ЕФЕКТИВНИЙ МЕТОД ВИЯВЛЕННЯ СТРУКТУР ЗАЛЕЖНОСТЕЙ В СТАТИСТИЧНИХ ДАНИХ

О.С.Балабанов

Інститут програмних систем НАН України,
03187 м. Київ, пр-кт академіка Глушкова,40,
факс: (380)(44)266-62-63, телефон (44)266-62-49, 266-51-39
E-mail: bas@isofts.kiev.ua

Розглянуто проблеми індуктивного виводу (відтворення) структур моделей ймовірнісних залежностей в класі ациклічних орієнтованих графів та в підкласі монопотокових моделей (де кожний цикл має два або більше колайдерів). Досліджено властивості монопотокових моделей. Розроблено метод “Proliferator-C” (узагальнений і вдосконалений варіант метода Chow&Liu), який відтворює структуру монопотокової моделі, спираючись на знання колайдерних змінних та тести умовної незалежності першого порядку, та алгоритм ‘Collifinder’, який ідентифікує всі колайдерні змінні. Порівняно з відомими методами “Proliferator-C” є менш критичним до розміру відборки даних, а за складністю – близький до відомих алгоритмів для лісів (дерев) залежностей.

Problems of recovery of probabilistic graphical model structures in class of acyclic directed graphs (DAG) and their subclass of ‘mono-streams’ models (i.e. digraphs with restriction that each cycle have two or more colliders) are considered. Properties of ‘mono-streams’ models are examined. The method for learning structure of any ‘mono-streams’ model from statistical data are developed. The method is more reliable and robust to sample size then known methods while its complexity is comparable with that for tree-like dependency model recovery algorithms. The method consist of the algorithm ‘Collifinder’ for identification all colliders and method “Proliferator-C”, an extended version of well-known Chow&Liu method.

1. Вступ

Ймовірнісні моделі систем залежностей на основі графів – актуальна тема сучасних досліджень на стику багатомірного статистичного аналіза, теорії графів, теорії інформації і штучного інтелекта. Ймовірнісні графові моделі залежностей відіграють роль точної і строгої мови репрезентації знань з невизначеністю і компактного запису систем залежностей [1-8]. Це – ефективний апарат описа задач штучного інтелекта (зокрема, в експертних системах нового покоління) та розв'язання аналітичних задач у різноманітних предметних галузях. Ймовірнісні графові моделі залежностей приваблюють тим, що їх можна ідентифікувати індуктивно, на основі статистичних даних. Тому такі моделі задіяно в методології відкриття знань у базах даних [4,6,7]. Найбільшої уваги зажили моделі на базі ациклічних орієнтованих графів (АОГ-моделі). Достоїнства АОГ-моделей становлять наочність, компактність, здатність відображати причинно-наслідкові зв'язки, обчислювальна ефективність ймовірнісного висновка від свідчень [1,2,3,6,9,10]. Ці властивості забезпечують ефективне розв'язання задач медичної і технічної діагностики, розпізнавання мови, прогнозування наслідків рішень і дій людини, робота, програмного агента і т.д. Найбільшого поширення набули два види АОГ-моделей: 1)моделі з номінальними змінними, тобто баєсівські мережі чи тенета; 2)лінійні моделі з неперервними змінними та нормальними дистурбаціями, тобто гауссові мережі.

Будувати модель для потреб реальних досліджень, грунтуючись на експертних думках і суб'єктивних уявленнях, як правило є неприйнятним. В більшості практичних ситуацій істинна модель є апріорі невідома. Отож, актуальна задача – ідентифікація моделі “об'єктивними” методами. Як правило, об'єкт (система), для описа якого потрібна модель, реально існує, і його можна спостерігати. Або принаймні ми маємо задовільний імітатор цього об'єкта. В таких ситуаціях модель в принципі можна ідентифікувати “індуктивно” з даних вимірювань (спираючись на кілька методологічних постулатів). Але часто доводиться спиратися на дані пасивних спостережень за об'єктом (системою) через те, що активні експерименти з об'єктом є неможливі або невиправдані практично, економічно, морально і т.д. Така ситуація є характерна, наприклад, для соціально-економічних і медико-біологічних досліджень, астрономії, а подекуди – і для інженерно-технічних проблем. Крім того, в таку ситуацію ми поставлені в технологіях відпрацювання даних (data mining). Ідентифікація ймовірнісних графових моделей на основі статистичних даних спостережень, коли на початку дослідження майже нічого не відомо про структуру моделі – надзвичайно трудомістка задача. Відомі методи здатні надійно розв'язувати такі задачі лише за вимушених нереалістичних обмежень чи за наявності відборки даних дуже великого обсягу.

2. Базові поняття та визначення АОГ-моделей

З огляду на взаємо-однозначну відповідність між змінними та вершинами графа моделі будемо вживати терміни з*мінна* та *вершина* графа взаємозамінно як еквіваленти. Якщо в графі існує дуга ($X \rightarrow Y$), то змінну X називають "батьком" змінної Y, а Y – дитиною X. Дугу називають ребром, коли не береться до уваги його орієнтація. Корінь орграфа – це вершина, до якої не надходить жодної дуги. Орпуть або оршлях (тобто строго орієнтований шлях) – це шлях, на якому всі ребра орієнтовані в напрямку одного й того самого кінця шляха. Вершину X називають предком вершини Y в орграфі, якщо існує орпуть від X до Y. Будемо позначати через $F(X)$ множину батьків змінної X, через $Anc(X)$ – множину предків змінної X.

АОГ-модель визначається як орієнтований ациклічний граф (де кожній вершині графу відповідає змінна) з приписаними параметрами. У баєсівських мережах параметри задаються як локальні умовні розподілення ймовірностей значень змінних $P(X_i \mid F(X_i))$. А в гауссових мережах – як коефіцієнти лінійних рівнянь (для ребер) та дисперсії відхилень (для вершин).

Визначення 1. **Колайдером** в орграфі називається фрагмент із двох суміжних дуг вида $X \rightarrow Y \leftarrow Z$, з тим, що вершини X та Z – не суміжні в орграфі. Якщо колайдер $X \rightarrow Y \leftarrow Z$ є частиною шляху π в орграфі, то Y називається **колайдерною змінною на шляху** π. (Водночас змінна Y може бути неколайдерною на деякому іншому шляху.) **Безколайдерний шлях** в орграфі – це шлях, що не містить жодного колайдера.

Відношення між структурою моделі і фактами умовної незалежності, відображеними в моделі, формалізовані через критерій *d-сепарації* [1,2,5].

Визначення 2 [2]. Шлях π в АОГ-моделі називають *d-закритим* (*d-блокованим*) за допомогою множини вершин $\boldsymbol{S}$, якщо і тільки якщо

(1) на шляху π лежить принаймні одна неколайдерна змінна, приналежна $\boldsymbol{S}$, або

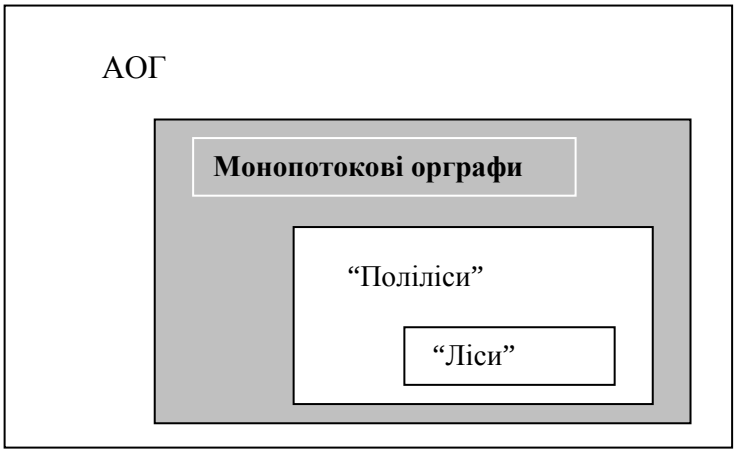
(2) існує колайдерна змінна C, $C \in \pi$, $C \notin \boldsymbol{S}$, і з тим ніякий нащадок змінної C не належить до $\boldsymbol{S}$, тобто $Anc(S) \cap C = \varnothing$.

Множина вершин $\boldsymbol{S}$ ***d-сепарує*** вершини X та Y якщо і тільки якщо всі шляхи між X та Y d-закриті за допомогою множини вершин $\boldsymbol{S}$.

Будемо позначати таку d-сепарацію термом $Ds(X \perp \boldsymbol{S} \perp Y)$, або скорочено ($X \perp \boldsymbol{S} \perp Y$). Множина змінних $\boldsymbol{S}$ - сепаратор для пари змінних X, Y ($X, Y \notin \boldsymbol{S}$).

З факту d-сепарації слідує відповідна імовірнісна умовна незалежність [10].

Спеціальними випадками або підкласами АОГ-моделей є монопотокові моделі, "поліліси" та ліси (дерева). Підкласи АОГ-моделей визначаються за "топологічними" обмеженнями на структуру графів моделі. АОГ – це орграф, у якому відсутні безколайдерні цикли. Монопотоковий орграф – це орграф, у якому кожний цикл суміжності має два або більше колайдерів. "Поліліс" – це орграф, у якому немає циклів суміжності. Ліс – це граф, у якому немає жодного колайдера чи цикла. Ієрархія цих моделей показана на схемі.



Ієрархія підкласів АОГ-моделей

Специфічна властивість монопотокових моделей – відсутність циклів з кількістю колайдерів менше двох – формалізується наступною аксіомою [11]

$$(X \in Anc(Y)) \,\&\, (Y \in Anc(Z)) \Rightarrow (X \perp Y \perp Z) \quad (1)$$

Говорячи неформально, у будь-яку задану змінну з кожного її предка веде єдиний оршлях. Тобто немає оршляхів ("потоків"), які дублюють один одного. З аксіоми монопотокових моделей (1) випливають наступні твердження

$$X \in \mathrm{F}(Y) \,\&\, Z \in \mathrm{F}(Y) \Rightarrow (X \perp\!\!\!\perp Z); \quad (2)$$

$$(X \in F(Y)) \,\&\, (Z \in F(Y)) \,\&\, R \in Anc(Z) \Rightarrow (X \perp\!\!\!\perp R) \,. \quad (3)$$

Властивості моделей зі структурою монопотокових орграфів залежностей (МПОГЗ) зручно виражати в апараті генотипів змінних [11]. Генотип вершини Г(*) – це множина її кореневих предків. Популяція – це множина змінних однакового генотипа. В монопотоковій моделі кожний ген надходить до змінної одним шляхом (“генеалогічною гілкою”) і успадковується змінною тільки від одного з батьків. З (2) слідує

$$(\Gamma(a) \cap \Gamma(b) \neq \varnothing) \;\&\; (a \rightarrow X) \Rightarrow \neg(b \rightarrow X). \quad (4)$$

Якщо в моделі монопотокової структури існує дуга ($X \rightarrow Y$), з тим, що X та Y належать до одної популяції, то не існує іншої дуги ($Z \rightarrow Y$). Таким чином, якщо змінна Y має батька, який належать до тої самої популяції, то інших батьків Y не має. Відтак виникає поняття клонколонії. Клонколонія $\boldsymbol{\Lambda}$ - це така множина змінних однієї популяції, що всі «зовнішні» дуги, що надходять до клонколонії $\boldsymbol{\Lambda}$, надходять до одної змінної $X=\boldsymbol{\psi}(\boldsymbol{\Lambda})$, а всі інші змінні клонколонії $\boldsymbol{\Lambda}$ є множиною потомків змінної X в складі цієї популяції. Змінна $X=\boldsymbol{\psi}(\boldsymbol{\Lambda})$ називається фундатором клонколонії $\boldsymbol{\Lambda}$. Клонколонія має структуру дерева, коренем якого є фундатор. (В одногенних популяціях немає фундаторів.)

3. Проблеми відтворення АОГ-моделей з статистичних даних. Основи сепараційного підходу

Відомо, що поліліси та ліси (дерева) залежностей ідентифікуються алгоритмами квадратичної складності [9,12], а задача ідентифікації структури залежностей в загальному випадку АОГ-моделі є NP-складною [4,9,13].

Всі індуктивні методи ідентифікації структури АОГ-моделі можна поділити на два основних підходи – “сепараційний” та “оптимізаційний” (або апроксимаційний) [3-9]. За оптимізаційного підхода максимізують критерій якості моделі в процесі підбора склада множини батьків кожної змінної. Головні проблеми, що з ними стикається цей підхід – пошук у величезному багатомірному просторі можливих структур моделі, де критерій має багаточисельні локальні максимуми [4,9].

За сепараційного підхода виявляють твердження умовної незалежності (за допомогою статистичного тестування гіпотез), з яких виводиться структура моделі [3,5,15,16]. Цей підхід має перевагу у швидкості і наочності. Процес відтворення моделі сепараційними методами складається з двох задач. Спочатку ідентифікують структуру моделі, а потім обчислюють її параметри. Перша задача розв'язується у дві фази: 1)ідентифікація всіх ребер (дуг) графа; 2) визначення орієнтацій (спрямування) дуг.

Нагадаємо поняття, важливі для сепараційного підходу. Відоме в теорії статистики [14] відношення умовної незалежності змінних X та Y при фіксації значень набора змінних S будемо виражати формулою $\mathrm{Ind}(X \perp \boldsymbol{S} \perp Y)$. При ймовірнісному трактуванні ця умовна незалежність означає, що

$$\mathrm{p}(Y|\boldsymbol{S},X) = \mathrm{p}(Y|\boldsymbol{S}). \quad (5)$$

Безумовна незалежність змінних є просто спеціальним випадком умовної незалежності з порожньою умовою, тобто $\mathrm{Ind}(X \perp \varnothing \perp Y)$, що будемо записувати стисло як $\mathrm{Ind}(X \perp\!\!\!\perp Y)$ чи через міру взаємної інформації $\mathrm{Inf}(X,Y) \approx 0$.

Статистичний факт умовної незалежності X від Y у наявних даних $\mathtt{D}$ при фіксації значень (блокуванні) $\boldsymbol{S}$ будемо позначати як $\mathrm{T}(X \perp \boldsymbol{S} \perp Y)|\mathtt{D}$. Загальновизнаним є *прагматичний постулат тестування*: з факта умовної незалежності повинне випливати відповідне кількісне співвідношення для оцінок імовірностей у відборці даних $\mathtt{D}$, а відтак коректно виконаний статистичний тест на незалежність у цій відборці повинен дати позитивну відповідь. Формально це виглядає як правило

$$(X \perp \boldsymbol{S} \perp Y) \Rightarrow \mathrm{T}(X \perp \boldsymbol{S} \perp Y)|\mathtt{D}\ . \quad (6)$$

Критерій d-сепарації, його імовірнісна інтерпретація та прагматичний постулат тестування в сукупності імплікують результати відповідних тестів незалежності за правилом

$$Ds(X \perp \boldsymbol{S} \perp Y) \Rightarrow \mathrm{T}(X \perp \boldsymbol{S} \perp Y)|\mathtt{D}. \quad (7)$$

(Надалі будемо опускати символ $\mathtt{D}$.) Кардинальність умови $\boldsymbol{S}$ визначає порядок умовної незалежності і порядок відповідного теста.

Процес ідентифікації моделі зі статистичних даних має йти від результатів статистичного тестування незалежності, тобто вимагає правил зворотного вида. Контрапозиція правила (7) дає правило ідентифікації шляхів залежності (“міцних”). Але нам треба знайти ребра графа. Фундаментом для потрібних правил може бути припущення неоманливості. Теоретики і розробники сепараційних методів ідентифікації АОГ-моделей приймають припущення неоманливості розподілу імовірностей АОГ-моделі [3,5], яке можна виразити як

$$\mathrm{Ind}(X \perp S \perp Y) \Rightarrow Ds(X \perp S \perp Y). \quad (8)$$

Проста (безумовна) форма припущення неоманливості виражається правилом

$$\mathrm{Ind}(X \perp\!\!\!\perp Y) \Rightarrow Ds(X \perp\!\!\!\perp Y) \quad (8a)$$

і стверджує, що відсутність асоційованості змінних означає відсутність безколайдерного шляха між цими змінними в орграфі моделі. Відомі методи й алгоритми [3,5,15,17] сепараційного підхода використовують самодостатнє правило скасування ребра

$$\exists S : \mathrm{T}(X \perp \boldsymbol{S} \perp Y) \Rightarrow \neg(X–Y). \quad (9)$$

Ребро (X–Y) ідентифікується, коли не існує сепаратора $\boldsymbol{S}$ такого, що чинне $\mathrm{T}(X \perp \boldsymbol{S} \perp Y)$. Оскільки сепаратор $\boldsymbol{S}$ в загальному випадку може бути множиною змінних будь-якої кардинальності (звісно, не більше за (n-2)), то головна проблема цих методів – експоненційний перебір підмножин змінних та велика кількість тестів умовної незалежності. Опріч комбінаторної складності, є актуальна проблема надійності.

Один з головних недоліків сепараційного підхода – ризик помилок при ідентифікації ребер моделі, – випливає з ненадійності результатів тестування тверджень умовної незалежності за недостатньо великого обсяга відборки даних. По мірі зростання кардинальності умови $\boldsymbol{S}$ відбувається по суті роздрібнення відборки даних. Фактор роздрібнення відборки має порядок величини ($\|X\|*\|Y\|*\|S\|$), де $\|S\| = \Pi_i \|Z_i\|$, ($Z_i \in S$). Для надійного відтворення істинної структури моделі необхідно по можливості задовільнятися тестами якомога малого порядка. Це бажано також і з точки зору трудомісткості алгоритмів.

Проблема дроблення відборки даних стосується не тільки загального випадка АОГ-моделей, але і їхнього підкласу – монопотокових моделей (МПОГЗ). Дійсно, у МПОГЗ, попри відсутність циклів з одним колайдером, кардинальність сепараторів не обмежена нічим, окрім кількості вершин графа. Застосування апарата генотипів змінних [11,16,17] дає змогу при пошуку сепаратора звузити кількість змінних-кандитатів, а отже - значно обмежити перебор і складність тестів під час ідентифікації АОГ-моделей та МПОГЗ.

4. Проблеми індуктивного відтворення монопотокових орграфів залежностей

Проблема дроблення відборки даних ускладнює ідентифікацію і АОГ-моделей залежностей, і монопотокових моделей. Відомі універсальні методи сепараційного підходу – алгоритми PC, SGS, IC та TPDA [5,12,14,18,20], – при відтворенні структур в підкласі МПОГЗ можуть потребувати тестів умовної незалежності високого порядка.

У [15] подано спеціалізований алгоритм ідентифікації структур МПОГЗ, який для кожної пари змінних виконує два тести – у форматі $\mathrm{T}(X \perp \varnothing \perp Y)$ та у форматі $\mathrm{T}(X \perp U \backslash \{X,Y\} \perp Y)$, де U – множина усіх змінних. Таким чином, у цьому алгоритмі запобігається комбінаторний перебір у пошуку сепараторів. Однак коли кількість змінних моделі є велика, тести останнього формата будуть дуже високого порядка, і, отже – ненадійними. Наприклад, коли дві змінні з'єднані шляхом із двох, трьох чи чотирьох ребер, тест умовної незалежності для цих змінних може дати негативний результат, і алгоритм [15] помилково з'єднає ці змінні ребром.

Необхідно задіяти властивості МПОГЗ, які дають змогу не тільки обмежити перебір, а й зменшити складність тестів. На використанні цих властивостей МПОГЗ та апарата генотипів змінних побудовано метод “Генеалогія”, що спирається на тести умовної незалежності тільки першого і нульового порядка. Варіанти реалізації метода – алгоритми “Генеалог-1” і “Генеалог-2” [16,17]. Обидва ці алгоритма включають процедуру “Геном-1” як свою складову. (Процедура “Геном-1” ідентифікує всі популяції та їхні генотипи у довільній АОГ.) На вхід цієї процедури треба задавати бінарне відношення неспорідненості всіх змінних, визначене як $\mathrm{NC}(X,Y) \Leftrightarrow (X \perp\!\!\!\perp Y)$. Тобто треба знати сукупність всіх фактів безумовної незалежності змінних.

Досить прийняти безумовну форму припущення неоманливості (8а), і тоді відношення неспорідненості змінних повинне ідентифікуватися тестами нульового порядка. Однак на практиці припущення неоманливості не є достатнім, а потрібно правило неоманливості відбіркового розподілу імовірностей, яке є критичне до розміру відборки і, отже, ненадійне. При відтворенні відношення спорідненості (чи неспорідненості) змінних нам необхідно знати безумовну незалежність точно і цілком. Якщо відборка даних невелика, а безколайдерний шлях між двома змінними є досить довгим, то результат статистичного тестування покаже, що ці змінні - безумовно незалежні. У такій ситуації відношення (не)спорідненості змінних або повинно бути задано апріорі, або необхідно залучати більш обачний і ретельний спосіб ідентифікації безколайдерних шляхів.

Як показав аналіз [19], навіть в класі МПОГЗ індуктивне виведення відношення спорідненості з сукупності фактів парних асоціацій змінних – нетривіальна задача. Підхід до її розв'язання принципово залежить від обсяга даних та від характера поведінки залежностей. Сприятливе і досить реалістичне припущення – статистична індикативність двореберних залежностей. Воно означає, що з факта існування між двома змінними безколайдерного шляха довжиною у два ребра неодмінно слідує, що ці дві змінні є статистично асоційованими. (Це – редукована форма припущення неоманливості.) Тоді задача виведення спорідненості змінних зводиться до прокладання (подовження) безколайдерних шляхів через злуку дотичних асоціацій. Досить виконати конкатенацію дволінкових та однолінкових залежностей так, аби компонований шлях залишався безколайдерним. Критична проблема полягає у розпізнаванні колайдерних змінних. В монопотокових орграфах залежностей ця задача полегшується тою обставиною, що для кожного колайдера ($X \rightarrow Y \leftarrow Z$) маємо $\mathrm{Ind}(X \perp\!\!\!\perp Z)$. Те ж саме вірно і для квазіколайдера, коли замість дуг маємо два шляхи, що стикаються стрілками. Однак квазіколайдерний патерн не завжди слугує доказом неспорідненості змінних, бо відбіркова статистична незалежність змінних може означати слабку залежність (коли шлях – довгий). Тобто вершина Y може вести себе як псевдоколайдерна просто через те, що розбиває “слабкий” шлях на два “дужих”

фрагмента шляха. І навпаки, не можна поспішно робити висновок про спорідненість змінних на тій підставі, що фрагменти шляха стикаються без квазіколайдерних патернів на стиках, бо є можливими помилки внаслідок проходження шляха через декілька “нереберних” асоціацій між фундаторами.

Таким чином, існує потреба в розробці нового індуктивного метода ідентифікації МПОГЗ, який не є критичний до розміра відборки даних і не потребує апріорного знання відношення спорідненості змінних. Ми пропонуємо такий метод.

5. Розробка надійного метода ідентифікації моделей залежностей в класі монопотокових орграфів

5.1 Ідентифікація колайдерних вершин в МПОГЗ

Наведені залежності в монопотокових моделях. Для ідентифікації структур монопотокових моделей нам буде потрібно виявляти і використовувати спеціальні паттерни в даних. Будемо називати паттерн $T(X\perp\!\!\!\perp Z)\&\neg T(X\perp Y\perp Z)$ *наведеною* (*спровокованою*) *залежністю* між X та Z, і при цьому змінну Y будемо називати *активатором* наведеної залежності. Будемо позначати через $Pvc(X)$ множину наведених (спровокованих) залежностей для активатора X, тобто множину усіх пар змінних, між якими змінна X наводить залежність.

Твердження 1. Якщо в МПОГЗ існує колайдер $X\rightarrow Y\leftarrow Z$, то чинне $T(X\perp\!\!\!\perp Z)\&\neg T(X\perp Y\perp Z)$.

Доведення цього твердження можна розбити на два кроки: 1) з посилки випливає проміжне $(X\perp\!\!\!\perp Z)\&\neg(X\perp Y\perp Z)$; 2) із цього проміжного твердження випливає остаточне $T(X\perp\!\!\!\perp Z)\&\neg T(X\perp Y\perp Z)$. Коректність імплікації $(X\perp\!\!\!\perp Z)\Rightarrow T(X\perp\!\!\!\perp Z)$ виправдовується прагматичним постулатом тестування, і то з високою статистичною надійністю з огляду на порожню умову незалежності. Імплікація $\neg(X\perp Y\perp Z)\Rightarrow\neg T(X\perp Y\perp Z)$ виправдовується припущенням неоманливості, локальністю фрагмента графа моделі (змінні X та Y поєднані двома суміжними ребрами) і простотою умови (у позиції сепаратора – одна змінна). Таким чином, нам залишається довести, що з посилки твердження випливає $(X\perp\!\!\!\perp Z)\&\neg(X\perp Y\perp Z)$. З умов твердження та з (2) випливає

$$(X\perp\!\!\!\perp Z). \qquad (10)$$

Залишається довести $\neg(X\perp Y\perp Z)$. Доказ негайно слідує з критерія d-сепарації. Однак цей критерій – досить незвичний для широких кіл аналітиків, тому даємо більш зрозумілий варіант доказу без залучення цього поняття. Натомість скористаємось тільки одною аксіомою АОГ-моделей – “слабкої транзитивності” [2]:

$$(X\perp S\perp Z)\&(X\perp S\cup\{Y\}\perp Z)\Rightarrow(X\perp S\perp Y) \text{ або } (Z\perp S\perp Y). \qquad (11)$$

Доведення побудуємо “від протилежного”. Нехай за умов твердження 1 буде

$$(X\perp Y\perp Z). \qquad (12)$$

З існування ребер $(X–Y)$ і $(Y–Z)$ негайно слідує

$$\neg(X\perp\!\!\!\perp Y), \qquad \neg(Y\perp\!\!\!\perp Z). \qquad (13)$$

Покладемо $S=\varnothing$ і тоді (10) і (12) дають ліву частину аксіоми “слабкої транзитивності” (11), з чого випливає $(X\perp\!\!\!\perp Y)$ або $(Y\perp\!\!\!\perp Z)$. Однак це суперечить (13). Отже, припущення (12) є невірне. □

Існує [18] і самодостатнє доведення твердження 1 (без звертання до жодних аксіом).

Легко бачити, що як змінна Y наводить залежність між X та Z, то $Y\notin Anc(X)$, $Y\notin Anc(Z)$. В монопотокових моделях кожна колайдерна змінна є активатором як мінімум однієї наведеної залежності. Однак не кожен активатор наведеної залежності є колайдерною змінною, оскільки відповідно до d-сепарації активатором для безумовно незалежних змінних X,Z може виявитися якийсь нащадок спільного родича X та Z.

Алгоритм ідентифікації колайдерних вершин. Відповідно до твердження 1, в монопотокових моделях можна знайти *всіх* кандидатів у колайдерні змінні (жоден колайдер не буде втрачений). Після цього задача зводиться до відсівання неколайдерних змінних з множини активаторів наведених залежностей.

З'ясуємо деякі корисні відносини у поводженні графічно “наближених” активаторів. Нехай між двома колайдерними змінними існує орпуть π, на якому немає третьої колайдерної змінної. Для цих змінних будуть чинні певні співвідношення на множинах наведених залежностей. (Через π позначатимемо і орпуть, і множину вершин, через які він проходить, включаючи початкову і кінцеву вершину.) Наступні твердження і наслідки доведено в [18,19].

Твердження 2. Нехай у МПОГЗ існує орпуть $\boldsymbol{\pi}$ від колайдерної вершини Y до колайдерної вершини W, і з тим усі вершини $X\in\boldsymbol{\pi}$, $X\neq Y$, $X\neq W$ є неколайдерні. Тоді є чинне

$$\neg(Pvc(Y)\supseteq Pvc(W)) \text{ і} \qquad (14)$$

$$Pvc(Y)\supseteq Pvc(X) \text{ для всіх } X\in\boldsymbol{\pi},\ X\neq Y,\ X\neq W. \qquad (15)$$

Наслідок 1. Нехай Y – колайдерна вершина в МПОГЗ, та $X\in Anc(Y)$. Як всі змінні на оршляху від X до Y (крім самої Y) мають по одному батьку, так $\neg(Pvc(Y)\subseteq Pvc(X))$.

Випливає з зіставлення (14) і (15).

Наслідок 2. Якщо в МПОГЗ маємо $\neg(Pvc(Y)\subseteq Pvc(X))$ і існує орпуть від X до Y, то X – неколайдерна вершина або на шляху від X до Y лежить як мінімум одна колайдерна вершина.

Наслідок 3. Якщо в МПОГЗ чинне $\neg(Pvc(X)\subseteq Pvc(Y))$ та $\neg(Pvc(Y)\subseteq Pvc(X))$, то неможливим є такий орпуть $\boldsymbol{\pi}$ між вершинами X та Y, на якому усі вершини, крім першої, є неколайдерними.

Встановлені твердження і наслідки дозволяють в множині активаторів наведених залежностей в МПОГЗ виділити ті змінні, що можуть бути неколайдерними. Для кожної неколайдерної змінної X, що є активатором, знайдеться така Y, що $Pvc(X) \subseteq Pvc(Y)$. Аби точно ідентифікувати (розділити) колайдерні і неколайдерні вершини серед активаторів, необхідно "підозрілі" активатори піддати тестам.

Твердження 3 (Розпізнавання неколайдерних змінних у МПОГЗ).
Змінна Y є неколайдерною, якщо для кожної X такої, що
$\exists Z$: $T(X \perp\!\!\!\perp Z) \& \neg T(X \perp Y \perp Z)$ можна знайти змінну W таку, що $T(X \perp W \perp Y)$.

Це твердження дозволяє виділити всі колайдерні вершини серед множини активаторів. З метою економії кількості виконуваних тестів буде достатнім перевірити зазначені умови не для всіх змінних X, а для однієї, бажано – для тісно (чи найдужче) асоційованої з активатором Y. Крім того, ясно, що змінну W можна шукати винятково серед активаторів, які (відповідно до твердження 2) задовільняють умові (15) для X. Відтак отримуємо

Твердження 4 (Ідентифікація колайдерних змінних у МПОГЗ).
Змінна Y є колайдерною, якщо $Pvc(Y) \neq \varnothing$ та для кожної X такої, що $\exists Z$: $T(X \perp\!\!\!\perp Z) \& \neg T(X \perp Y \perp Z)$ & $X = \operatorname{argmax}\{Inf(X,Y)\}$ чинне

$$\forall W : (Pvc(W) \neq \varnothing \;\&\; Pvc(Y) \supseteq Pvc(W)) \Rightarrow \neg T(X \perp W \perp Y) \,. \quad (16)$$

Подані твердження обґрунтовують алгоритм виявлення колайдерних вершин у МПОГЗ. Алгоритм спочатку визначає множину всіх активаторів наведених залежностей, а потім відсіває з цієї множини неколайдерні змінні згідно твердження 4.

Алгоритм "Collifinder"

1. Знайти всі наведені залежності; очистити список колайдерних змінних CL:={}
2. Помістити всі активатори в список A і упорядкувати його за зменшенням кардинальності ||Pvc(*)||
3. Розбити список A на декілька таким чином, аби у кожнім списку для сусідніх елементів X_i та X_{i+1} виконувалося $Pvc(X_i) \supseteq Pvc(X_{i+1})$, задля чого
   3.1. Утворити список списків ListA і внести в нього {A}
   3.2. **Для всіх** елементів X_{i+1} з A виконати
      Якщо $(X_i,*) \notin Pvc(X_{i+1})$ або $\neg Pvc(X_i) \supseteq Pvc(X_{i+1})$, **то** знайти такий список $L_j \in$ ListA, що для останнього елемента $Y \in L_j$ буде виконуватися умова $(Y,*) \in Pvc(X_{i+1}) \& Pvc(Y) \supseteq Pvc(X_{i+1})$, та помістити X_{i+1} у хвіст L_j , а **якщо** такого списку немає, започаткувати новий список L_k у складі ListA та помістити в нього X_{i+1} та нарахувати $k := k+1$
   Повторювати
4. Одержали k списків змінних. Перший елемент кожного списку занести в список CL колайдерних змінних
5. **Для всіх** списків $L_j \in$ ListA (j=1,…,k) виконати
   Для всіх елементів $X_{i+1} \in L_j$ виконати
      Знайти Z таку, що $(Z,*) \in Pvc(X_{i+1})$, $Z = \operatorname{argmax}\{Inf(Z, X_{i+1})\}$;
      Шукати у списку L_j серед колайдерних змінних, попередніх до Z, таку
      $X_{(i-r)} \in L_j$, що $T(Z \perp X_{(i-r)} \perp X_{i+1})$;
      якщо такої $X_{(i-r)}$ там немає, **то** занести X_{i+1} у список CL колайдерних змінних
   Повторювати

Повторювати

Позаяк кожна колайдерна змінна в МПОГЗ є фундатором клонколонії, то алгоритм "Collifinder" розв'язує ключову задачу для відтворення монопотокових структур залежностей. Знаючи всі колайдерні змінні, можна побудувати алгоритм ідентифікації структури МПОГЗ, за складністю близький до відомого алгоритма Chow&Liu [12] для лісів (дерев) залежностей.

5.2 Метод "Proliferator-C" для відтворення структури МПОГЗ. Най з'ясуємо, що заважає застосувати відомий алгоритм Chow&Liu у випадку монопотокової моделі. Є дві перепони:

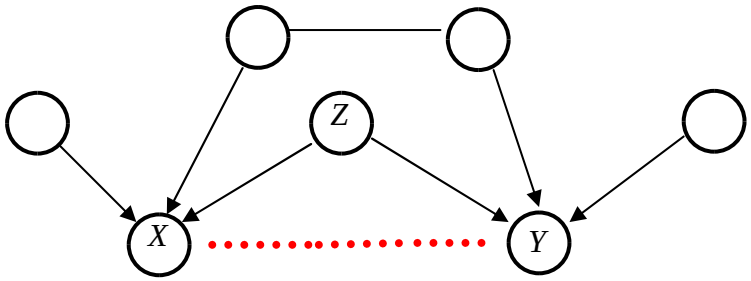
1) в МПОГЗ є можливими синергетичні асоціації, тобто такі парні асоціації змінних, що їм не відповідає жодне ребро, а проте ці асоціації – дужчі за асоціації, які відповідають ребрам, що сукупно формують синергетичну асоціацію [6];
2) в МПОГЗ є можливими цикли суміжності (з певними обмеженнями).

Однак попри ці ускладнення можна скористатися головним принципом методу Chow&Liu, і з певними вдосконаленнями втілити його в алгоритмі відтворення структур МПОГЗ. Базова ідея метода – послідовне встановлення ребер в порядку зменшення величини парних асоціацій змінних. Проблему розпізнавання «нереберних» асоціацій, що виникає при цьому, можна ефективно розв'язати, спираючись на знання колайдерних змінних (фундаторів клонколоній). В МПОГЗ існує три види «нереберних» асоціацій: транзитивні; синергетичні; комбіновані (компоновані). Розглянемо задачу розпізнавання кожного з них.

Розпізнавання транзитивних асоціацій. Транзитивна асоціація – це асоціація між змінними X та Y, коли між ними існує один безколайдерний шлях, що складається з двох, трьох або більше ребер. Вочевидь таку

асоціацію можна легко розпізнати як у дереві залежностей (метод Chow&Liu). (Доцільно однак залучити до реалізації модифіковану техніку генотипів.)

Розпізнавання синергетичних асоціацій. Нехай асоціація між змінними X та Y – не транзитивна. Синергетична асоціація в МПОГЗ – це дужа (локально-домінуюча) асоціація між двома фундаторами, не поєднаними ребром. Як розпізнати, чи є асоціація між змінними X та Y синергетичною? Тобто як відрізнити її від ребра? Якщо маємо $X \rightarrow Y$, то неодмінно буде $(X,*) \in Pvc(Y)$ та $(Y,*) \notin Pvc(X)$. А в разі синергетичної асоціації між X та Y таке є малоймовірним. Але припустимо воно все ж таки трапилось. Тоді вдаємося до іншого свідчення. Щойно маємо $X \rightarrow Y$, наразі для всіх Z таких, що $Z \in Pvc(Y) \& \neg T(X \perp\!\!\!\perp Z)$ мусить бути вірне $T(Z \perp\!\!\!\perp X \perp\!\!\!\perp Y)$. А як натомість асоціація між X та Y – синергетична, так мусить існувати Z, з якою означена імплікація не чинна (дивись фігуру).



Розпізнавання комбінованих (*компонованих*) *асоціацій*. Компонована асоціація складається з двох частин (ланок), одна з яких – синергетична асоціація між двома фундаторами клонколоній, а інша – оршлях чи ребро, що відходить від фундатора клонколонії. (Авжеж, також є можливим випадок, коли комбінована асоціація виникає на базі трьох ланок, дві з яких є оршляхами, що відходять від кожного з двох фундаторів. Проте такий випадок принципово нічим не відрізняється від дволанкового чи транзитивного.)

Нехай комбінована асоціація виникає між фундатором клонколонії X та неколайдерною змінною Y. Вочевидь існує змінна $\psi(Y)$ – фундатор клонколонії, до якої належить змінна Y, і існує оршлях від $\psi(Y)$ до Y. Зрозуміло, що на момент аналіза компонованої асоціації між X та Y оршлях від $\psi(Y)$ до Y вже буде встановлено, а отже, дуга $(X \rightarrow Y)$ буде заборонена. В той же час дуга $(Y \rightarrow X)$ не заборонена, але в разі її існування мусить бути вірне співвідношення: асоціація $\mathrm{Inf}(X,Y)$ дужча за асоціацію $\mathrm{Inf}(X,\psi(Y))$ як транзитивну. Якщо ж дуга $(Y \rightarrow X)$ не існує, а асоціація між змінними X та Y є комбінованою, то співвідношення тісноти тих асоціацій буде зворотнім.

Таким чином, озброївшись механізмом розпізнавання «нереберних» асоціацій, можна побудувати алгоритм відтворення структур монопотокових моделей залежностей на базі ідеї метода Chow&Liu. Орієнтації дуг можна ідентифікувати за аналогом відомого “колайдерного” правила. Будемо називати цей метод “Proliferator-C”.

Блок-схема метода “Proliferator-C”

1) Обчислити парні асоціації змінних
2) Впорядкувати пари у списку ***L*** за зменшенням асоціацій
3) **Для всіх** пар змінних (X,Y) зі списка ***L***
   Якщо обидві змінні – неколайдерні, і асоціація не є ані транзитивною, ані компонованою, **то** встановити ребро $(X–Y)$, **інакше**
   Якщо одна змінна – фундатор, а інша – неколайдерна, і асоціація не є ані транзитивною, ані компонованою, **то** встановити ребро $(X–Y)$, **інакше**
   Якщо обидві змінні – фундатори і асоціація не є ні транзитивною, ні синергетичною, ні компонованою, **то** встановити ребро $(X–Y)$
   ;
   ;
   ;

Орієнтувати ребра за правилом

$$(X–Y–Z) \;\&\; (X,Z) \in Pvc(Y) \Rightarrow X \rightarrow Y \leftarrow Z$$

Орієнтувати ребра за правилом

$$(X \rightarrow Y–Z) \;\&\; (X,Z) \notin Pvc(Y) \Rightarrow X \rightarrow Y \rightarrow Z$$

Повторювати

Таким чином, поєднання методів “Collifinder” та “Proliferator-C” дає надійний і ефективний метод відтворення структур монопотокових моделей залежностей “з чистого аркушу” без тестів вище першого порядка.

6. Підсумки

В монопотокових моделях, як і в АОГ-моделях, кардинальність сепараторів може бути великою. Розроблений раніше метод “Генеалогія” для відтворення структур монопотокових моделей потребує вичерпного знання бінарного відношення спорідненості змінних. Проте в умовах невеликого обсягу даних індуктивно вивести відношення спорідненості змінних – непросто. Ми знайшли інший підхід, який не потребує ані апріорного знання відношення спорідненості змінних, ані повного його індуктивного виведення. Розроблено метод і алгоритм “Collifinder”, який ідентифікує всі колайдерні змінні в монопотоковій моделі. Метод грунтується на феномені наведених залежностей між змінними і застосовує мінімальний обсяг тестів умовної незалежності першого порядка.

Показано, що для відтворення структур монопотокових моделей залежностей можна взяти базову ідею метода Chow&Liu. Задача розпізнавання «нереберних» асоціацій, що виникає при цьому, ефективно розв'язується на основі знання колайдерних змінних. Отже алгоритм “Collifinder” разом з методом “Proliferator-C” (який є узагальненим і вдосконаленим варіантом метода Chow&Liu) повністю розв'язує проблему ідентифікації структури монопотокових моделей за допомогою тестів умовної незалежності першого порядка. Наш метод не потребує апріорних знань про структуру і працює в умовах відносно невеликого обсяга даних, а за складністю – близький до відомих алгоритмів для дерев залежностей.

Ефективність запропонованого метода грунтується на системній організації процеса ідентифікації дуг графа моделі як послідовного уточнення знань про структуру через виконання лише доконче необхідних тестів у найпростішому форматі на підставі аналіза всієї проміжної інформації.

Література

1. *Lauritzen S.L.* Graphical Models.- Clarendon Press, Oxford, UK, 1996.
2. *Pearl J*. Probabilistic reasoning in intelligent systems: networks of plausible inference. – Morgan Kaufmann, 1988.
3. *Spirtes P., C.Glymour, and R.Scheines.* Causation, prediction and search. (2-nd Ed.), – New York: MIT Press, 2001. – 496p.
4. *Heckerman D*. Bayesian networks for data mining// Data Mining and Knowledge Discovery, 1997. – Vol.**1**. – N 1. – P.79-119.
5. *Cheng J., R.Greiner, J.Kelly, D.Bell and W.Liu*. Learning bayesian networks from data: an information-theory based approach// Artificial Intelligence, 2002, Vol. 137, P.43-90.
6. *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.
7. *Андон Ф.И., А.С.Балабанов*. Выявление знаний и изыскания в базах данных: подходы, модели, методы и системы (обзор). – Материалы 2-й междунар. конф. “УкрПРОГ-2000”// Проблемы программирования, 2000. – № 1-2. – С.513-526.
8. *Балабанов А.С*. Выделение знаний из баз данных – передовые компьютерные технологии интеллектуального анализа данных// Математические машины и системы, 2001. – №1/2. – С.40-54.
9. *Heckerman D., D.Geiger, and D.M.Chickering*. Learning bayesian networks: the combination of knowledge and statistical data // Machine Learning, 1995. – Vol.**20**. – P.197-243.
10. *Verma T. and J.Pearl*. Causal Networks: Semantics and Expressiveness/ in: R. Shachter, T.S.Levitt, and L.N.Kanal (Eds.), Uncertainty in AI, **4**, Elsevier Science Publishers, 1990.– P.69-76.
11. *Балабанов А.С*. Восстановление структур систем вероятностных зависимостей из данных. Аппарат генотипов переменных// Проблемы управления и информатики.– 2003.– № 2. – С.91-99.
12. *Chow C.K., and C.N.Liu.* Approximating discrete probability distributions with dependence trees//IEEE trans. on Information Theory, 1968. – Vol.14. – N3. – P.462-467.
13. *Chickering D.M*. Learning Bayesian networks is NP-complete//In: Fisher, D. and Lenz, H., editors, *Learning from Data: Artificial Intelligence and Statistics* **5**, pages 121-130. Springer-Verlag, 1996.
14. *Dawid A.P*. Conditional independence in statistical theory (with discussion)// Journal of Roy. Statist. Soc., ser.**B.,** Vol.**41,** P.1-31. (1979).
15. *Geiger D., A.Paz, J.Pearl*. Learning simple causal structures// Intern. Journal of Intelligent Systems. – 1993, **8**, – N2, P.231-247.
16. *Балабанов О.С.* Новий метод відтворення ймовірнісних графових моделей залежностей// Праці 1-ї міжнар. конф. з індуктивного моделювання “МКІМ-2002”, т.**1**, С.118-124, Львів, 20-25 травня 2002. (Proceedings of “ICIM-2002”, Vol.**1**, P.118-124, Lviv, UA).
17. *Балабанов А.С.* Индуктивный метод восстановления монопотоковых вероятностных графовых моделей зависимостей// Проблемы управления и информатики. – 2003.- № 5. – С.75-84.
18. *Балабанов А.С.* К выводу структур моделей вероятностных зависимостей из статистических данных – (2003). (подано в журнал “Кибернетика и системный анализ”).
19. *Балабанов О.С.* “Дослідження і розробка методологічних основ та комп'ютерних методів виділення знань з даних, індуктивного виводу закономірностей і моделей предметних галузей для прикладних прогнозно-аналітичних досліджень”– Звіт з НДР 3/01К.02-02. – ІПС НАН України, Київ, 2003. – 28с.