УДК 007:681.3.00

О.С. БАЛАБАНОВ

СИСТЕМИ ІМОВІРНІСНИХ ЗАЛЕЖНОСТЕЙ: ГРАФОВІ ТА СТАТИСТИЧНІ ВЛАСТИВОСТІ

Abstract: *It is given a brief introduction to probabilistic dependency models in a class of acyclic directed graphs (DAG). It is compared expressiveness of different special subclasses of these models. A class of "mono-flow" models (where instruments factors of common effect are mutually independent) is thoroughly examined. It is rigorously derived relations between structural and Markov properties of 'mono-flow' models. Spurious association phenomenon is shown. The proposed analytical instruments are useful for constructing reasoning schemes and learning model structure from data.*
Key words: *probabilistic dependency models, causal relationship, 'mono-flow' models, Markov properties.*

Анотація: *Дається стисле введення в моделі імовірнісних залежностей на основі ациклічних орграфів (баєсівські мережі), і представлені їх можливості з опису системи факторів та ефектів для об'єкта моделювання. Показано місце спеціальних підкласів моделей. Теоретично досліджено властивості підкласу монопотокових моделей, які визначені обмеженням, що фактори спільного ефекту є взаємно незалежні. Строго виведено взаємовідношення структурних ознак цих моделей з марковськими властивостями. Описано феномен обманних асоціацій. Розроблено аналітичний інструментарій роботи з такими моделями, корисний для побудови схем міркувань від свідчень та для методів відтворення моделі з даних.*
Ключові слова: *моделі імовірнісних залежностей, каузальні зв'язки, монопотокові моделі, марковські властивості.*

Аннотация: *Дано краткое введение в модели вероятностных зависимостей на основе ациклических орграфов (байесовские сети), и представлены их возможности по описанию системы факторов и эффектов для моделируемого объекта. Показано место специальных подклассов моделей. Теоретически исследованы свойства подкласса монопотоковых моделей, которые определены ограничением, что факторы общего эффекта взаимно независимы. Строго выведены взаимоотношения структурных признаков этих моделей с марковскими свойствами. Описан феномен обманных ассоциаций. Разработан аналитический инструментарий работы с такими моделями, полезный для построения схем рассуждений от свидетельств и для методов восстановления модели из данных.*
Ключевые слова: *модели вероятностных зависимостей, каузальные связи, монопотоковые модели, марковские свойства.*

1. Вступ

Розглядається новий клас моделей, який є привабливим апаратом математичного моделювання і ефективним інструментом комп'ютерних технологій для розв'язання аналітичних задач у багатьох предметних галузях. Ми описуємо загальні властивості цих моделей, корисні для різноманітних конкретних розробок та практичних застосувань.

Імовірнісні моделі систем залежностей на основі графів – актуальна тема сучасних досліджень на перетині багатовимірного статистичного аналізу, апарату графів, теорії ймовірностей, теорії інформації та штучного інтелекту. Імовірнісний характер цих моделей є необхідним для відбиття поведінки об'єкта в умовах неповної спостережуваності (або недовизначеності). Графові моделі залежностей забезпечують більш системну репрезентацію, ніж традиційні види моделей, – логічні формули, регресійні рівняння, правила класифікації, діагностики чи розпізнавання образів і т.д. При цьому мова орієнтованих графів дозволяє відображати спрямований вплив і, в принципі, здатна виражати каузальні (причинно-наслідкові) відношення у предметній галузі. Тому у літературі їх часто звуть каузальними мережами. Іноді їх називають діаграмами впливу. На відміну від нейромережевих та регресійних моделей графові моделі залежностей є багатоцільові. Вони забезпечують виведення (міркування) в потрібному напрямку (пряме, зворотне та комбіноване виведення) та у будь-якому форматі «посилки – висновки».

Найбільшої уваги привертають моделі на базі ациклічних орієнтованих графів (АОГ-моделі), достоїнства яких – це наочність, компактність, здатність відображати причинно-наслідкові зв'язки, обчислювальна ефективність імовірнісного виведення від свідчень [1–7]. Ці властивості забезпечують ефективне розв'язання задач медичної, технічної і комп'ютерної діагностики, розпізнавання мови, прогнозування наслідків рішень і дій людини, робота чи агента. АОГ-моделі становлять апарат імовірнісних експертних систем [3, 5]. Графові моделі залежностей можуть бути побудовані на основі експертних знань або ідентифіковані на основі статистичних даних. Моделі для міркувань з суб'єктивними імовірностями конструюють експертно і називають belief networks. Як приклад втілення можна назвати систему для діагностики та керування автономними рухомими засобами [8]. Доречно також згадати експертну систему підтримки рішень фірми “Hugin“, впроваджену для фінансової аналітики у страховій компанії Trygg-Hansa, – одній із найбільших у Швеції. Приклади застосування АОГ-моделей згадуються, зокрема, в [3, 6]. Одна з перших демонстрацій застосування – система “Hailfinder”, яка допомагає передбачати сильний літній град на північному сході Колорадо. Ця модель має 56 дискретних змінних та 66 дуг (зв'язків). Відомий програмний пакет для роботи з АОГ-моделями – комерційний продукт Netica фірми “Norsys“.

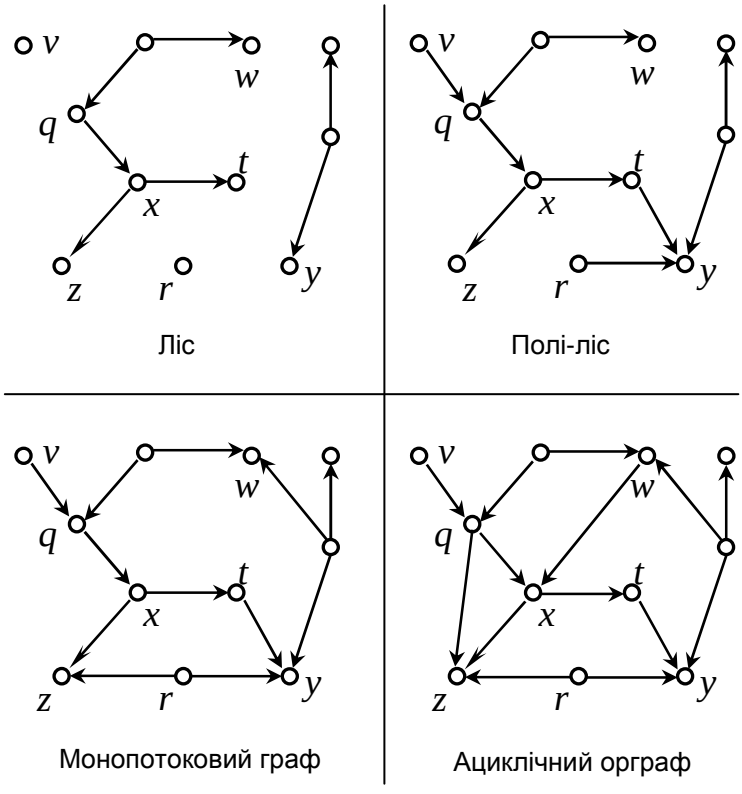

Рис. 1. Структури моделей у порядку зростання складності

З одного боку, АОГ-моделі – це ефективний апарат опису задач штучного інтелекту та роботи з суб'єктивними імовірностями; з іншого – інструмент глибокого аналізу даних та моделювання. Як статистичні моделі (на основі частотних імовірностей) АОГ-моделі підлягають верифікації, ідентифікації та індуктивному виведенню (з даних). Відтворення АОГ-моделі з статистичних даних спостережень (за об’єктом) стало привабливим напрямком досліджень та застосувань. Таке відтворення сприяє „інсайту” щодо структури та зв'язків процесів у досліджуваній предметній області і дозволяє виявити структурно адекватну систему впливів. Індуктивне виведення моделі стає інструментом для пізнавальних та науково-дослідницьких завдань (зокрема, для аналізу механізмів експресії генів) і становить один із підходів до «відкриття (каузальних) знань» в базах даних.

АОГ-моделі різняться складністю структури (рис. 1) і, відповідно, складністю відтворення та застосування. Після викладу загальних основ АОГ-моделей ми більш докладно характеризуємо

підклас монопотокових моделей, які зберігають значні експресивні можливості і в той же час забезпечують простоту багатьох процедур аналізу та відтворення моделі.

2. Ациклічні орієнтовані моделі. Побіжний огляд

У графових моделях кожній змінній відповідає вершина графа. АОГ-модель визначена як $< G, \theta >$, де G – ациклічний орієнтований граф, а θ – сукупність локально заданих параметрів. Орграф *G* визначено множинами вершин **V** та дуг **A**. Параметри θ описують умовні розподіли імовірностей значень змінних $p(x \mid F(x))$, де $F(x)$ слід розуміти як множину всіх тих вершин z, що для кожної є дуга $z \to x$. Сумісне розподілення імовірностей всіх змінних АОГ-моделі описується у формі

$$p(x_1,...,x_n) = \prod_i p(x_i \mid F(x_i)). \quad (1)$$

Формула (1) ґрунтується на базових марковських властивостях цих моделей. Поєднання графа з статистичним описом моделі здійснюється за допомогою поняття умовної незалежності (марковської властивості).

Відношення умовної незалежності змінних x та y при фіксації значень (блокуванні) набору змінних **Z** будемо виражати предикатом у формі Pr($x \perp \mathbf{Z} \perp y$), де $x, y \notin \mathbf{Z}$. Для дискретних змінних незалежність Pr($x \perp \mathbf{Z} \perp y$) означає $p(y, x \mid \mathbf{Z}) = p(y \mid \mathbf{Z}) \cdot p(x \mid \mathbf{Z})$. Це можна розуміти так, що коли відома інформація **Z**, змінна x не несе (додаткової) інформації про змінну y (і також обернено). Безумовну незалежність змінних x та y будемо виражати як Pr $(x \perp\!\!\!\perp y)$. В лінійних моделях (умовна) незалежність виражається тим, що відповідний часний коефіцієнт кореляції дорівнює нулю.

Відомими різновидами АОГ-моделей є: 1) моделі з номінальними (дискретними) змінними, тобто баєсівські мережі (чи тенета); 2) лінійні моделі з неперервними змінними та нормальними дистурбаціями, тобто гаусові мережі. В гаусових мережах параметрами є коефіцієнти регресійних рівнянь; ці коефіцієнти прив'язані до відповідних дуг орграфа. В баєсових мережах параметри – це умовні розподіли імовірностей. У дискретних моделях дуга $x \to y$ має окремий параметр тільки тоді, коли немає іншої дуги $z \to y$ до тієї самої вершини y. (Крім того, можна визначити окремий параметр для кожного фактора серед кількох у разі спеціальної форми взаємодії факторів. Це так звана каузальна незалежність, згадана в підрозд. 2.1.) Баєсівські мережі застосовуються у штучному інтелекті, зокрема, як апарат експертних систем. Гаусові мережі звичайно описують не як (1), а системою рівнянь для змінних. Гаусові мережі, поміж іншого, застосовуються як економетричні моделі. Вони відомі також під назвою рекурсивні системи лінійних структуральних рівнянь.

Нагадаємо елементарні поняття. Позначатимемо дугу знаком $\to$. Ребро – це дуга, орієнтація (спрямування) якої може бути невідома чи ігнорується. Позначатимемо ребро у вигляді $x — y$. Вершини графа звуться суміжними, якщо вони поєднані ребром. Шлях (суміжності) у графі – це послідовність дотичних ребер (будь-якої орієнтації), де всі вершини різні. Як виняток, крайні вершини шляху можуть збігатися, і тоді цей шлях називають циклом. Оршлях (тобто строго орієнтований шлях) – це шлях, на якому всі ребра орієнтовані узгоджено, тобто в напрямку того

самого кінця шляху. Орієнтований цикл (орцикл, «циклон») – це оршлях, на якому початкова вершина співпадає з останньою. Якщо в орграфі G є дуга $x \to y$, то вершина x зветься "батьком" вершини y, а вершина y зветься дитиною вершини x. Якщо існує оршлях $x \to \cdots \to y$ з x до y, то вершина x зветься предком вершини y, а вершина y – нащадком x. З огляду на взаємно-однозначну відповідність, терміни змінна (моделі) та вершина (графа) вживаються як взаємозамінні.

Визначення 1. Колізором (колайдером) в орграфі зветься фрагмент із двох суміжних дуг вигляду $x \to y \leftarrow z$. Колізор $x \to y \leftarrow z$ зветься шунтованим (екранованим), якщо у графі є ребро $x — z$, інакше – нешунтованим. Якщо колізор $x \to y \leftarrow z$ є частиною шляху **T** в орграфі, то y називають колізорною (колайдерною) вершиною на шляху **T**. (Зазначимо, що вершина y одночасно може бути неколізорною на якомусь іншому шляху.) Безколізорний шлях або ланцюг в орграфі – це шлях, який не містить жодного колізора.

Підкласи АОГ-моделей визначаються "топологічними" обмеженнями на структуру графа моделі. Зокрема, відомі: монопотокові моделі, полі-ліси і ліси. АОГ – це орграф, у якому немає циклонів (орциклів). Монопотоковий граф – це орграф, у якому кожен цикл суміжності має два чи більше колізорів. Полі-ліс – це орграф, у якому немає циклів суміжності. Ліс – це орграф, у якому немає циклів суміжності і колізорів (рис. 1). Послідовно підсилюючи обмеження на структуру орграфа, отримуємо більш спеціальні підкласи. Візьмемо такі градації формальних обмежень:

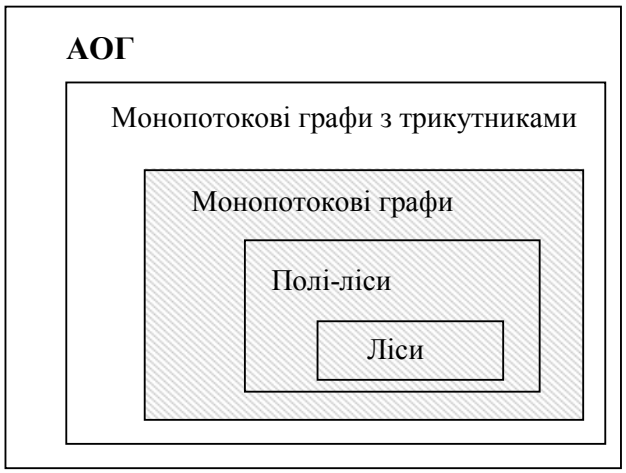

Рис. 2. Підкласи АОГ-моделей

- відсутність циклонів;
- відсутність циклонів, циклів з одним колізором та циклів з чотирма чи більше ребрами;
- відсутність циклів з кількістю колізорів 0 чи 1;
- відсутність циклів (суміжності);
- відсутність циклів і колізорів.

Згідно з цими обмеженнями, отримуємо систему вкладених множин (класів) моделей, показану на рис. 2. Пропонована класифікація моделей – не лише формальна. Ці підкласи різняться експресивними можливостями, статистичними властивостями і принципами індуктивного виведення [2, 3, 9 – 13].

У лісі (дереві) кожна вершина має не більше одного батька. Відомо, що спрямування окремого ребра в лісі не відтворюється з даних [3, 9, 11]. Полі-дерево здатне виразити паттерн залежностей, який надійно характеризує каузальний вплив ("Y-конфігурація") [3, 7, 9]. Але в реальних ситуаціях такий фрагмент часто є «занурений» у контекст оточуючих факторів, внаслідок чого структура моделі опиняється у класі АОГ.

У монопотоковій моделі, на відміну від полі-лісу, кілька факторів можуть мати кілька "паралельних" спільних ефектів. Наприклад, на рис. 1 (унизу) маємо ефекти y, z з факторами x, r. Задача виведення висновків у монопотокових моделях у гіршому випадку є NP-важкою, як і в

АОГ-моделях [14]. Але типові схеми прямого та зворотного виведення виконуються в монопотоковій моделі набагато простіше (розповсюдження інформації через ланцюги, без об'єднання змінних у “кліки”). Трохи більш загальними графами є монопотокові графи, збагачені трикутниками. В цих графах дозволені цикли з трьох ребер (тобто шунтовані колізори), але одноколізорні цикли з 4-х чи більше ребер – заборонені.

Відомий простий варіант дерева – Наївний Баєсівський класифікатор (NB). Ця модель складається з кореня (змінна класу) і кількох його дітей (ознаки). Наївний Баєсівський класифікатор приваблює обчислювальною ефективністю, яка є наслідком простої структури. Для кожної ознаки є свій параметр, яких оцінюється незалежно від інших. Але NB-класифікатор може виявитися неприйнятним не лише через неадекватність припущення умовної незалежності ознак, але і з огляду на його неспроможність відобразити ієрархію ознак. NB-класифікатор не дозволяє адаптувати класифікацію до запитів з неповним набором (форматом) ознак та складною взаємодією між ознаками. З цих причин було запропоновано узагальнити модель. У результаті отримали Баєсівський класифікатор з долученням дерева ознак – tree-augmented bayes classifier, або скорочено TAN-classifier [15]. TAN-classifier позиціонується як спеціальний гібридний підклас. Структура TAN обертається на дерево після видалення вершини класу (цільової змінної), тобто фіксації його значення. Як більш гнучка і експресивна, ця модель дозволяє, зокрема, усувати надлишкову інформацію вхідного запиту. Водночас вона залишається ефективною з обчислювальної точки зору.

Ліси та полі-ліси відтворюються з даних алгоритмами квадратичної складності, базованими на тестах першого рангу (або на кількісній мірі парної асоціації). Для відтворення АОГ-моделей є відомі алгоритми з кількістю тестів порядку N^4, але ті тести – складного формату. Монопотокові моделі посідають проміжну позицію в щойно оглянутій ієрархії і можуть пропонуватися для впровадження в простих експертних системах діагностики, в системах пошуку інформації у Web, в інтелектуальних автономних агентах та роботах тощо.

Відношення між структурою АОГ-моделі і фактами умовної незалежності, представленими цією структурою, формалізовано за допомогою критерію d -сепарації.

Визначення 2. Шлях π в АОГ-моделі називають d -закритим (d -блокованим) за допомогою (кондиціонування) множини вершин **S**, якщо і тільки якщо

1) існує вершина x, $x \in \mathbf{S}$, $x \in \pi$, причому на шляху π є дуга $x \to$ чи $\leftarrow x$ або

2) на шляху π лежить хоча б один колізор $\to y \leftarrow$, причому $y \notin \mathbf{S}$ й вершина y не є предком ніякої вершини в **S**, тобто немає $y \to \cdots \to z$, де $z \in \mathbf{S}$.

Множина вершин **S** d -сепарує вершини x та y, якщо і тільки якщо всі шляхи між x та y є d -закритими за допомогою множини вершин **S**. Будемо позначати таку d -сепарацію предикатом $Ds(x \perp \mathbf{S} \perp y)$. При цьому будемо називати **S** сепаратором для (x, y). Якщо хоча б один шлях між x та y не є d -закритим (тобто є d -відкритим), то говорять, що вершини x та y d -з'єднані.

Відомо, що критерій d -сепарації визначає марківські властивості АОГ-моделі, тобто

$$Ds(x \perp \boldsymbol{Z} \perp y) \Rightarrow \Pr(x \perp \boldsymbol{Z} \perp y). \qquad (2)$$

Цей критерій допомагає зчитати з графа АОГ-моделі всі твердження умовної незалежності, які чинні за будь-якої параметризації моделі.

Методи відтворення АОГ-моделей з даних базуються на припущенні каузальної необманливості (faithfulness) розподілу імовірностей щодо генеративної АОГ-моделі [2, 3]. Це припущення виражається як імплікація, обернена відносно (2). Чинність припущення необманливості забезпечує ізоморфізм структури і поведінки моделі. Але в скінченній відбірці даних такий ізоморфізм витримується тільки у локальному сенсі (і то не завжди).

Відомі методи «сепараційного» підходу до виводу структури моделі і, зокрема, PC алгоритм [2, 3], спираються на таке правило ідентифікації відсутності ребра:

$$\exists \boldsymbol{Z}: \Pr(x \perp \boldsymbol{Z} \perp y) \Rightarrow \text{немає } x - y. \quad (3)$$

Кардинальність (ранг) умови **Z** визначає ранг умовної незалежності $\Pr(x \perp \mathbf{Z} \perp y)$. Із зростанням рангу відбувається, власне кажучи, дроблення відбірки даних. Необхідно уникати складних сепараторів та екстенсивного їх пошуку. З огляду на це введено таке поняття [16, 17].

Визначення 3. Локально-мінімальним сепаратором в АОГ для пари вершин (x, y) зветься такий сепаратор **S**, що після вилучення з **S** будь-якого його члена (елемента) він перестає бути сепаратором для (x, y). Формально це записується як $Ds(x \perp \mathbf{S} \perp y) \& \forall z \in \mathbf{S}: \neg\, Ds(x \perp \mathbf{S} \setminus \{z\} \perp y)$.

Будемо позначати локально-мінімальний сепаратор для пари вершин (x, y) як $S_{lom}(x, y)$. Неформально $S_{lom}(x, y)$ не містить зайвих (необов'язкових) членів. Мінімальність сепараторів важлива не тільки для відтворення моделі з даних, а також для алгоритмів розмірковування на моделях (в експертних системах). При виведенні від свідчення x до цільової змінної y інформація має проходити через кожну $z \in S_{lom}(x, y)$ кожного $S_{lom}(x, y)$.

3. Дослідження монопотокових моделей залежностей

3.1. Базові положення

Моделі, що розглядаються в цьому розділі, відомі під назвами "прості" та "спрощені" [12, 18]. Однак назва "прості" сприймається неточно, і до того ж термін "прості графи" має інший зміст. Зустрічається також назва directed-path singly-connected Bayes networks. Можна було б назвати цей клас моделями з незалежними причинами. Проте найменування «каузальна незалежність» (causal independence) вже зафіксовано в літературі і розуміється інакше: як незалежність впливів («внесків») різних причин у формування спільного ефекту [19]. Мається на увазі обмеження функції взаємодії факторів, тобто звуження множини припустимих значень локальних параметрів баєсової мережі. При цьому дискретні фактори взаємодіють, як у лінійній моделі. У пропонованих нами моделях причини (аргументи) заданого ефекту (відгуку) надходять взаємно незалежно (як у класичному факторному аналізові), але ніяких рамкових обмежень на параметри формування ефекту не накладається. Будемо називати цей клас моделей □исовувань□ві□и графами залежностей (МПГЗ). Простим прикладом МПГЗ є дворядний орграф за схемою «хвороби→симптоми», причому всі хвороби є взаємно незалежні.

Коли точна структура є загальним випадком АОГ-моделі, монотопотокові структури залежностей можна застосувати в апроксимаційному режимі, тобто для приблизного моделювання і здійснення висновувань. Але в такому разі нижченаведені властивості не будуть чинні в емпіричних даних, отриманих з справжньої генеративної моделі (з об'єкта).

Визначення 4. Монопотоковим графом є такий АОГ, в якому між батьками однієї й тієї ж самої вершини немає жодного ланцюга.

Формально це обмеження виражається через таку аксіому.

Аксіома МПГЗ:

$$\forall x, y, z : (x \to y) \,\&\, (z \to y) \Rightarrow Ds\left(x \perp\!\!\!\perp z\right). \tag{4}$$

Це означає, зокрема, що всі колізори нешунтовані.

Альтернативний варіант аксіоматизації МПГЗ демонструється таким твердженням.

Твердження 1. В рамках АОГ наступні два обмеження еквівалентні:

(i) $\forall x, y, z : (x \to y) \,\&\, (z \to y) \Rightarrow Ds\left(x \perp\!\!\!\perp z\right)$;

(ii) $\forall r, q, w : (r \to q) \,\&\, (q \to w) \Rightarrow Ds\left(r \perp q \perp w\right)$.

Доведення. Імплікація (*i*) $\Rightarrow$ (*ii*). Припустимо протилежне, нехай маємо $r \to q$, $q \to w$, але при цьому буде невірним $Ds(r \perp q \perp w)$. Це означає (згідно з d-сепарацією), що або (а) існує ланцюг між r та w, який не проходить через вершину q, або (б) є ланцюг вигляду $q \leftarrow \cdots - w$. У випадку (а) отримуємо цикл $w - \cdots - r \to q \to w$, який водночас є ланцюгом. Тобто отримуємо або циклон $w \to \cdots r \to q \to w$, або ланцюг вигляду $w \leftarrow \cdots - r \to q \to w$, де два батьки вершини w є d-залежні, що суперечить (i). Випадок (б) теж призводить до утворення циклону чи циклу з одним колізором (на вершині w).

Доведемо імплікацію (ii) $\Rightarrow$ (i). Припустимо протилежне: нехай маємо $x \to y$, $z \to y$ та $\neg Ds(x \perp\!\!\!\perp z)$. Останнє значить, що існує ланцюг між x та z. На тому ланцюгу мусить бути дуга $v \to x$ або $u \to z$. Згідно з d-сепарацією, буде $\neg Ds(v \perp x \perp y)$ та $\neg Ds(u \perp z \perp y)$ відповідно, що суперечить (ii).

Варто зауважити, що між цими варіантами (i), (ii) аксіоматизації МПГЗ існує прихована різниця, яка виявляється при переході до статистичних фактів і пов'язана з формами (варіантами) припущення необманливості.

Аксіома монопотокових моделей тягне такі наслідки.

Наслідок 1. У МПГЗ чинне

$$(x \to \cdots \to y) \;\&\; (z \to \cdots \to y) \Rightarrow Ds(x \perp\!\!\!\perp z).$$

Доведення. На оршляхах $x \to \cdots \to y$ та $z \to \cdots \to y$ лежать відповідно вершини q та r, які є батьками для вершини y. Якби існував якийсь ланцюг між x та z, то існував би ланцюг між вершинами q та r, що суперечить аксіомі (4).

Цей наслідок легко прочитати як заборону одноколізорних циклів. Еквівалентно ніякий цикл не може бути ланцюгом.

Наслідок 2. Якщо в МПГЗ існують оршляхи $r \to \cdots \to x$ та $r \to \cdots \to y$, то не існує жодної вершини w такої, що існують оршляхи $x \to \cdots \to w$ та $y \to \cdots \to w$. Зокрема, виключаються також випадки $w \equiv x$ та $w \equiv y$. Тобто за умови існування оршляхів $r \to \cdots \to x$ та $r \to \cdots \to y$ неможливі ні оршлях $x \to \cdots \to y$, ні оршлях $x \leftarrow \cdots \leftarrow y$, ні ребро $x — y$.

Доведення – тривіальне, бо заперечення висновку призводить до існування одноколізорного циклу.

Зауваження. Наслідки 1 та 2 спричиняють неможливість в МПГЗ трикутника з трьох оршляхів.

Наслідок 3. У МПГЗ неможливий трикутник ребер.

Наслідок 4. Якщо в МПГЗ між вершинами x та y існує оршлях, то між x та y не існує ніякого іншого ланцюгу.

Доведення – тривіальне, бо заперечення висновку імплікує існування одноколізорного циклу.

З наслідку 4 негайно випливає наступне.

Наслідок 5. У монопотокових моделях чинне

(*i*) $(x \to \cdots \to y) \& (y \to \cdots \to z) \Rightarrow Ds(x \perp y \perp z)$;

(*ii*) якщо між вершинами x та y існує ланцюг, який не є оршляхом, то між x та y не існує жодного оршляху;

(*iii*) з заданої вершини x до кожного її нащадка веде тільки один відповідний оршлях; іншими словами, в кожну задану змінну з кожного її предка веде лише один оршлях.

Наслідок 6. Якщо в МПГЗ між вершинами x та y існує оршлях, який не є ребром, то для кожної вершини z на цьому оршляху чинне $Ds(x \perp z \perp y)$.

Доведення. Оскільки оршлях δ, який поєднує вершини x та y, не є ребром, то на ньому лежить принаймні одна якась вершина z, яка не співпадає ні з x, ні з y. Вершина z d-закриває шлях δ. Згідно з наслідком 4, між вершинами x та y не існує жодного іншого ланцюга, крім оршляха δ. Отже, вершина z d-сепарує вершини x та y.

Наслідок 7. Якщо в МПГЗ між вершинами x та y існує більше одного ланцюга, то серед тих ланцюгів немає жодного оршляху, а всі ці ланцюги мають вигляд $x \leftarrow \cdots \to y$ (тобто всі вони закінчуються стрілкою в кінцеві вершини).

Доведення – випливає з наслідку 4.

3.2. Конструювання монопотокових моделей

Визначальна ознака монопотокових структур моделей – в них немає оршляхів («потоків»), які дублюють один одного. Тобто не існує двох чи більше оршляхів, які розпочинаються з певної вершини x і закінчуються в певній вершині y. В задану змінну з кожного її предка веде єдина “генеалогічна гілка”, і це пояснює назву класу структур – монопотокові графи. Для аналізу та конструювання МПГЗ (з експертних знань) потрібні зручні поняття та апарат. В [17] було запропоновано апарат генотипів змінних та інші поняття, які ми уточнюємо нижче.

Ген змінної (вершини) в АОГ – це її кореневий предок. Генотип змінної – це множина її кореневих предків. Популяція змінних (вершин) – це множина змінних, які мають одну й ту саму

множину кореневих предків. Зрозуміло, що по дузі та по оршляху “успадковуються” усі кореневі предки. Отже, популяція – це множина змінних однакового генотипу. В МПГЗ кожний ген успадковується змінною тільки від одного з своїх батьків. Таким чином, якщо змінна y заданої популяції має батька, який належить до тієї самої популяції, то інших батьків y не має. Відтак логічно виникає поняття клонколонії (або “клану”). Клонколонія – це така множина **L** змінних однієї популяції, що всі “зовнішні” дуги, які надходять до клонколонії **L**, надходять (безпосередньо) до однієї змінної x, а всі інші змінні клонколонії **L** є нащадками змінної x. Така змінна x називається фундатором клонколонії **L**. Зрозуміло, що клонколонія має структуру дерева, коренем якого є фундатор. Діти та нащадки фундатора (фундаторів), які самі не є фундаторами, називаються клонами.

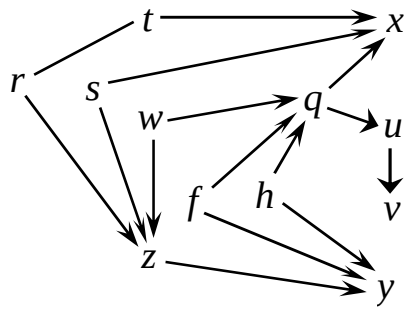

Рис. 3. МПГЗ з 5-ма генами та 4-ма фундаторами

Ясно, що в МПГЗ кожна колізорна змінна є фундатором клонколонії. І, навпаки, кожний фундатор клонколонії є колізорною змінною. Виняток становлять фундатори клонколоній, які мають лише одного кореневого предка. Такий фундатор не отримує “зовнішніх” дуг, а відтак, коли ця клонколонія складається з кількох вершин, неможливо розпізнати серед них фундатор (на підставі лише статистичних даних). Отже, фундатором такої клонколонії можна вважати довільну змінну цієї клонколонії. Тому можна не виділяти фундатора серед змінних одногенної клонколонії. Всі змінні такої клонколонії можна називати квазі-кореневими. Іншими словами: кожний ідентифіко-спроможний фундатор клонколонії є колізорною змінною. Фундаторами будемо називати тільки вершини, які мають кілька батьків. Усі інші змінні – це квазі-кореневі змінні, а також діти та нащадки колізорів. Таким чином, всі змінні (вершини) МПГЗ-моделі поділяються на: 1) фундатори; 2) клони; 3) кореневі та квазі-кореневі змінні. Наприклад, на рис. 3 маємо: вершини q, x, y, z – фундатори; вершини u, v – клони; вершини f, h, s, w – кореневі; вершини r, t – квазі-кореневі. Популяція є об'єднанням клонколоній. Кожна змінна входить до складу лише однієї клонколонії. Кожна однокоренева популяція складається тільки з однієї клонколонії. У типових випадках названі роди змінних (вершин) можна розуміти так: кореневі змінні – це «початкові» фактори (наприклад, захворювання); фундатори – це ефекти сумісного впливу кількох факторів; клони – це індикатори (синдроми) відповідного ефекту. При побудові МПГЗ корисно керуватися правилом: кожний свій ген змінна отримує тільки від одного з батьків.

Якщо обмеження на топологію, властиві МПГЗ-моделям, стають неприйнятними, можна обійти їх за допомогою технічного трюку («вигадки»), залишаючись у класі МПГЗ. А саме там, де виникає потреба запровадити одноколізорний цикл, вводять дублікат клонколонії, створюють фрагмент моделі з забороненим циклом, а потім клонколонія з найбільшим генотипом має поглинути всі клонколонії (крім одногенних) на оршляху, який утворює заборонений цикл. Внаслідок такої трансформації модель розростається вшир. Зрозуміло, що так трансформована модель частково втрачає структурну адекватність і в ній треба утриматись від одночасних операцій з кількома фрагментами, які (неявно) містять дублікати один одного. Якщо модель і після такої

трансформації не задовольняє аналітика, можна спробувати побудувати потрібну модель у підкласі монопотокових моделей, збагачених трикутниками. У цих моделях процедури виведення від свідчень майже такі самі, як у МПГЗ.

3.3. Характеристика та аналіз монопотокових моделей

Дослідимо графові та марковські властивості МПГЗ.

Твердження 2. Якщо в МПГЗ вершини x та y не поєднані ребром, $\neg Ds(x \perp\!\!\!\perp y)$ і не існує жодної такої вершини z, що чинне $Ds(x \perp z \perp y)$, то між вершинами x та y існує більше одного ланцюга, причому всі ці ланцюги мають вигляд $x \leftarrow \cdots \rightarrow y$.

Доведення. Згідно з критерієм d-сепарації, $\neg Ds(x \perp\!\!\!\perp y)$ означає існування одного чи більше ланцюгів між вершинами x та y. З факту, що не існує жодної такої вершини z такої, яка забезпечує $Ds(x \perp z \perp y)$, випливає, що між вершинами x та y існує більше одного ланцюга. Отже, з наслідку 7 випливає, що всі ці ланцюги мають вигляд $x \leftarrow \cdots \rightarrow y$.

Твердження 3. Якщо в МПГЗ існує ребро $x - z$ й існують ланцюги $\lambda: x - \cdots - y$ та $\tau: y - \cdots - z$, то ланцюги λ та τ мають на кінцях дуги, спрямовані до вершини y, тобто мають вигляд відповідно $x - \cdots \rightarrow y$ та $y \leftarrow \cdots - z$, і принаймні один з ланцюгів λ чи τ має стрілки на обох кінцях, тобто має бути $x \leftarrow \cdots \rightarrow y$ або $y \leftarrow \cdots \rightarrow z$.

Доведення. Розглянемо цикл, утворений ребром $x - z$ і ланцюгами λ та τ. Цикл повинен мати щонайменше два колізори. Ці два колізори не можуть бути на вершинах x та z одночасно, бо ребро $x - z$ має стрілку (вістря) тільки на одному своєму кінці. Звідси один з колізорів має бути утворений у точці стикання ланцюгів λ та τ (це або вершина y, або її предок), а другий колізор буде на вершині x чи z. Отже, зважаючи на твердження 2, отримуємо потрібне.

Твердження 4 (про пару дотичних ребер). Якщо в МПГЗ існують ребра $x - y$ та $y - z$ і чинне $\neg Ds(x \perp\!\!\!\perp z)$ та $\neg Ds(x \perp y \perp z)$, то буде: а) $x \leftarrow y \rightarrow z$; б) існує щонайменше один інший ланцюг $x \leftarrow \cdots \leftarrow w \rightarrow \cdots \rightarrow z$, де $Ds(y \perp\!\!\!\perp w)$; в) вершина y входить до складу всіх сепараторів для пари вершин (x, z); г) всі ланцюги між вершинами x та z мають вигляд $x \leftarrow \cdots \rightarrow z$.

Доведення. Якби вершина y була колізорною на шляху $x - y - z$, то факт $\neg Ds(x \perp\!\!\!\perp z)$ суперечив би аксіомі МПГЗ. Якби шлях $x - y - z$ був би оршляхом, то, згідно з наслідком 5(і), було б $Ds(x \perp y \perp z)$. Залишається єдиний варіант $x \leftarrow y \rightarrow z$. З факту $\neg Ds(x \perp y \perp z)$ випливає існування щонайменше ще одного ланцюга між x та z. Згідно з наслідком 7, всі ланцюги між вершинами x та z мають вигляд $x \leftarrow \cdots \rightarrow z$.

Твердження 5 (про трикутник ланцюгів). Якщо в МПГЗ існують три ланцюги, які утворюють цикл, тобто маємо ланцюги $\lambda: x - \cdots - y$, $\tau: y - \cdots - z$ й $\eta: z - \cdots - x$, то принаймні один з цих трьох ланцюгів (назвемо π) стикається з двома іншими, утворюючи колізори, і на ланцюзі π лежить

вершина r, $r \notin \{x, y, z\}$ така, що чинне відповідно $Ds(r \perp\!\!\!\perp x)$ або $Ds(r \perp\!\!\!\perp y)$, або $Ds(r \perp\!\!\!\perp z)$.

Доведення. З наслідку 3 випливає, що принаймні один з ланцюгів циклу (λ, τ чи η) не є ребром. Нехай ланцюг λ не є ребром, тобто маємо $x - \cdots - q - \cdots - y$. Згідно з визначенням МПГЗ, цикл, який ми розглядаємо, мусить мати щонайменше два колізори. Ці колізори можуть бути лише на вершинах x, y чи z. Якщо, наприклад, вершина z не є колізорною, то, по-перше, колізорними є x та y, а, по-друге, приєднання ланцюга τ до η дає ланцюг між вершинами x та y, який не проходить через q. Тоді повинно бути $Ds(q \perp\!\!\!\perp z)$, бо інакше буде ланцюг між q та z, який не проходить ні через x, ні через y, і тоді отримаємо одноколізорний цикл з колізором на x чи на y.

Якщо ж у циклі з ланцюгів λ, τ та η неколізорною є вершина x, то, по-перше, колізорними є вершини y та z, а, по-друге, маємо ланцюг $y - \cdots - q - \cdots x - \cdots - z$ (конкатенація ланцюгів λ та η). Тоді, згідно з наслідком 7, маємо уточнити ланцюг τ як $y \leftarrow \cdots \rightarrow z$. Отже, на ланцюгу τ існує якась вершина w і маємо $y \leftarrow \cdots - w - \cdots \rightarrow z$. Тоді мусить бути $Ds(w \perp\!\!\!\perp x)$, бо інакше існує якийсь ланцюг ϖ між w та x, який не проходить через y або z, і тоді отримаємо або одноколізорний цикл, який створюється з ϖ, λ та з відтинку $y \leftarrow \cdots - w$, або одноколізорний цикл, який створюється з ланцюгів ϖ, η та відтинку $w - \cdots \rightarrow z$. Нарешті, якщо в циклі з ланцюгів λ, τ та η неколізорною є вершина y, то доведення буде аналогічне.

3.4. Двійникова асоціація та інші види обманних асоціацій

Цікава статистична властивість монопотокових моделей – це можливість виникнення “двійникової” (“близнюкової”) асоціації між змінними. Асоціація зветься “двійниковою”, якщо вона дужча за реберні асоціації, які її формують. Силу (міць) асоціації змінних у баєсівських мережах зазвичай вимірюють за допомогою взаємної інформації (Шеннона), яку будемо позначати як $\mathrm{Inf}(x, y)$. Наступні визначення поширюються і на гаусові мережі, але там асоціацію можна вимірювати коефіцієнтом кореляції.

Визначення 5 (twin-association). У довільному орграфі залежностей асоціацію між змінними x та y назвемо “*двійниковою*”, якщо між x та y немає жодного оршляху і на кожному ланцюзі між x та y існує щонайменше одне ребро q — z таке, що виконується $\mathrm{Inf}(x, y) > \mathrm{Inf}(q, z)$.

Визначення 5а (twin-association). У МПГЗ асоціація між змінними x та y зветься “*двійниковою*”, якщо на кожному ланцюзі між x та y існує щонайменше одне ребро q — z таке, що $\mathrm{Inf}(x, y) > \mathrm{Inf}(q, z)$.

Зокрема, зараховується випадок $q \equiv x$ або $z \equiv y$. У МПГЗ факт відсутності ребра x — y імплікується рештою умови визначення 5а. Позначатимемо двійникову асоціацію між вершинами x та y, як $x \diamond\!\!\!\!\diamond y$.

Визначення 6 (super-twin-association). У МПГЗ асоціацію між змінними x та y будемо називати "супер-двійниковою", якщо для всіх ребер $q — z$ на кожному ланцюзі між x та y виконується $\mathrm{Inf}(x, y) > \mathrm{Inf}(q, z)$.

Супер-двійникова асоціація є нереберна асоціація, яка переважає силою всі реберні асоціації, що її формують.

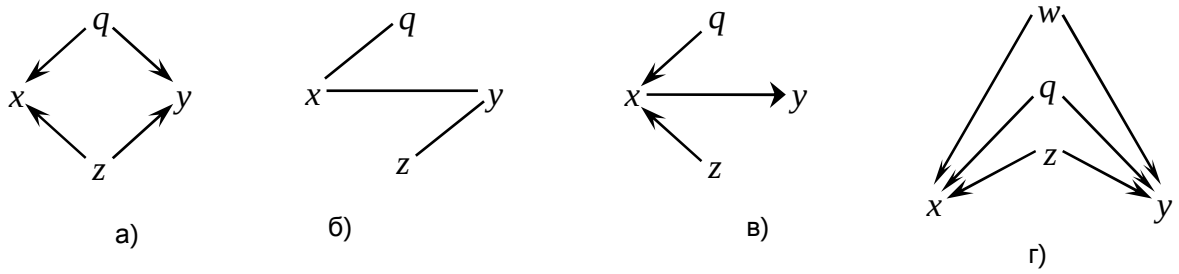

Рис. 4: а) найпростіша МПГЗ; б) апроксимація деревом;
в) апроксимація полі-деревом; г) МПГЗ з трьома генами

Двійникова та супер-двійникова асоціації спричиняють складності для відтворення структури моделі з даних і, зокрема, не дозволяють прямо застосувати алгоритми, подібні до Chow&Liu [11, 13]. Нехай генеративна модель, зображена на рис. 4а, має двійникову асоціацію $x <> y$. Тоді результатом роботи алгоритму типу Chow&Liu буде сурогат у вигляді дерева (рис. 4б) або полі-дерева (рис. 4в), чи симетричні варіанти.

Пояснимо механізм виникнення двійникової асоціації. Наведемо простий приклад, де змінні x та y мають кілька спільних батьків, як зображено на рис. 4а. Нехай маємо гаусову мережу, яка описується системою залежностей (рівнянь):

$$x := q + z + \varepsilon_1, \qquad y := q + z + \varepsilon_2. \tag{5}$$

Нехай змінні q та z розподілені нормально з дисперсією D_0, а дистурбації ε_1 та ε_2 – нормально з дисперсією D. Тоді легко вивести коефіцієнт кореляції змінних x та y, буде $\rho_{xy} = 2D_0 / (2D_0 + D)$. В той же час коефіцієнт кореляції для всіх ребер моделі, тобто для пари змінних (x, q), пари (y, z) і т.д., дорівнює $\rho_e = 1/\sqrt{2+(D/D_0)}$.

При $D_0 = D$ реберна залежність ρ_e дорівнює $1/\sqrt{3}$, а залежність між змінними x та y дорівнює 2/3, тобто в $\sqrt{4/3}$ рази сильніша за реберну. При наближенні D до нуля реберна залежність наближується до $1/\sqrt{2}$, а залежність між x та y наближується до 1, тобто переважає реберну в $\sqrt{2}$ рази. Таким чином, у розглянутих випадках асоціація між змінними x та y є не тільки двійникова, а й супер-двійникова. Легко показати, що коли в подібній моделі змінні x та y будуть мати k спільних батьків, то двійникова асоціація буде переважати реберну в $\sqrt{k}$ раз. Наприклад, для рис. 4г величина ρ_{xy} буде в $\sqrt{3}$ рази міцніше за реберну асоціацію.

У дискретних моделях описані диспропорції асоціацій можуть бути ще більш значними. Це пояснюється неадитивністю залежності змінної від кількох факторів (батьків). Реберна залежність може ослаблюватись і наближатись до нуля, в той час, як двійникова асоціація може бути максимальною. Наприклад, для тієї ж самої структури (рис. 4а) нехай всі змінні будуть бінарні і нехай значення змінних x та y формуються як сума q та z по модулю два. Тоді, в разі рівномірних розподілень змінних x та y, буде $\mathrm{Inf}(x,y)=1$, $\mathrm{Inf}(x,q)=0=\mathrm{Inf}(x,z)$.

Отже, двійникова асоціація виникає між подібними (один до другого) ефектами кількох спільних факторів, коли специфічні для кожного ефекту відхилення є порівняно слабкими. Існують й інші форми сильних непрямих асоціацій, зокрема,

Визначення 7 (Структурно-немонотонна асоціація). Асоціацію між змінними x та y в МПГЗ будемо називати структурно-немонотонною, якщо немає ребра $x — y$ і для кожної змінної z на кожному ланцюзі між x та y виконується співвідношення $\mathrm{Inf}(x,y)>\mathrm{Inf}(x,z)$ або $\mathrm{Inf}(x,y)>\mathrm{Inf}(z,y)$.

Структурно-немонотонна асоціація може й не бути двійниковою. В той же час двійникова асоціація водночас (за рідкими виключеннями) є структурно-немонотонною.

У практиці аналізу даних виникнення двійникових асоціацій є правдоподібним, зокрема, між спорідненими індексами в економетриці та соціометриці. При цьому не існує прямого каузального впливу одного на інший, щоб вірно розтлумачити їхній взаємозв'язок, треба шукати спільні фактори.

Розпізнати асоціацію (залежність) як двійникову важливо навіть для “пасивної предикції” [6] (зокрема, у класифікації). Дійсно, коли, при чинності моделі (5), ми зробимо регресію y на змінні q, z, то отримаємо помилку з дисперсією D. А регресія y на x дає дисперсію помилки $2D$, тобто вдвічі більшу.

Феномен двійникової асоціації – це особливість монопотокових моделей, якої немає в простіших класах моделей (в лісах та полі-лісах). Легко показати, що двійникова асоціація $x <> y$ є неможлива, коли між вершинами x та y існує лише один ланцюг (як у дереві). Для цього візьмемо, наприклад, змінну z, яка лежить на єдиному ланцюзі між x та y. Тоді буде $Ds(x \perp z \perp y)$, з чого випливає для взаємної інформації $\mathrm{Inf}(x,(zy))=\mathrm{Inf}(x,z)$ та $\mathrm{Inf}(y,(zx))=\mathrm{Inf}(y,z)$. З іншого боку, з теорії інформації відомо (в загальному випадку), що

$$\mathrm{Inf}(y,x)\le\mathrm{Inf}(y,(zx)) \quad \text{та} \quad \mathrm{Inf}(x,y)\le\mathrm{Inf}(x,(zy)).$$

Підставляючи вищенаведені рівняння у відповідні нерівності, отримуємо

$$\mathrm{Inf}(y,x)\le\mathrm{Inf}(y,z), \tag{6}$$

$$\mathrm{Inf}(x,y)\le\mathrm{Inf}(x,z). \tag{7}$$

Співвідношення (6) та (7) можна назвати монотонністю взаємної інформації на марковському ланцюзі залежностей. Повторюючи наведені викладки для всіх ребер ланцюга між z та y, можна показати, що всі реберні асоціації на єдиному ланцюзі між x та y переважають асоціацію між

змінними x та y. Саме завдяки цій властивості ліси (дерева) та полі-ліси залежностей вірно відтворюються з даних алгоритмом Chow&Liu або подібними до нього [11, 13].

Твердження 6. Якщо в АОГ-моделі маємо двійникову асоціацію $x <> y$, то між вершинами x та y існують принаймні два ланцюги вигляду $x \leftarrow \cdots \rightarrow y$, які не мають жодного спільного ребра.

Нагадаємо, що коли в МПГЗ між двома заданими вершинами існує більше одного ланцюга, то всі ці ланцюги мають вигляд $x \leftarrow \cdots \rightarrow y$. Отже, коли в МПГЗ є двійникова асоціація між двома змінними, жодна з цих змінних каузально не впливає на другу.

3.5. Відтворення МПГЗ та сепарація

При відтворенні моделі з даних та виведенні від свідчень виникає проблема через те, що в монопотокових моделях (як і загалом в АОГ-моделях) кардинальність сепараторів може бути доволі великою. Зокрема, в моделі, що має N змінних, єдиним сепаратором для змінних x та y може бути сепаратор з N–2 змінних. (Але доцільно зауважити, що при цьому всі члени такого сепаратора будуть взаємно незалежні.) Застосування універсальних алгоритмів (на кшталт ‘PC’) до відтворення МПГЗ є неефективним, бо може потребувати великих сепараторів та складних тестів. Наприклад, на рис. 3 сепараторами для змінних x та y є $\{r,s,w,f,h\}$, а також $\{f,h,z\}$ та $\{q,z,w\}$. На рис. 4г єдиний сепаратор для пари (x,y) – це $\{q,z,w\}$.

Твердження 7. Якщо в МПГЗ маємо $\neg Ds(x \perp\!\!\!\perp y)$ і не існує ребра $x — y$, то існує такий локально-мінімальний сепаратор для вершин (x,y), що всі його члени є безумовно незалежні один від другого.

Доведення. Коли x та y поєднані єдиним ланцюгом, твердження виконується тривіально, бо $S_{lom}(x,y)$ складається з одного елемента (наслідок 6). Нехай тепер вершини x та y поєднані декількома ланцюгами. В такому разі, згідно з наслідком 7, всі ті ланцюги мають вигляд $x \leftarrow \cdots \rightarrow y$. Тож зрозуміло, що $S_{lom}(x,y)$ можна скласти, зокрема, з тих батьків змінної x, які лежать на ланцюгах між вершинами x та y. Тоді залишається згадати аксіому МПГЗ.

Зрозуміло, коли довжина всіх ланцюгів між вершинами x та y є більше трьох ребер, то є можливість знайти потрібний локально-мінімальний сепаратор для вершин (x,y), який не складається з батьків змінної x чи y. Дійсно, оскільки всі такі ланцюги між вершинами x та y мають вигляд $x \leftarrow \cdots \rightarrow y$, то на кожному з цих ланцюгів є вершина w_k, від якої розбігаються дуги відповідного ланцюга, тобто $\leftarrow\leftarrow w_k \rightarrow\rightarrow$. Ясно, що можна скласти потрібний локально-мінімальний сепаратор з тих вершин w_k $(k = 1,2,...)$. Кожні два члени цього сепаратора є безумовно незалежні, бо в іншому разі отримаємо одноколізорні цикли з колізором на x чи на y відповідно.

Але треба зауважити, що не завжди всі локально-мінімальні сепаратори в МПГЗ складаються з змінних, які незалежні один від другого. Наприклад, на рис. 3 маємо локально-

мінімальний сепаратор $S_{lom}(x,y)=\{q,z,w\}$. При цьому чинне $\neg Ds(q \perp\!\!\!\perp z)$, $\neg Ds(z \perp\!\!\!\perp w)$ й $\neg Ds(q \perp\!\!\!\perp w)$. (Така ситуація пояснюється відкриванням колізорів.) У цій моделі також існують сепаратори $S_{lom}(x,y)=\{q,r,s\}$ та $S_{lom}(x,y)=\{f,h,z\}$, члени яких взаємонезалежні.

Достоїнством МПГЗ є можливість відтворення структури з даних за допомогою простих тестів. Шукати сепаратор для кожної пари змінних не є ефективною тактикою, бо в МПГЗ кардинальність сепараторів може бути великою. Але можна комбінувати результати різних тестів. У [18] була показана можливість відтворення монопотокових моделей тестами першого рангу. В [20] запропоновано економічний алгоритм "Генеалог" для реконструкції МПГЗ. Однак цьому алгоритму треба надати вичерпну й точну інформацію про всі парні асоціації. Для цього потрібно мати асимптотично-велику відбірку даних або апріорні знання про безумовні асоціації (замінити ці апріорні знання могло б додаткове джерело даних про парні асоціації). В [21] було показано можливість відтворення структури МПГЗ без апріорних знань та без сепарації змінних.

Можна ідентифікувати ребро без перевірки правдоподібних сепараторів, комбінуючи дотичні непрямі свідчення. В [16] обґрунтовано необхідні вимоги до членів локально-мінімальних сепараторів у АОГ. Зокрема, маємо

Твердження 8. В АОГ вершина w не входить до складу жодного локально-мінімального сепаратора для пари вершин (x,y), якщо (а) маємо $Ds(w \perp x \perp y) \,\&\, \neg Ds(w \perp\!\!\!\perp y)$ ("відсторонення" вершини) або (б) існує така змінна z, що вірне $Ds(w \perp z \perp \{x,y\})$ (відсікання "апендикса"), або (в) $Ds(w \perp\!\!\!\perp x)$ та $Ds(w \perp\!\!\!\perp y)$.

У монопотокових моделях умову (в) можна розширити: якщо $Ds(w \perp\!\!\!\perp x)$ або $Ds(w \perp\!\!\!\perp y)$, то $w \notin S_{lom}(x,y)$.

Як показано в [16], для АОГ-моделей в умовах припущення необманливості чинне таке правило "швидкої" ідентифікації ребра. Якщо змінні x та y асоційовані і для всіх інших змінних z вірне

$$\Pr(z \perp\!\!\!\perp x) \vee \Pr(z \perp\!\!\!\perp y) \vee \Pr(z \perp x \perp y) \vee \Pr(z \perp y \perp x), \qquad (8)$$

то ребро $x — y$ існує. ($\vee$ означає диз'юнкцію.)

На жаль, умова (8) не є необхідною для існування ребра не тільки для АОГ, але й для МПГЗ. Зворотне (тобто, якщо ребро $x — y$ існує, то вірне (8)) завжди чинне тільки в лісах та полі-лісах. У МПГЗ перевірка (8) не дасть усіх ребер. (Якась вершина z може бути колізорною у циклі.) Та все ж умова (8) забезпечує процедуру ідентифікації багатьох ребер у МПГЗ, і ця процедура ефективна, бо застосовує тести лише першого рангу і не потребує перевірки всіх можливих сепараторів для змінних x та y.

Але в ситуації з невеликою відбіркою даних умова (8) може стати навіть недостатньою і призводити до помилкової ідентифікації неіснуючого ребра. Дійсно, коли асоціація між змінними x та y є структурно-немонотонна (визначення 7) чи двійникова, може статися $\Pr(z \perp\!\!\!\perp x) \vee \Pr(z \perp\!\!\!\perp y)$ (тобто ці залежності статистично незначущі) для всіх змінних z на

кожному ланцюзі між x та y. Це буде сприйнято процедурою як свідчення ребра $x — y$. В таких ситуаціях більш ефективними можуть виявитися процедури на базі провокованих залежностей. Одним із варіантів є алгоритм Proliferator-D [22].

У монопотокових моделях легше, ніж в АОГ, передбачити та спланувати деякі локально-мінімальні сепаратори. Зокрема, локально-мінімальний сепаратор для пари вершин (x, y) можна сформувати у три такі кроки:

1) зібрати множину **Z** всіх таких змінних z, що $\neg Ds(z \perp\!\!\!\perp x) \& \neg Ds(z \perp\!\!\!\perp y)$;

2) відсіяти з множини **Z** всі ті змінні, які можна відсіяти правилом відсікання апендикса та правилом відсторонення вершини (згідно з твердженням 8);

3) залишити в отриманій множині тільки взаємно незалежні змінні (згідно з твердженням 7).

Якщо замінити предикати Ds(*) на Pr(*), то ця процедура стане алгоритмом виведення локально-мінімальних сепараторів з даних. Але такий алгоритм буде ненадійним при застосуванні до невеликої відбірки даних.

У [21] показано метод відтворення структури МПГЗ без сепарації змінних. Для цього застосовано інструмент провокованої залежності [10, 21]. Це один із варіантів розвитку ідеї ідентифікації структури через непрямі емпіричні свідчення.

Провокованою залежністю між x та z називається паттерн $\Pr(x \perp\!\!\!\perp z) \& \neg \Pr(x \perp y \perp z)$. Для переважної більшості АОГ-моделей чинне таке твердження [10]. Якщо маємо колізор $x \to y \leftarrow z$ з тим, що $\Pr(x \perp\!\!\!\perp z)$, то вірне $\neg \Pr(x \perp y \perp z)$, тобто провокована залежність ідентифікується. Це твердження можна розглядати як послаблену (локальну) форму припущення необманливості.

Для ефективного розпізнавання дуг та двійникових асоціацій на підставі емпіричних свідчень важлива така констатація (рис. 5).

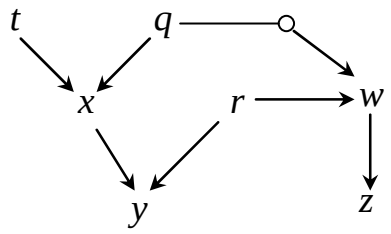

Рис. 5. Ілюстрація до твердження 9

Твердження 9 (дуга зі складним провокуванням). Нехай у МПГЗ існує дуга $x \to y$ і існують вершини z, t такі, що $Ds(z \perp\!\!\!\perp t)$, $\neg Ds(z \perp x \perp t)$, з тим, що $\mathrm{Inf}(y, z) > \mathrm{Inf}(x, z) > 0$. Тоді усі ланцюги між z та $\{x, y\}$ закінчуються стрілкою, тобто виглядають, як $z \leftarrow \cdots \to x$ та $z \leftarrow \cdots \to y$ відповідно, й існує щонайменше один ланцюг $z \leftarrow \cdots \to y$, який не проходить через x.

Доведення. З факту $Ds(z \perp\!\!\!\perp t) \& \neg Ds(z \perp x \perp t)$ випливає існування ланцюгів $t - \cdots \to x$ та $z - \cdots \to x$. З існування останнього ланцюга випливає існування ланцюга τ вигляду $y \leftarrow x \leftarrow y \cdots - z$. Далі з факту $\mathrm{Inf}(y, z) > \mathrm{Inf}(x, z)$ випливає $\neg \Pr(z \perp x \perp y)$. Оскільки вершина x не є колізорною на ланцюзі $y \leftarrow x \leftarrow \cdots - z$, то, з огляду на $\neg \Pr(z \perp x \perp y)$, існує якийсь ланцюг λ між y та z, який не проходить через x. Це є щонайменше другий ланцюг між y та z. Згідно з наслідком 7, якщо між двома заданими вершинами існує більше одного ланцюга, то всі вони закінчуються стрілкою в кінцеві вершини цих

ланцюгів. Отже, ланцюги τ та λ закінчуються стрілкою в y та z. Ланцюг λ має вигляд $y \leftarrow \cdots \rightarrow z$ і не проходить через x. Ланцюг τ має вигляд $y \leftarrow x \leftarrow \cdots \rightarrow z$.

Наслідок 8. В умовах твердження 9 мають існувати такі q, r, що $Ds(q \perp\!\!\!\perp r)$, $\mathrm{Inf}(q,x) > \mathrm{Inf}(q,y)$, $Ds(x \perp\!\!\!\perp r)$, $Ds(q \perp x \perp r) \& \neg Ds(q \perp y \perp r) \& \neg Ds(q \perp z \perp r)$ та $\neg Ds(x \perp y \perp r) \& \neg Ds(x \perp z \perp r)$.

Цей інструмент застосовано в алгоритмі Proliferator-D.

4. Заключення

Імовірнісні моделі залежностей на основі графів як засіб інформаційних технологій проникли у різноманітні сфери діяльності, від корпоративного бізнесу і технічних служб до науково-дослідницьких установ. Застосування баєсівських мереж та рекурсивних систем структуральних рівнянь набуває різних форм: від побудови імовірнісних експертних систем до методології і інструментарію глибокого аналізу даних. Моделі залежностей на базі ациклічних орієнтованих графів (АОГ-моделі) мають переваги над традиційними моделями у здатності адекватно, наочно, компактно і системно відображати складні системи зв'язків та впливів (включно з причинно-наслідковими) в умовах неповної спостережуваності, а також в ефективності механізмів розмірковувань. Аналіз даних може виявити фрагмент (паттерн) справжнього каузального впливу, навіть якщо дані збиралися за схемою пасивних спостережень. Знання автентичної структури причинних впливів, яке несе АОГ-модель, дозволяє прогнозувати наслідки активних втручань людини або робота в об'єкт, що моделюється.

У статті стисло викладено теоретичні основи АОГ-моделей залежностей. Баєсові та гаусові мережі варіюють за складністю від простих схем класифікації до великорозмірних моделей управління. Такі моделі можуть відігравати роль аналітичного портрета об’єкта і водночас бази знань для імовірнісних висновків. У цьому широкому спектрі ми виокремили клас моделей з помірно-складною топологією – монопотокові моделі, які забезпечують відносно прості процедури виведення і водночас зберігають багаті експресивні можливості. Зокрема, монопотокові моделі спроможні відобразити взаємодію незалежних факторів, “двійникові” асоціації змінних та каузальні відношення. Показано механізм виникнення двійникової асоціації та інших «обманних» асоціацій, де локально-домінуюча асоціація змінних виникає без прямого причинного зв'язку. Виявлення таких асоціацій важливо для коректної інтерпретації моделі та прогнозування.

Дано аксіоматизацію та характеристику монопотокових моделей. Досліджено зв'язок між структурними ознаками монопотокових моделей, марковськими та іншими статистичними властивостями. Знання цих зв'язків важливо для методів відтворення моделей з даних і допоможе у практичних розробках та застосуваннях. Показано, як можна знаходити локально-мінімальні сепаратори для заданої пари змінних, що є важливо для побудови схеми міркування від свідчень. Показано способи ідентифікації ребер моделі (безпосередніх зв'язків) за допомогою простих тестів. У майбутній статті планується показати механізми міркувань за допомогою цих моделей.

У цілому клас імовірнісних орієнтованих моделей залежностей і притаманні їм методи є достатні, щоб на їх основі втілити як комп'ютерну технологію увесь закінчений цикл робіт за схемою {вимірювання, спостереження} → дані → модель → {аналіз, прогноз, керування}.

СПИСОК ЛІТЕРАТУРИ

1. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. – San Mateo: Morgan Kaufmann, 1988. – 552 p.
2. The TETRAD Project: Constraint Based Aids to Causal Model Specification / R. Scheines, P. Spirtes, C. Glymour et al. // Multivariate Behavioral Research. – 1998. – Vol. 33, N 1. – P. 65 – 118.
3. Neapolitan R.E. Learning Bayesian Networks. – Upper Saddle River, N.J.: Prentice Hall, 2003. – 694 p.
4. Pearl J. Causal diagrams for empirical research // Biometrika. – 1995. – Vol. 82, N 4. – P. 669 – 688.
5. Probabilistic Networks and Expert Systems. Exact Computational Methods for Bayesian Networks / R.G. Cowell, A.P. Dawid, S.L. Lauritzen et al. – Springer, 2007. – 324 p.
6. Андон Ф.И., Балабанов А.С. Структурные статистические модели: инструмент познания и моделирования // Системні дослідження та інформаційні технології. – 2007. – № 1. – С. 79 – 98.
7. Балабанов А.С. Выделение знаний из баз данных – передовые компьютерные технологии интеллектуального анализа данных // Математичні машини і системи. – 2001. – № 1/2. – С. 40 – 54.
8. Madsen A.L., Kjærulff U.B. Applications of HUGIN to Diagnosis and Control of Autonomous Vehicles / P.J.F. Lucas, J.A. Gamez, A. Salmeron (eds.) // Advances in probabilistic graphical models (Studies in fuzziness and soft computing). – Berlin/Heidelberg: Springer, 2007. – Vol. 213. – P. 313 – 332.
9. Балабанов О.С. Відкриття структур залежностей в даних: від непрямих асоціацій до каузальності // Материалы 3-й междунар. конф. "УкрПРОГ'2002". Проблемы программирования. – 2002. – № 1–2. – С. 309 – 316.
10. Балабанов А.С. К выводу структур моделей вероятностных зависимостей из статистических данных // Кибернетика и системный анализ. – 2005. – № 5. – С. 19 – 31.
11. Балабанов О.С. Індуктивне відтворення деревовидних структур систем залежностей // Проблемы программирования. – 2001. – № 1–2. – С. 95 – 108.
12. Geiger D., Paz A., Pearl J. Learning simple causal structures // Intern. Journal of Intelligent Systems. – 1993. – Vol. 8, N 2. – P. 231 – 247.
13. Chow C.K., Liu C.N. Approximating discrete probability distributions with dependence trees // IEEE trans. on Information Theory. – 1968. – Vol.14, N 3. – P. 462 – 467.
14. Cooper G.F. The computational complexity of probabilistic inference using Bayesian belief networks // Artificial Intelligence. – 1990. – Vol. 42, N 2–3. – P. 393 – 405.
15. Friedman N., Geiger D., Goldszmidt M. Bayesian networks classifiers // Machine Learning. – 1997. – Vol. 29. – P. 131 – 163.
16. Балабанов А.С. Минимальные сепараторы в структурах зависимостей. Свойства и идентификация // Кибернетика и системный анализ. – 2008. – № 6. – С. 17 – 32.
17. Балабанов А.С. Восстановление структур систем вероятностных зависимостей из данных. Аппарат генотипов переменных // Проблемы управления и информатики. – 2003. – № 2. – С. 91 – 99.
18. De Campos L.M., Huete J.F. On the Use of Independence Relationships for Learning Simplified Belief Networks / Tech. Report #DESCAI-960227, Dep. de Ciencias de la Computacion e Inteligencia Artificial, E.T.S.I. Informatica, Univ. de Granada, 1996. – June. – 26 p.
19. Zhang N.L., Poole D. Exploiting Causal Independence in Bayesian Network Inference // Journal of Artificial Intelligence Research. – 1996. – Vol. 5. – P. 301 – 328.
20. Балабанов А.С. Индуктивный метод восстановления монопотоковых вероятностных графовых моделей зависимостей // Проблемы управления и информатики. – 2003. – № 5. – С. 75 – 84.
21. Балабанов О.С. Ефективний метод виявлення структур залежностей в статистичних даних // Материалы 4-й междунар. научно-практ. конф. по програм. "УкрПРОГ'2004". Проблемы программирования. – 2004. – № 2–3. – С. 312 – 319.
22. Балабанов О.С. Дослідження шляхів підвищення обчислювальної ефективності методів ідентифікації моделей залежностей: Звіт з НДР (проміжний), Шифр 3/01К.02-02. – К.: Інститут програмних систем НАН України, 2005. – 36 с.

Стаття надійшла до редакції 04.11.2008