УДК 004.855:519.216

КЛАСИ КАУЗАЛЬНИХ СТРУКТУР, ЩО ІДЕНТИФІКУЮТЬСЯ ТЕСТАМИ ПРОСТОГО ФОРМАТУ

О.С. Балабанов

Тема доповіді стосується теоретичної проблематики відтворення каузальних моделей з даних (без апріорних знань) методами, основаними на незалежності. Розглянуто задачу ідентифікації моделі на основі результатів тестів незалежності виключно 0-го та 1-го рангу (тобто безумовної незалежності та незалежності з однією змінною в умові). Дано визначення класу 1-ідентифікабельних структур моделей. Викладено ідею розпізнавання завершеності відтворення коректної (адекватної) структури моделі, коли не задано апріорних обмежень на структуру. Показано, що підходящим апаратом для розв'язання проблеми 1-ідентифікабельності є апарат локально-мінімальної сепарації. Побудовано декілька підкласів 1-ідентифікабельних моделей; дано структурні обмеження цих підкласів і відповідні критерії завершеності відтворення моделі. Показано приклади структур, які виходять за межі класу 1-ідентифікабельних моделей.

Ключові слова: каузальна мережа, 1-ідентифікабельні моделі, безумовна незалежність, умовна незалежність 1-го рангу, локально-мінімальний сепаратор, колізор, цикл, ланцюг, d-сепарація, полі-ліс, трикутник ребер.

Тема доклада относится к теоретической проблематике восстановления каузальных моделей из данных (без априорных знаний) методами, основанными на независимости. Рассмотрена задача идентификации модели на основе результатов тестов независимости исключительно 0-го и 1-го ранга (т.е. безусловной независимости и независимости с одной переменной в условии). Дано определение класса 1-идентификабельных структур моделей. Изложена идея распознавания завершенности восстановления корректной (адекватной) структуры модели, когда не задано априорных ограничений на структуру. Показано, что подходящим аппаратом для решения проблемы 1-идентификабельности является аппарат локально-минимальной сепарации. Построено несколько подклассов 1-идентификабельных моделей; даны структурные ограничения этих подклассов и соответствующие критерии завершенности восстановления модели. Показаны примеры структур, которые выходят за пределы класса 1-идентификабельных моделей.

Ключевые слова: каузальная сеть, 1-идентификабельные модели, безусловная независимость, условная независимость 1-го ранга, локально-минимальный сепаратор, коллайдер, цикл, цепь, d-сепарация, поли-лес, треугольник ребер.

We tackle some theoretical problems of constraint-based approach to causal network inference from data (without prior restrictions). Our interest is to recover a model structure from independence tests of zero and first rank only. Class of 1-identifiable causal structures is defined. An idea to recognize whether model recovery is successfully completed (i.e. adequate model structure is outputted) is suggested. The framework of locally minimal separation in DAG is shown to be appropriate instrument to tackle the problem. A few subclasses of class of 1-identifiable structures are specified; corresponding structural restrictions and criteria of recovery completeness are given. We present some causal structures which are not 1-identifiable.

Key words: causal network, 1-identifiable model, unconditional independence, conditional independence of 1st rank, locally minimal separator, collider, cycle, chain, d-separation, poly-tree, edge triangle.

Вступ

Каузальна мережа – це модель залежностей між змінними (заданого набору), яка адекватно відображає структуру спрямованих впливів (зазвичай – в умовах неповної спостережуваності). Каузальна мережа описується як пара (G,Θ), де G – граф, що специфікує структуру моделі, Θ – параметри, прив'язані до G, які описують кількісний аспект моделі. В практичних моделях структура не має орієнтованих циклів, тобто оргграф G – ациклонний. У доповіді розглядаються мережі з одно-орієнтованими ребрами, тобто моделі на основі ординарних ациклонних графів (оАОГ-моделі), а також визначаються й аналізуються певні підкласи оАОГ-моделей. Параметри оАОГ-моделі описуються як сукупність локальних параметрів, заданих для кожної змінної. Зокрема, для мережі, що показана на рис. 1, для змінної Y опис може бути заданий у формі $y=\mathrm{f}(q,w,z)+\varepsilon_Y$. (В даному разі використано адитивний гамір ε_Y, але це не обов'язково.) Через неповну спостережуваність об'єкту опис залежностей має ймовірнісний характер. Відомості про каузальні мережі можна отримати з [1–4]. Надалі розглядається виключно структурний аспект каузальних мереж (опис параметрів не використовується). (Альтернативними інтерпретаціями для обраного формалізму є багатозначні залежності в теорії баз даних, апарат умовної незалежності в не-ймовірнісних численнях [5] тощо.)

Мета роботи виникла в рамках проблематики індуктивно-емпіричного виведення каузальних мереж (породжена проблемами відтворення моделей зі статистичних даних).

Зафіксуємо елементарні визначення. Якщо є ребро $X\to Y$, то X є батьком для Y, а Y – дитина для X. За наявності оршляху $X\to\cdots\to Y$ вершина X є пращуром, а Y – нащадком. Кінець ребра $X\to Y$ біля вершини X назвемо «хвіст», а кінець цього ребра біля Y назвемо «вістря». Фрагмент вигляду $X\to Y\leftarrow Z$ називають колізором (коллайдером, collider). Ланцюг – це безколізорний шлях. Цикл – це шлях, де перша й остання вершина тотожні. Циклон – це строго орієнтований цикл. Очевидно, відсутність циклонів тотожна відсутності циклів без колізорів. Ребро, спрямування якого не є суттєвим, позначаємо $X—Y$.

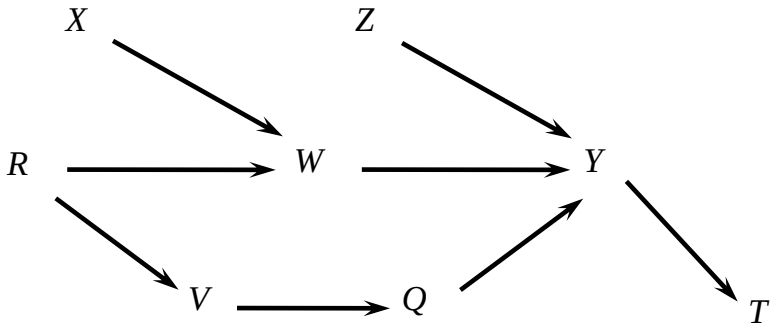

Рис. 1. Приклад структури мережі

«Якісними» властивостями каузальної мережі є її марковські властивості, тобто сукупність фактів умовної незалежності, інваріантних до параметризації моделі. Всі марковські властивості моделі визначаються виключно структурою моделі й можуть бути зчитані з графу G моделі за допомогою критерія d-сепарації [1]. d-залежність (за умови $\mathbf{S}$) означає, що між відповідними вершинами існує відкритий шлях за заданої умови. Факт d-сепарації (d-незалежності) вершин тягне (імплікує) ймовірнісну умовну незалежність однойменних змінних. Далі будемо оперувати переважно фактами безумовної (не)залежності та (не)залежності з однією змінною (вершиною) в умові. Дві вершини є безумовно d-залежні, якщо й тільки якщо між ними є принаймні один ланцюг, інакше вони безумовно d-сепаровані. Три наступні констатації є еквівалентними: 1) змінні X та Z є умовно незалежні за умови на Y; 2) всі шляхи між X та Z є d-блоковані за допомогою вершини Y; 3) вершини X та Z є d-сепаровані за допомогою Y. Останнє позначають як $\mathrm{DS}(X,Z\,|\,Y)$ і кажуть, що Y – це d-сепаратор для пари (X,Z). Наприклад, на рис. 1 вершина W є d-сепаратором для пари (X,Y). Якщо між вершинами X та Z немає жодного ланцюга, то пишуть $\mathrm{DS}(X,Z)$ (безумовна d-сепарація).

Базова проблема і формулювання задачі

Коренева (первинна) проблема, звідки походить обрана задача: відтворити модель (G,Θ) зі статистичних даних (без апріорних знань або із «слабкими» знаннями на вході). Типовою є ситуація, коли не задано обмежень на структуру моделі (за замовчуванням, структура відноситься до ординарних ациклонних орграфів). Кількість можливих структур в класі оАОГ-моделей – супер-експоненційна. Поширений й найбільш популярний підхід до відтворення моделі – оснований на незалежності («сепараційний») [2–4]. Задача розв'язується за етапами: 1) ідентифікація всіх ребер графу G; 2) обґрунтована орієнтація (деяких) ребер; 3) оцінка параметрів Θ для отриманої структури (кількісний опис моделі).

В ході виконання 1-го етапу ключовою процедурою є наступна резолюція: коли є умовна незалежність змінних X та Y за підходящої умови $\mathbf{S}$, робимо висновок, що немає ребра $X - Y$. Отже, проблема зводиться до пошуку (підбору) сепараторів. Пошук сепараторів ведеться у порядку зростання їх складності (починаючи з порожніх, 0-го рангу). Для уникнення надто складного комбінаторного пошуку використовують наступний принцип пошуку гіпотетичних сепараторів. Множину (умову) $\mathbf{S}$ формують як підмножину вершин, суміжних до X, а також як підмножину вершин, суміжних до Y. Але й за такого звуження перебір може бути складним. Ранг тесту незалежності з умовою $\mathbf{S}$ дорівнює кардинальності (потужності) множини $\mathbf{S}$. Алгоритми відтворення моделі стикаються з практичними проблемами, коли кардинальність умов $\mathbf{S}$ зростає, що збільшує ризик помилок тестування умовної незалежності та складність обчислень. Особливо гостро проблема постає у випадку неперервних змінних та невідомих форм залежностей. Доводиться виконувати «важкі» непараметричні тести, які є також ненадійні. Пропонується розв'язати (або радше пом'якшити) проблему за рахунок максимально повного використання результатів тестів малого рангу, а саме – тестів нульового та першого рангів. Тобто бажано відтворити модель, спираючись на факти безумовної незалежності та умовної незалежності з однією змінною в умові.

Вдамося до ідеалізації задачі (оберемо теоретичну постановку). Розглянемо проблему в апараті графів. Будемо оперувати фактами d-залежності (d-сепарації), а не фактами емпіричної залежності (незалежності). При інтерпретації результатів в емпіричному сенсі треба мати на увазі, що приймається припущення, що факт залежності виявляється за будь-якої довжини відкритого шляху (хоча практично залежність через довгий шлях може бути надто слабкою). Втім, ця ідеалізація – не критична (тим більше – коли розглядаються и структур, де кардинальність d-сепараторів не перевищує одиницю).

Отже, наявність принаймні одного ланцюгу між вершинами X та Z означає заперечення безумовної d-сепарації, тобто означає $\neg\mathrm{DS}(X,Z)$. Згідно нашої ідеалізації, факт $\neg\mathrm{DS}(X,Z)$ інтерпретується як

безумовна залежність змінних X та Z, що будемо виражати як $\mathrm{DP}(X,Z)$. («Безумовно залежні» тотожно «зв'язані ланцюгом».) Якщо X та Z зв'язані ланцюгом, але не ребром, то можливо, що на тому ланцюгу є певна вершина Y, що її кондиціонування блокує той ланцюг, і тоді чинне $\mathrm{DS}(X,Z|Y)$. Якщо кондиціонування вершини Q не блокує якийсь ланцюг між X та Z, то маємо $\mathrm{DP}(X,Z|Q)$.

Мета й задача формулюються наступним чином. Нехай виконано всі (актуальні) тести нульового та першого рангів. Які висновки можна зробити з цих результатів? Якщо відомо, що адекватна (генеративна) модель належить до класу, де результати таких тестів є достатні для відтворення генеративної моделі, то треба відтворити модель і повідомити користувачу про завершеність відтворення моделі. Але якщо генеративна модель виходить за межі того класу моделей, то бажано розпізнати цей факт і повідомити користувачу про можливу незавершеність отриманого результату виведення. (Зазвичай в такому випадку виведена структура містить зайві ребра.) Вказана мета й задача потребують відповідного визначення.

Визначення 1. Клас моделей залежності Ψ назвемо *1-ідентифікабельним*, якщо й тільки якщо:

1) будь-які дві нееквівалентні моделі K та M з класу Ψ різняться результатом принаймні одного тесту нульового або першого рангу;

2) у випадку, якщо генеративна (автентична) модель виходить за межі класу Ψ, то цей факт виходу неодмінно можна розпізнати на основі результатів тестів нульового та першого рангу (або на основі конфігурації ребер, утриманих після таких тестів).

Визначення 2. Клас моделей залежності Ω назвемо 1-відтворюваним, якщо й тільки якщо будь-які дві нееквівалентні моделі K та M з класу Ω різняться результатом принаймні одного тесту нульового або першого рангу.

До визначення 2 доцільно надати коментар (тлумачення): якщо на вході методу виведення моделі задано (як апріорне знання), що генеративна (автентична) модель належить до класу Ω, то коректний метод (гарантовано) відтворить генеративну модель (з точністю до класу еквівалентності [1, 2, 4]) з сумісного розподілення змінних, використовуючи результати тестів 0-го та 1-го рангу. Проте факт належності моделі до класу Ω часто невідомий. Інтуїтивно ясно, що клас Ω є ширшим за клас Ψ.

Обрана задача має наступну прагматику. Виведення моделі може зупинитися після тестів першого рангу за таких обставин:

1) вимушено (через формат даних);

2) «зумисне», волюнтаристськи, за директивою аналітика (з огляду на…);

3) внаслідок обґрунтованого висновку, що виведення моделі завершене.

Коректний метод (алгоритм) мусить: або 1) вивести точну модель (її клас еквівалентності); або 2) виявити і повідомити, що точну модель неможливо вивести за результатами тестів 0-го та 1-го рангу. Коректний алгоритм, який задовольняє вказаним вимогам, можна побудувати, якщо генеративна модель належить до класу Ψ. Визначені поняття дозволяють показати можливість побудови методів з бажаними властивостями (без розгляду деталей методів).

Наскільки відомо автору, вказана задача досі не ставилася. Факт 1-ідентифікабельності для (під)класів каузальних мереж дотепер не було встановлено. Клас 1-ідентифікабельних моделей не було побудовано. Відомим дотичним результатом є наступний. Давно встановлено, що тестами 1-го рангу відтворюються дерева (ліси), але встановлено тільки факт 1-відтворюваності лісів. Алгоритм, який по-суті вирішує цю задачу – алгоритм Chow&Liu [6], – видає апроксимацію довільного сумісного розподілення на основі дерева, найкращу з точки зору суми взаємної інформації всіх пар змінних. Цей алгоритм легко поширюється на клас полі-лісів [4, 7, 8]. У строгому сенсі факт 1-ідентифікабельності було встановлено тільки для полі-лісів (проте декларативного критерія завершеності в явному вигляді не було сформульовано). Ширші підкласи 1-ідентифікабельних каузальних мереж дотепер не ідентифіковано.

Ідея розв'язання задачі. Найпростіший підклас 1-ідентифікабельних моделей

Як можна розпізнати завершеність виведення коректної (адекватної) структури моделі (не знаючи її клас)? В рамках підходу, основаного на незалежності, етап ідентифікації всіх ребер моделі завершується, коли вичерпано всі можливості спростування «утриманих» (гіпотетичних) ребер. Отже, для розпізнавання завершеності виведення (першого етапу) треба знайти ефективні засоби характеристики множини можливих сепараторів й розпізнавання вичерпаності пошуку сепараторів. Підходящий теоретичний апарат для аналізу 1-ідентифікабельності – апарат локально-мінімальної сепарації в оАОГ-моделях [4, 9, 10]. Було встановлено необхідні вимоги до кожного члена локально-мінімального сепаратора (ЛоМС). Доведено низку резолюцій (правил) для формування множини кандидатів у члени ЛоМС. Всі опубліковані резолюції формування ЛоМС використовують результати тестів нульового та 1-го рангу. Нагадаємо найбільш важливі (для нашої задачі) результати.

Перше. Встановлено необхідність «стрижня» сепаратора. Доведено, що до складу кожного не порожнього ЛоМС для пари вершин (X,Y) входить, як мінімум, одна вершина Z , така, що чинне $\mathrm{DP}(X,Z)$, $\mathrm{DP}(Y,Z)$, $\mathrm{DP}(X,Z|Y)$ та $\mathrm{DP}(Y,Z|X)$.

Друге. Доведено резолюцію «відсторонення»: якщо в оАОГ-моделі вірне $\mathrm{DS}(X,Z|Y)$, то вершина Z не є членом жодного ЛоМС для пари вершин (X,Y).

Зрозуміло, якщо (після тестів 1-го рангу й застосування набору резолюцій) для «утриманого» ребра залишається тільки один або жодного кандидата в члени сепаратора, то це ребро – «остаточне» (автентичне), тобто в ході подальшого виведення моделі це ребро не буде видалено. Це дає «процедурний» критерій завершеності виведення моделі. Але бажано мати явний компактний критерій 1-ідентифікабельності моделі. Також бажано дати компактний конструктивний опис класу 1-ідентифікабельних моделей та його підкласів.

Необхідність стрижня сепаратора підказує, що шуканий критерій треба формулювати в термінах сепарацій в межах кожної кліки залежності. Кліка залежності – це максимальна (по включенню) множина вершин, де всі вершини попарно безумовно залежні.

Будемо послідовно узагальнювати критерій. Будемо поступово розширювати підкласи 1-ідентифікабельності, уточнюючи критерій та послаблюючи структурні обмеження (SR) моделі.

Певно, найпростіший критерій для 1-ідентифікабельності – критерій «тотальна розв'язуваність у трійках» (позначимо його «Cr1»). Критерій «Cr1»: в кожній трійці взаємозалежних вершин X,Y,Z чинна одна з трьох сепарації: $\mathrm{DS}(X,Y|Z)$, $\mathrm{DS}(X,Z|Y)$ або $\mathrm{DS}(Y,Z|X)$. Підклас 1-ідентифікабельних структур з таким критерієм – «прості полі-ланцюги». Приклад структури в цьому підкласі зображено на рис. 2. Цей підклас визначається наступними структурними обмеженнями:

«SR-1»: кожний цикл має не менше 4-х колізорів.

«SR-2»: якщо із спільною вершиною Q контактують три ребра чи більше, то з-поміж всіх тих ребер тільки одне може контактувати із Q «хвостом».

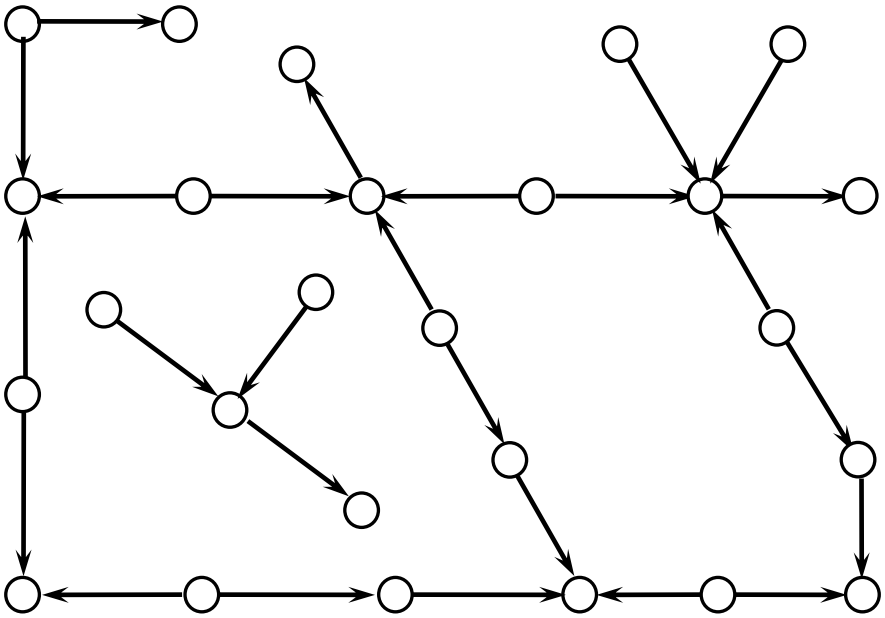

Рис. 2. Приклад мережі в підкласі «прості полі-ланцюги»

Можна незначною мірою модифікувати критерій й розширити підклас моделей так, щоб дозволити трикутники ребер. Оновлений критерій формулюється як «часткова розв'язуваність у трійках». Критерій «Cr2»: в кожній кліці залежностей із 4-х чи більше вершин для кожної трійки X,Y,Z чинна одна з трьох сепарації: $\mathrm{DS}(X,Y|Z)$, $\mathrm{DS}(X,Z|Y)$ або $\mathrm{DS}(Y,Z|X)$. Якщо в кліці із трьох вершин X,Y,Z не чинна жодна з трьох сепарації $\mathrm{DS}(X,Y|Z)$, $\mathrm{DS}(X,Z|Y)$, $\mathrm{DS}(Y,Z|X)$, то не існує жодної «четвертої» вершини Q, яка залежить від двох вершин з трійки X,Y,Z.

Відповідний підклас 1-ідентифікабельних структур з таким критерієм – «полі-ланцюги із трикутниками». Структурні обмеження для цього підкласу трохи відрізняються від вищенаведених і формулюються наступним чином.

«SR-1a»: кожний цикл або є трикутником, або має не менше 4-х колізорів.

«SR-2a»: якщо із спільною вершиною Q контактують три ребра чи більше, причому жодне з тих ребер не входить до трикутника, то тільки одне з тих ребер може контактувати із Q «хвостом»; якщо вершина Q входить до якогось трикутника і контактує з трьома чи більше ребрами, то всі ребра, дотичні до Q, контактують з Q «стрілками».

З обмеження «SR-1a» випливає, що кожний цикл має або три ребра (трикутник), або не менше 8 ребер. З обмеження «SR-2a» випливає: якщо певна вершина трикутника контактує з якоюсь вершиною поза цим трикутником, то інші вершини цього трикутника не можуть контактувати з жодною вершиною поза цим трикутником. Структурне обмеження «SR-2a» ілюструється прикладами, поданими на рис. 3. Ребро, яке контактує «хвостом» з вершиною трикутника, заборонене (рис. 3, (ii)). Двом ребрам заборонено контактувати «хвостами» з вершиною, дотичною до третього ребра (рис. 3, (iii)). Легітимні в цьому підкласі конструкції з трикутників зображені на рис. 3, (iv) та рис. 3, (v). Для останніх двох моделей встановлення факту завершеності відтворення потребує правила «чужого гена» або Факту 10 з [9], а також Факту 6 з [9] або Пропозиції 3 з [10]. (Названі правила зібрано й редаговано в [4].) Щойно визначені підкласи охоплюють невеличку підмножину 1-ідентифікабельних моделей. Вони навіть не включають клас лісів (хоча й виходять за межі лісів).

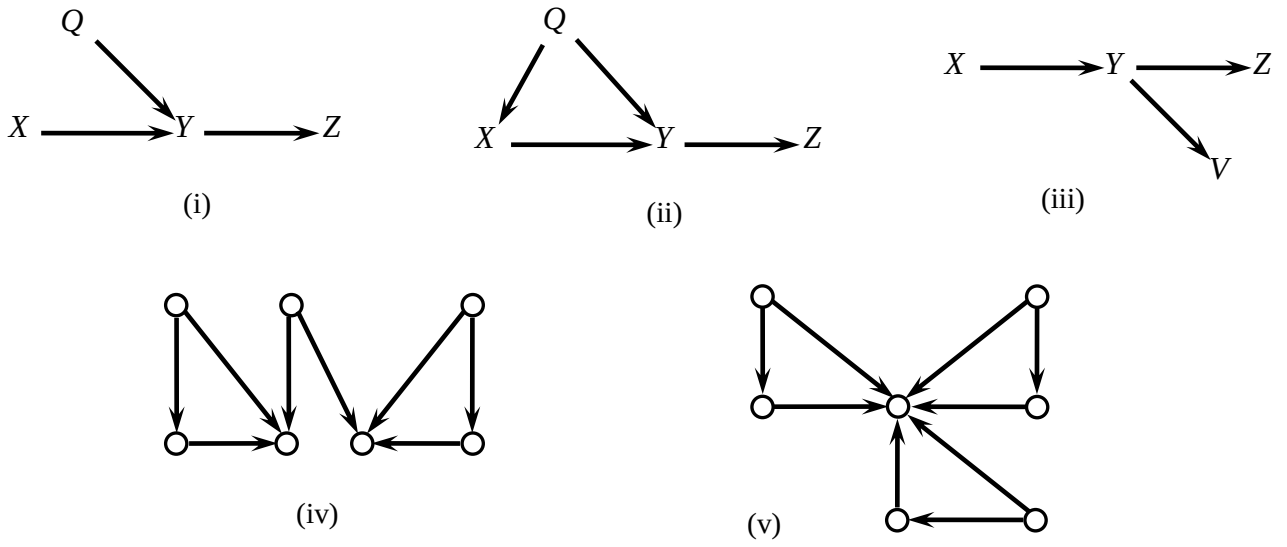

Рис. 3. Ілюстрація обмежень для підкласу «полі-ланцюги із трикутниками»:
(i) – легітимне з'єднання вершин; (ii) та (iii) – не легітимне з'єднання;
(iv) – легітимне з'єднання 2-х трикутників; (v) – легітимне з'єднання 3-х трикутників

Узагальнення критерія та уточнення меж 1-ідентифікабельних моделей

Зрозуміло, що можна розширити визначений вище підклас 1-ідентифікабельних моделей за рахунок включення всіх полі-лісів (тобто дозволити розгалуження оршляхів, як у дереві). Водночас доцільно дозволити трикутникам контактувати з «іншими» ребрами з будь-якого кінця (тобто дозволити ребру контактувати «хвостом» з вершиною трикутника). За цією логікою, заразом також доцільно дозволити цикли з трьома колізорами. Отже, декларуємо підклас 1-ідентифікабельних моделей «полі-ліси з трикутниками». Оскільки три взаємозалежні вершини дерева можуть розташовуватися на різних гілках дерева, для таких трьох вершин треба шукати сепаратор серед «четвертих» вершин. Відтак, критерій 1-ідентифікабельності для цього підкласу формулюється наступним чином.

Критерій «Cr3»: якщо для трійки взаємозалежних вершин X,Y,Z не чинна жодна з трьох сепарації: $\mathrm{DS}(X,Y|Z)$, $\mathrm{DS}(X,Z|Y)$ або $\mathrm{DS}(Y,Z|X)$, то маємо або 1) для кожної пари вершин з цієї трійки існує «четверта» вершина, яка є сепаратором для цієї пари, або 2) для кожної вершини Q, залежної від трьох X,Y,Z, виконується $\mathrm{DS}(Q,Y|X)\,\&\,\mathrm{DS}(Q,Z|X)$, або $\mathrm{DS}(Q,X|Y)\,\&\,\mathrm{DS}(Q,Z|Y)$, або $\mathrm{DS}(Q,X|Z)\,\&\,\mathrm{DS}(Q,Y|Z)$. Далі, якщо для трійки взаємозалежних вершин X,Y,Z не чинна жодна з трьох сепарації: $\mathrm{DS}(X,Y|Z)$, $\mathrm{DS}(X,Z|Y)$ або $\mathrm{DS}(Y,Z|X)$, і якщо для жодної пари вершин з цієї трійки не існує сепаратора, то не існує жодної вершини W, залежної точно від двох вершин з трьох X,Y,Z.

Прокоментуємо дозволені та заборонені в цьому підкласі типи конкатенації фрагментів, згідно критерія «Cr3». Легітимні конкатенації трикутника з ребром відображено на рис. 4, (i), (ii), (iii). На рис. 4, (ii)

ребра трикутника можуть бути орієнтовані довільно, але так, щоб не утворювати циклон. Заборонена конструкція показана на рис. 4, (iv). Тут вершина Z залежна точно від двох вершин Q,Y трикутника. (Така структура відома як модель інструментальної змінної.) Для сепарації Z та Y необхідний тест 2-го рангу.

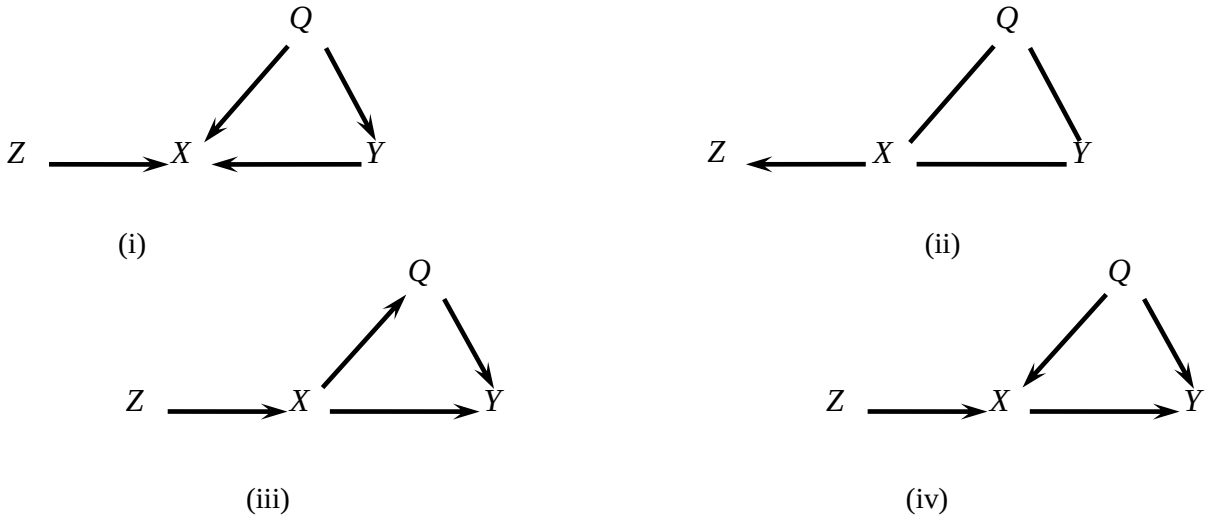

Рис. 4. Ілюстрація обмежень на під'єднання трикутників у підкласі «полі-ліси з трикутниками»: (i), (ii), (iii) – легітимне під'єднання; (iv) – не легітимне

Сформулюємо структурні обмеження для класу «полі-ліси з трикутниками». Обмеження «SR-1a» та «SR-2a» замінюються іншими, а саме.

«SR-3»: кожний цикл або є трикутником, або має не менше 3-х колізорів.

«SR-4»: якщо у трикутнику одне ребро контактує з вершиною X «вістрям», а інше – «хвостом», то жодне ребро поза цим трикутником не може контактувати з вершиною X «вістрям».

З обмеження «SR-3» випливає, що кожний цикл має або три ребра (трикутник), або не менше 6 ребер. Два трикутники не можуть мати спільного ребра, бо тоді виник би цикл з 4-х ребер.

В цьому підкласі 1-ідентифікабельних моделей довільний набір ребер може бути дотичним до однієї вершини і «вістрями», і «хвостами» (якщо не порушується обмеження «SR-4»). Завдяки розширеним можливостям конкатенації трикутників в цьому підкласі трикутники можна з'єднувати у послідовність («квазі-ланцюг»). На рис. 5 показано приклад структури в класі «полі-ліси з трикутниками».

Звісно, клас 1-ідентифікабельних моделей є ширшим за «полі-ліси з трикутниками». Наведемо деякі відомості, що допоможуть провести межу класу 1-ідентифікабельних моделей.

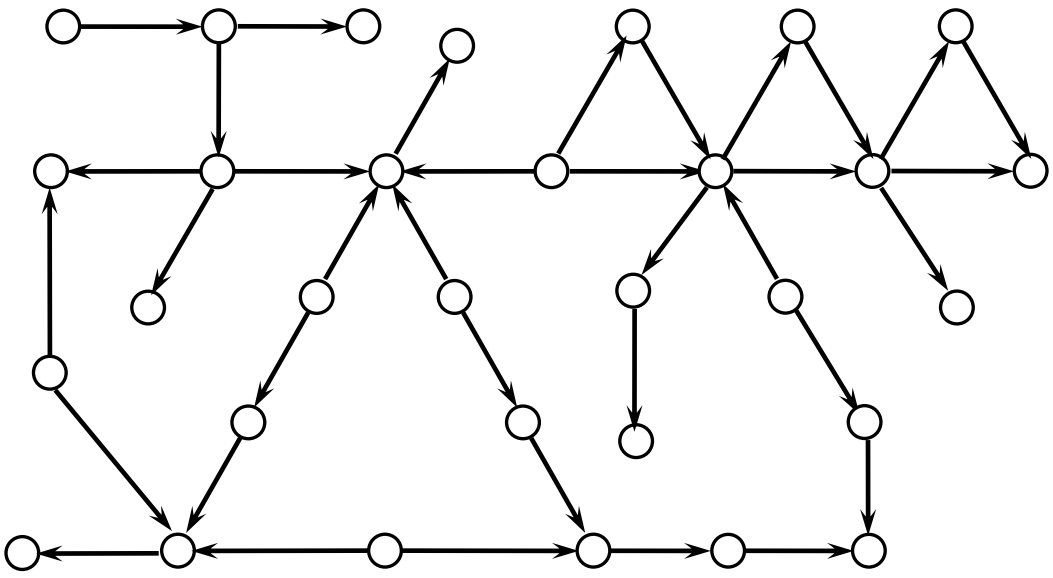

Рис. 5. Приклад мережі в підкласі ««полі-ліси з трикутниками»

Проміжне місце між полі-лісами та загальним випадком оАОГ посідає клас так званих монопотокових графів залежностей (МПГЗ) [4, 7, 8, 11]. Моделі з класу МПГЗ можуть мати цикли, але кожний цикл містить не менше двох колізорів. (Трикутники заборонені в МПГЗ.) Було показано, що для відтворення монопотокових моделей достатньо спиратися на тести 0-го та 1-го рангу. Мабуть, найбільш ефективним алгоритмом відтворення моделей в класі МПГЗ є «Proliferator-D» [4, 11]. Але цей алгоритм (й інші спеціалізовані методи) спирається на знання, що генеративна модель належить класу МПГЗ. Клас МПГЗ не входить до класу 1-ідентифікабельних моделей. Водночас багато моделей, що містять цикли з двома колізорами, входять до класу 1-ідентифікабельних моделей. Розглянемо ситуації, показані на рис. 6. Для всіх варіантів на даному рисунку (включаючи й ті, що містять ребро $Q — Y$, й ті, що не містять його) для набору вершин Q,X,Y,Z маємо одну сепарацію нульового рангу ($\mathrm{DS}(X,Z)$) і жодної сепарації 1-го рангу. Якщо ребро $Q — Y$ відсутнє, структури рис. 6, (i), (ii) попадають в клас монопотокових моделей. Обидва варіанти структури для рис. 6, (iii) містять одноколізорний цикл. Вищевказаний набір фактів сепарації узгоджується і з присутністю ребра $Q — Y$, і з його відсутністю. Додання двох ребер й вершин, як показано на рис. 6, (i), не змінює ситуацію, тобто для вирішення питання щодо ребро $Q — Y$ потрібен тест 2-го рангу. Натомість за наявності двох додаткових ребер іншої орієнтації $R \to Q$ та $Y \leftarrow W$ (рис. 6, (ii)) ситуація суттєво змінюється. Нехай ребро $Q — Y$ відсутнє в моделі, але це невідомо методу. Така модель вичерпно ідентифікується тестами нульового рангу. Дійсно, якби існувало ребро $Q \to Y$, то Y залежала би від R. Якби існувало ребро $Q \leftarrow Y$, то Q залежала би від W. Але маємо факти $\mathrm{DS}(Y,R)$ та $\mathrm{DS}(Q,W)$. Отже, факти свідчать, що ребро $Q — Y$ неможливе. Питання щодо цього ребра (коли його в дійсності немає) розв'язується непрямо, через логічний аналіз сукупності результатів тестів 0-го рангу. Така резолюція сформульована як правило не-поглинання (твердження 1) в [9]. (В інших ситуаціях відсутність ребра можна з'ясувати за подібним принципом, використовуючи правило квазіінструментальної пари (пропозиція 9) з [10].)

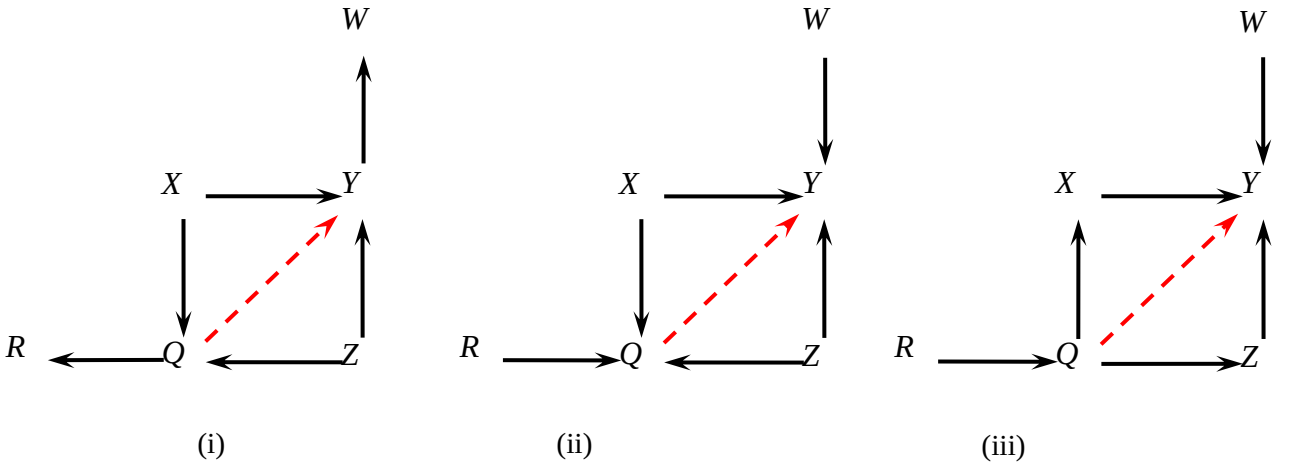

Рис. 6. Варіанти структур із 2-колізорними та 1-колізорними циклами:
(i) та (iii) – моделі, що не є 1-ідентифікабельні;
(ii) – 1-ідентифікабельна модель (без ребра $Q — Y$)

Натомість моделі, що містять цикл з чотирьох чи більше ребер, з одним колізором, не є 1-ідентифікабельними. Наприклад, в моделі, яка зображена на рис. 6, (iii), для розв'язання невизначеності щодо ребра $Q — Y$ необхідно виконати тест 2-го рангу. (За будь-якого «контексту» неможливо створити правило, яке би вирішило питання щодо такого ребра $Q — Y$ на підставі результатів тестів 0-го та 1-го рангів. Втім, це не означає, що задачу не можна вирішити в інший спосіб, аналізуючи розподілення ймовірностей змінних.)

Нарешті, покажемо, що існує клас 1-відтворюваних моделей, відмінний від класу 1-ідентифікабельних моделей. Підкласом класу Ω є клас монопотокових моделей. Повернемося до моделі, відображеної на рис. 6, (i), коли ребра $Q — Y$ немає. Для зручності обговорення нехай також немає вершин R та W. Така модель попадає в клас монопотокових. Як вказано вище, для четвірки вершин Q,X,Y,Z маємо $\mathrm{DS}(X,Z)$ і жодної сепарації 1-го рангу. На відміну від розглянутої вище ситуації, нехай тепер на вході методу виведення моделі задано знання, що генеративна модель входить до класу монопотокових. Відтак, трикутники ребер заборонені. Множина можливих варіантів моделі складається з наступних структур: ланцюги; прості полі-ланцюги; ліси; полі-ліси; цикл з двома колізорами. Варіанти з ізольованою вершиною відмітаються (спростовуються), бо породжують три сепарації 0-го рагу. Ланцюги та дерева, що покривають 4 вершини,

відмітаються, бо суперечать факту $\mathrm{DS}(X,Z)$. (Для побудови полі-лісу, який не є полі-ланцюгом, потрібно не менше п'яти вершин.) Полі-ланцюг, що складається з $X \to Q \leftarrow Z$ та $Q \leftarrow Y$, характеризується сепарацією $\mathrm{DS}(X,Y)$ та $\mathrm{DS}(Z,Y)$, що суперечить фактам. Полі-ланцюг, що складається з $X \to Q \leftarrow Z$ та $Q \to Y$, також відмітається, тому що тягне $\mathrm{DS}(X,Y \mid Q)$ та $\mathrm{DS}(Z,Y \mid Q)$, що суперечить фактам. Симетричні варіанти спростовуються аналогічно. Отже, залишається єдина модель, що узгоджується із заданим набором фактів – цикл з двома колізорами (зрозуміло, без ребра $Q — Y$). Таким чином, вказана структура є 1-відтворюваною, але не є 1-ідентифікабельною. Необхідно зауважити, що у наведених викладках було використано знання, що генеративна модель належить спеціальному підкласу моделей з класу Ω, з відомим конструктивним обмеженням.

Багато практичних задач потребують моделей, які описують зворотні зв'язки. Зокрема, це потрібно для моделювання багатокрокових процесів. Наприклад, в економетриці моделюються багатовимірні процеси з автокореляційними зв'язками і «регулярними» зворотними зв'язками. Каузальна мережа відображає зворотний зв'язок за допомогою одноколізорного циклу. Найпростіший варіант одноколізорного циклу втілюється трикутником ребер. Але для моделювання зворотних зв'язків в економетриці навряд чи буде достатньо використати послідовність («квазі-ланцюг») трикутників. Припустимо, вдалося описати зворотний зв'язок в такій моделі за допомогою трикутника ребер (і то завдяки тому, що вимірювання змінних в рамках одного кроку зсунуті у часі так, що описують залежність між субпроцесами через ребро трикутника $U \to V$). Інше ребро трикутника ($U \to U^*$) описує послідовні стани субпроцесу U. Але послідовні стани субпроцесу V теж зазвичай поєднані автокореляційними зв'язками, відтак, трикутники поєднуються у цикли з 4-х ребер і більше. Отже, в таких моделях необхідно використовувати одноколізорні цикли з 4-х та більше ребер. Такі структури не є 1-ідентифікабельними. Зрозуміло, клас 1-ідентифікабельних моделей не може задовольнити всі практичні потреби. Та в будь-якому разі бажано відтворити якомога більшу частину моделі простими тестами й локалізувати ділянки невизначеності, де необхідні складні тести.

Відзначимо деякі проміжні підсумки. Можна сказати, що наведені результати практично наблизили розуміння локалізації меж класу 1-ідентифікабельних моделей. Можна передбачати, що подальша (більш повна) характеристика класу 1-ідентифікабельних моделей і точне визначення його меж та критеріїв призведуть до некомпактних формулювань. (По-суті, доведеться вводити в тіло критерія формулювання певних правил локально-мінімальної сепарації.) Цікаво, що визначення підкласів моделей, виведення яких гарантовано завершується на етапі тестів 1-го рангу, породило нестандартні класи структур орграфів. Отримані результати відкривають перспективи подальших досліджень і узагальнень. Наприклад, можна визначити й аналізувати клас 2-ідентифікабельних моделей.

Висновки

Відтворення каузальних мереж зі статистичних даних є складною проблемою, особливо коли на вході методу майже нічого невідомо про структуру моделі. Практика вимагає зниження обчислювальних витрат та ризику помилок в ході відтворення моделі. Для досягнення такої мети (в рамках методів, основаних на незалежності) потрібно зменшувати кількість виконаних тестів та задовольнятися результатами тестів якомога простішого формату. Виведення каузальної моделі може бути припинено після тестів певного рангу вимушено (через формат даних) або за директивою аналітика (обґрунтованого певними міркуваннями). Критичним питанням в такій ситуації стає розпізнавання завершеності (не завершеності) відтворення моделі після виконання тестів заданого рангу. В доповіді показана принципова можливість побудови коректних методів й алгоритмів відтворення моделей з заданими властивостями на основі тестів нульового та першого рангів. Можливість побудови бажаних методів показана через поняття класу 1-ідентифікабельних моделей. Дано визначення класу 1-ідентифікабельних моделей та класу 1-відтворюваних моделей. Зіставлення цих двох понять розкриває «невловиму» (елюзивну), але критично важливу відмінність двох типів проблемної ситуації та двох типів методів відтворення моделей з даних.

Показано, які моделі можна вичерпно відтворити, спираючись на факти безумовної незалежності та умовної незалежності з однією змінною в умові (без апріорних знань). Описано доволі широкий (потужний) підклас класу 1-ідентифікабельних моделей – «полі-ліси з трикутниками», а також простіші підкласи. Полі-ліси з трикутниками охоплюють всі полі-ліси, включають багато моделей з циклами й навіть покривають певну підмножину моделей за межами класу монопотокових графів. Повна й точна характеристика класу 1-ідентифікабельних каузальних моделей – складна проблема. Не входять до класу 1-ідентифікабельних моделей всі структури, які мають цикл з одним колізором і чотирма чи більше ребрами, а також багато структур, які мають цикл з двома колізорами.

Література

1. Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge Univ. Press, 2000. 526 p.
2. Spirtes P. Introduction to causal inference. *J. of Machine Learning Research*. 2010. Vol. 11. P. 1643−1662.

3. Балабанов О.С. Відкриття знань в даних та каузальні моделі в аналітичних інформаційних технологіях. *Проблеми програмування*. 2017. № 3. С. 76–92.
4. Балабанов О.С. Каузальні мережі: аналіз, синтез та виведення з статистичних даних: Автореферат дис. … доктора фіз.-мат. наук (01.05.01– теоретичні основи інформатики та кібернетики). К.: Ін-т кібернетики ім. В.М. Глушкова НАНУ, 2014. 34 с.
5. Studeny M. Formal properties of conditional independence in different calculi of AI. In: Symbolic and Quantitative Approaches to Reasoning and Uncertainty (eds. M. Clarke, R. Kruse, S. Moral), Springer-Verlag, Berlin-Heidelberg, 1993. P. 341−348.
6. Chow C.K., Liu C.K. Approximating discrete probability distributions with dependence trees. IEEE trans. on Information Theory. 1968. Vol. 14. N 3. P. 462–467.
7. Балабанов О.С. Прискорення алгоритмів відтворення баєсових мереж. Адаптація до структур без циклів. *Проблеми програмування.* 2011. № 1. С. 63–69.
8. Балабанов О.С. Принципи та аналітичні засоби реконструкції структур ймовірнісних залежностей у спеціальному класі. *Проблеми програмування*. 2017. № 1. С. 97–110.
9. Балабанов А.С. Минимальные сепараторы в структурах зависимостей. Свойства и идентификация. *Кибернетика и системный анализ*. 2008. № 6. С. 17–32.
10. Балабанов А.С. Формирование минимальных d-сепараторов в системе зависимостей. *Кибернетика и системный анализ*. 2009. № 5. С. 38–50.
11. Балабанов А.С. Реконструкция модели вероятностных зависимостей по статистическим данным. Инструментарий и алгоритм. *Проблемы управления и информатики*. 2009. № 6. С. 90–103.

References

1. Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge Univ. Press, 2000. 526 p.
2. Spirtes P. Introduction to causal inference. J. of Machine Learning Research. 2010. Vol. 11. P. 1643−1662.
3. Balabanov O. S. (2017). Knowledge discovery in data and causal models in analytical informatics. Problems in Programming. (3), 96−112. [in Ukrainian]
4. 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]
5. Studeny M. (1993). Formal properties of conditional independence in different calculi of AI. In: Symbolic and Quantitative Approaches to Reasoning and Uncertainty (eds. M. Clarke, R. Kruse, S. Moral), Springer-Verlag, 341−348.
6. Chow C.K., Liu C.K. (1968). Approximating discrete probability distributions with dependence trees. IEEE trans. on Inform. Theory. 14, (3), 462–467.
7. Balabanov O.S. (2011). Accelerating algorithms for Bayesian networks recovery. Adaptation to structures without cycles. Problems in programming. (1), 63–69. [In Ukrainian]
8. Balabanov O.S. (2017). Principles and analytical tools for reconstruction of probabilistic dependency structures in special class. Problems in programming. (1), 97–110. [in Ukrainian]
9. Balabanov A.S. (2008). Minimal separators in dependency structures: Properties and identification. Cybernetics and Systems Analysis. 44, (6), 803–815.
10. Balabanov A.S. (2009). Construction of minimal d-separators in a dependency system. Cybernetics and Systems Analysis. 45, (5), 703–713.
11. Balabanov A.S. (2009). Reconstruction of the model of probabilistic dependences by statistical data. Tools and algorithm. Journal of Automation and Information Sciences. 41, (12), 32–46. (ISSN 1064-2315).

Про автора:

Балабанов Олександр Степанович,
доктор фізико-математичних наук,
провідний науковий співробітник Інституту програмних систем НАН України.
Кількість наукових публікацій в українських виданнях – 52.
Кількість наукових публікацій в зарубіжних виданнях – 10.
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