі з реальних даних

Для аналізу причин та наслідків бідності й темпів народжуваності в країнах, що розвиваються, о зібрано й підготовлено соціально-економічні дані [17]. Використовувалися дані Світового банку щодо 80-ти країн що розвиваються (включаючи Україну). Виходячи з припущення про лінійність залежностей, спочатку було обчислено матрицю парних коваріацій, а потім застосовано найбільш відомий алгоритм PC [2, 17]. Виведення моделі з тих самих даних було повторено [7] з використанням розроблений нами алгоритму Razor-1.2.

Оскільки вибірка даних дуже мала (80 випадків), виведення було повторене кілька разів з різними рівнями значущості тестів незалежності («альфа» – від 0,05 до 0,1). Виведена структура показана на рис. 6. Стабільні зв'язки (ті, що присутні за всіх рівнів «альфа»), показані неперервними лініями. Зв'язки, присутні у результатах більшості експериментів, показані пунктирно. Виявлено два каузальних зв'язки: (GDP→$2) та ($2→RU). У моделі присутні три стабільні суб-каузальні зв'язки: (AI∘→GDP), (IL∘→BR), (CM∘→BR). Зв'язок $2 з BR виявився нестабільним щодо напрямку. Отже, згідно нашої моде-



Позначення показників:

$2 – бідність_за_витратами (частка населення з витратами до двох доларів/день);
GI – коефіцієнт Джині (індекс концентрації доходів);
FR – індекс несвободи; AI – доходи сільського господарства;
LE – тривалість життя; RU – частка міського населення;
CM – дитяча смертність; IL – рівень неписьменності;
GDP – величина прибутку сімейного господарства на душу населення;
BR – народжуваність; UN – недоїдання;
IT – міжнародна торгівля. FA – допомога ззовні;

Рис. 6. Виведена модель факторів народжуваності й бідності

лі, GDP впливає на бідність за витратами, а через посередництво бідності, можливо, впливає на народжуваність. Згідно моделі, коефіцієнт Джині може впливати на GDP, але тільки через індекс несвободи. Коефіцієнт кореляції для впливу GDP→$2 дорівнює -0,61. Тобто зростання GDP призводить до зменшення бідності (за витратами). Коефіцієнт кореляції для зв'язку ($2→RU) дорівнює +0,61. Цей коефіцієнт оцінює прямий каузальний вплив, бо гіпотетичний шлях між $2 та RU через GDP та UN напевно є колізорним. Отже, зростання бідності (за витратами) призводить до зростання частки міського населення. Згідно моделі, обидва вказані коефіцієнти адекватно оцінюють каузальний ефект.

Оскільки в роботі [17] модель була виведена за допомогою алгоритму PC (який не відображає неповні орієнтації), отримані там результати відрізняються від наших. Але суперечностей немає. Обидві моделі згодні у тому, що GDP впливає на $2, і що безпосередніми факторами народжуваності, правдоподібно, є неписьменність та дитяча смертність.

Допоміжні та комплементарні засоби виведення та уточнення моделі

Останніми роками сформувалися нові напрямки досліджень, що розширюють можливості відтворення моделей новими засобами, використовуючи інші типи властивостей (не тільки марковські). Тим самим долається неспроможність традиційних методів розрізнити моделі в одному класі марковської еквівалентності. Забезпечується можливість розпізнати орієнтацію зв'язку, виходячи з характеру розподілення гамору [3]. Можна сказати, що використовується несиметрія сумісного розподілу пари змінних. (Зазначимо, що оригінальний спосіб використання несиметрії розподілу двох дискретних змінних для визначення напрямку впливу раніше був запропонований у [18]. Хоча не було дано обґрунтування того способу).

Співвідношення залежностей (нерівності або рівності), характерні для певних структур моделей, дають нові засоби верифікації та уточнення моделі. Використання співвідношень парних показників залежності, як більш надійний інструмент, може замінити тестування умовної незалежності. Особливо важливо, що виведені обмеження можуть допомогти в ситуації з прихованими змінними, коли стандартні методи не працюють. Наприклад, певні рівності підтримують виявлення прихованої спільної причини трьох, чотирьох чи більше змінних [19]. У роботі [20] знайдено нерівності для кореляцій в моделі; ці нерівності іноді можуть допомогти розпізнати присутність певного зв'язку в моделі в ситуації неповної спостережуваності.

Висновки

Каузальні мережі та методи їх виведення з емпіричних даних є відповіддю на потреби аналізу рішень та прогнозу наслідків керування в процесі планування. Такі моделі – гнучкі й зручні для застосування. Методи виведення структур залежностей та відкриття каузальних зв'язків знаходяться на передовому фронті досліджень, розробок і технологій інформатики.

Індукцію каузальної моделі можна застосувати для виявлення статистичних зв'язків у будь-якому середовищі (безвідносно до фізичної чи матеріальної природи виникнення тих зв'язків). Треба тільки ретельно зібрати достатньо даних. Методи виведення каузальних мереж здатні ідентифікувати систему зв'язків настільки повно й точно, наскільки це дозволяють зібрані дані. Глибокий аналіз даних, втілений в розглянутих методах, за сприятливих обставин здатен відкривати причинно-наслідкові відношення без апріорних знань про об'єкт дослідження. Розроблені нами засоби підсилюють відомі методи і забезпечують підвищення адекватності каузальних моделей. Для каузальних мереж доступна техніка й процедури міркувань, що забезпечують практичне застосування моделей. Адекватність прогнозу каузального ефекту (за допомогою моделі) пояснюється тим, що каузальна мережа акумулює вміст оброблених даних у формі знань, які надають механізм швидкої мобілізації потрібного «зрізу» («про-

екції») даних для потенційно всіх запитів аналітика. Побіч того, каузальна мережа приваблює своєю наочністю, що стимулює візуальний «інсайт» аналітика і дослідника проблеми.

1. Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge Univ. Press, 2000. 526 p.
2. Spirtes P., Glymour C., Scheines R. Causation, prediction and search. New York: MIT Press, 2001. 543 p.
3. Spirtes P., Zhang K. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*. 2016. Vol. 3: 3. 28 p.
4. Андон Ф.И., Балабанов А.С. Выявление знаний и изыскания в базах данных: подходы, модели, методы и системы (обзор). *Проблемы программирования.* 2000. № 1–2. С. 513–526.
5. Андон Ф.И., Балабанов А.С. Структурные статистические модели: инструмент познания и моделирования. *Системні дослідження та інформаційні технології*. 2007. № 1. С. 79–98.
6. Балабанов О.С. Відтворення каузальних мереж на основі аналізу марковських властивостей. *Математичні машини та системи*. 2016. № 1. С. 16–26.
7. Балабанов О.С. Каузальні мережі: аналіз, синтез та виведення з статистичних даних: Автореферат дис. … доктора фіз.-мат. наук. К.: Ін-т кібернетики ім. В.М. Глушкова НАНУ, 2014.
8. Balabanov O. S. On perspectives of causal networks reconstruction by independence-based methods. Proceedings of the 4th Intern. Conf. on Inductive Modelling (ICIM'2013). Kyiv, September 16–20, 2013. Kyiv, Ukraine. P. 139–142.
9. Балабанов О.С. Від коваріацій до каузальності. Відкриття структур залежностей в даних. *Системні дослідження та інформаційні технології*. 2011. № 4 . С. 104–118.
10. Балабанов А.С. Логика минимальной сепарации в каузальных сетях. *Кибернетика и системный анализ*. 2013. № 2. С. 36–47.
11. Балабанов О.С. Правила підбору сепараторів в баєсівських мережах. *Проблеми програмування*. 2007. № 4. С. 33–43.
12. Балабанов А.С. Минимальные сепараторы в структурах зависимостей. Свойства и идентификация. *Кибернетика и системный анализ*. 2008. № 6. С. 17–32.
13. Балабанов А.С., Гапеев А.С., Гупал А.М., Ржепецкий С.С. Быстрый алгоритм вывода структур байесовых сетей из данных. *Проблемы управления и информатики*. 2011. № 5. С. 73–80.
14. Балабанов А.С. Индуцированная зависимость, взаимодействие факторов и дискриминация каузальных структур. *Кибернетика и системный анализ.* 2016. № 1. С. 10–22.
15. Балабанов А.С. Формирование минимальных d-сепараторов в системе зависимостей. *Кибернетика и системный анализ.* 2009. № 5. С. 38–50.
16. Балабанов О.С. Прискорення алгоритмів відтворення баєсових мереж. Адаптація до структур без циклів. *Проблеми програмування*. 2011. № 1. С. 63–69.
17. Bessler D. A. On world poverty: its causes and effects. Food and Agricultural Organization (FAO) of the United Nations. – Research Bulletin. Rome. 2003. 50 p.
18. Балабанов О.С. Індуктивне відтворення деревовидних структур систем залежностей. *Проблемы программирования.* 2001. № 1–2. С. 95–108.
19. Андон П.І., Балабанов О.С. До відкриття латентного бінарного фактора в статистичних даних категорного типа. Доповіді НАН України. 2008. № 9. С. 37–43.
20. Балабанов О.С. Про характерні співвідношення кореляцій в деяких системах лінійних структуральних рівнянь. Доповіді НАН України. 2016. № 12. С. 17–21.

References

1. Pearl J. (2000). Causality: models, reasoning, and inference. Cambridge: Cambridge Univ. Press. 526 p.
2. Spirtes P., Glymour C., Scheines R. (2001). Causation, prediction and search. New York: MIT Press. 543p.
3. Spirtes P., Zhang K. (2016). Causal discovery and inference: concepts and recent methodological advances // Applied Informatics. 3, (3). 28 p.
4. Andon P.I., and Balabanov O.S. (2000). Vyjavlenie znanij i izyskanija v bazah dannyh. Podhody, modeli, metjdy i sistemy.

Problems in programming. 2000, (1–2). P. 513–526. [In Russian].
5. Andon P.I., and Balabanov O.S. (2006). Structured statistical models: a tool for cognition and modeling. System Research and Information Technologies. 2006, (1). P. 79–98. [In Russian].
6. Balabanov O.S. (2016). Vidtvorennya kauzalnych merezh na osnovi analizu markovskich vlastyvostej [Reconstruction of causal networks via analysis of Markov properties]. Mathematical Machines and Systems. (2016). (1). P. 16–26. [In Ukrainian].
7. Balabanov O.S. (2014). 'Causal nets: analysis, synthesis and inference from statistical data', Doctor of math. sciences thesis, V.M. Glushkov Institute of Cybernetics, Kyiv, Ukraine. [In Ukrainian].
8. Balabanov O.S. (2013). On perspectives of causal networks reconstruction by independence-based methods. Proc. of 4th Intern. Conf. on Inductive Modelling (ICIM'2013). Kyiv, September 16–20. Kyiv, Ukraine. P. 139–142.
9. Balabanov O.S. (2011). From covariation to causation. Discovery of structures of dependency in data. System Research and Information Technologies. (2011). (4). P. 104–118. [In Ukrainian].
10. Balabanov O.S. (2013). Logic of minimal separation in causal networks. Cybernetics and Systems Analysis. 49. (2). P. 191–200.
11. Balabanov O.S. (2007). Rules for picking up separators in Bayesian networks. Problems in programming. (4). P. 33–43. [In Ukrainian].
12. Balabanov A.S. (2008). Minimal separators in dependency structures: Properties and identification. Cybernetics and Systems Analysis. 44. (6). P. 803–815.
13. Fast algorithm for learning the Bayesian networks from data / A.S. Balabanov, A.S. Gapyeyev, A.M. Gupal, S.S. Rzhepetskiy. J. Automation and Information Sciences. (2011). 43. (10). P. 1–9.
14. Balabanov O.S. (2016). Induced dependence, factor interaction, and discriminating between causal structures. Cybernetics and Systems Analysis. 52 (1). P. 8–19.
15. Balabanov A.S. (2009). Construction of minimal d-separators in a dependency system. Cybernetics and Systems Analysis. 45. (5). P. 703–713.
16. Balabanov O.S. (2011). Accelerating algorithms for Bayesian networks recovery. Adaptation to structures without cycles. Problems in programming. (1). P. 63–69. [In Ukrainian].
17. Bessler D.A. (2003). On world poverty: its causes and effects. Food and Agricultural Organization (FAO) of the United Nations. Research Bulletin. Rome, 2003. 50 p.
18. Balabanov O.S. (2001). Inductive recovery of structures of dependency trees. Problems in programming. (2001). (1–2). P. 95–108. [In Ukrainian].
19. Andon P.I., and Balabanov O.S. (2008). On revealing a latent binary factor in categorical data. Reports of Nat. Acad. of Sciences of Ukraine. (9). P. 37–43.
20. Balabanov O.S. (2016). On the intrinsic relations of correlations in some systems of linear structural equations. Dopov. Nac. akad. nauk Ukr. [Reports of Nat. Acad. of Sciences of Ukraine]. (12). P. 17–21.

Одержано 30.06.2017

Про автора:

Балабанов Олександр Степанович,
доктор фізико-математичних наук,
провідний науковий співробітник.
Кількість наукових публікацій в
українських виданнях – 50.
Кількість наукових публікацій в
зарубіжних виданнях – 9.
http://orcid.org/0000-0001-9141-9074.

***Місце роботи автора*:**

Інститут програмних систем
НАН України,
03187, м. Київ-187,
проспект Академіка Глушкова, 40.
Тел.: (044) 5263420.
E-mail: bas@isofts.kiev.ua