УДК 007:681.3.00

О.С. Балабанов

КВАЗІ-ЛІНЕАРНІСТЬ У ДИСКРЕТНИХ МОДЕЛЯХ ЗАЛЕЖНОСТЕЙ ТА ВІДКРИТТЯ ЛАТЕНТНОГО ФАКТОРА ТРЬОХ ЕФЕКТІВ

Для дискретних моделей залежностей з ланцюговою (або деревовидною) структурою показано, що коли проміжна (сепараторна) змінна є бінарною, можна факторизувати (декомпозувати) транзитивну залежність згідно відтинків ланцюга. Ця властивість ("квазі-лінеарність") для структури у формі "зірка з трьома променями" імплікує "тріад-стримування" – спеціальне обмеження типу рівність на добуток парних залежностей. Дотримання чинності тріад-стримування може правити за свідчення для ідентифікації прихованої бінарної змінної, яка є відповідальна за асоціацію трьох дискретних змінних.

Вступ

Результати роботи є внеском у дослідження та розробку методів глибокого аналізу даних (категорних) та відкриття структур залежностей у даних.

Добре відомо, що лінійно-нормальні системи залежностей мають цікаву властивість – мультиплікативність (факторизуємість) коефіцієнта кореляції на ланцюговій чи деревовидній структурі залежностей. Тобто для ланцюгів у складі лінійних моделей з дійсними змінними та нормальними розподіленнями відомо [1–3], що коефіцієнт кореляції для транзитивної залежності дорівнює добутку коефіцієнтів кореляції для всіх ланок, які утворюють ланцюг. Наприклад, у моделі, яка має "ланцюгову" структуру *X*–*Y*–*Z*, коефіцієнт кореляції між змінними *X* та *Z* дорівнює $r_{XZ} = r_{XY} \cdot r_{YZ}$ (добутку коефіцієнтів кореляції для відтинків "ланцюга"). Як показано в [4], аналогічна мультиплікативність чинна для коефіцієнта номінальної детермінації у моделі з бінарними змінними, як також у спеціальних випадках моделі з змінними довільної значності [5]. Покажемо, що така сама мультиплікативність (або квазі-лінеарність) залежності виконується і в інших випадках, тобто узагальнимо відомі результати і поширимо їх на інші моделі. Ця властивість оснащує емпіричне відкриття латентного фактору, спільного для кількох (не менше трьох) спостережуваних змінних.

1. Міра залежності номінальних змінних та властивості бінарних моделей

Коефіцієнт номінальної стохастичної детермінації (NSD-коефіцієнт) [5–7] вимірює силу залежності між двома випадковими змінними номінального (категорного) типу. Нагадаємо, що цей коефіцієнт визначається як

$$d_{\text{nom}}(Y \leftarrow X) = \\ = C_N * \sum_x \sum_y \left(p(Y \mid X) - \frac{1}{\left\| X \right\|} \sum_x p(Y \mid X) \right)^2, \quad (1)$$

де C_N – нормалізаційний коефіцієнт, $\|X\|$ – значність змінної *X*.

Була запропонована також інша форма (версія) NSD-коефіцієнта, яка відрізняється використанням модуля замість квадрата [5]. Модуль-форма коефіцієнта номінальної детермінації, яку будемо називати *індексом детермінації*, визначається як

$$di(Y(X)) = \\ = C_{N-a} \sum_x \sum_y \left| p(Y \mid X) - \frac{1}{\left\| X \right\|} \sum_x p(Y \mid X) \right|. \quad (2)$$

У випадку бінарних змінних формула для коефіцієнта номінальної стохастичної детермінації спрощується [7] до вигляду

$$d_{nom}(Y \leftarrow X)_{[2]} = \\ = \left(p(y_1 | x_1) - p(y_1 | x_2) \right)^2, \quad (3)$$

де індекси змінних вказують на значення бінарних змінних.

Надалі будемо користуватися індексом детермінації (2). Для бінарних змінних він спрощується до вигляду

$$di(Y(X)) = |\mathrm{p}(y_1|x_1) - \mathrm{p}(y_1|x_2)|. \quad (4)$$

(Нормалізаційний коефіцієнт $C_{N\text{-}a}$ в цьому випадку дорівнює 1/2.)

Можна перетворити індекс детермінації для бінарних змінних так:

$$di(Y(X)) = |p(y_1|x_1) - p(y_1|x_2)| = |\,[p(y_1,x_1)p(x_2) - p(y_1,x_2)p(x_1)]/p(x_1)p(x_2)\,| = |\,[p(y_1,x_1)p(y_2,x_2) - p(y_1,x_2)p(y_2,x_1)]/p(x_1)p(x_2)\,|.$$

$$di(X(Y)) = |p(x_1|y_1) - p(x_1|y_2)| = |\,[p(y_1,x_1)p(y_2) - p(y_2,x_1)p(y_1)]/p(y_1)p(y_2)\,| = |\,[p(y_1,x_1)p(y_2,x_2) - p(y_1,x_2)p(y_2,x_1)]/p(y_1)p(y_2)\,|.$$

Тепер беремо добуток отриманих формул

$$di(Y(X)) \cdot di(X(Y)) = [p(y_1,x_1)p(y_2,x_2) - p(y_1,x_2)p(y_2,x_1)]^2/[p(x_1)p(x_2)p(y_1)p(y_2)]. \quad (5)$$

Як бачимо, добуток індексів детермінації у прямому та зворотному напрямі дорівнює відомому в статистиці коефіцієнта контингенції [8].

У роботі [4] показано, що для ланцюгової структури бінарних змінних X–Y–Z є чинним співвідношення

$$d_{\text{nom}}(Z \leftarrow X) = d_{\text{nom}}(Y \leftarrow X) * d_{\text{nom}}(Z \leftarrow Y), \quad (6)$$

назване мультиплікативним послабленням транзитивної залежності. Далі будемо називати цю властивість квазі-лінеарністю моделі, або квазі-лінеарністю залежності на ланцюгу. Легко показати, що така сама квазі-лінеарність виконується і для індекса детермінації (на ланцюгу бінарних змінних). Далі покажемо, що має силу більш загальний результат.

2. Узагальнення чинності квазі-лінеарності ланцюгів залежностей у дискретних моделях

Нехай маємо ланцюгову (або каскадну) структуру залежностей у вигляді X–Y–Z, тобто таку структуру моделі, що змінна Z умовно незалежна [9] від змінної X за кондиціонування змінної Y; позначимо цю умовну незалежність $\Pr(X \perp Y \perp Z)$. (Тут є зовнішня подібність до Марковського ланцюга, але нагадаємо, що тут X, Y та Z є окремими змінними, а не послідовними станами одного процесу.) Нехай також медіаторна змінна Y буде бінарна, а змінні X та Z – дискретні довільної значності, r та q відповідно. Покажемо, що в цьому випадку теж виконується квазі-лінеарність залежності.

Для зручності (аби уникнути питання обчислення нормалізаційного коефіцієнта) у подальшому переважно будемо застосовувати *ненормалізовану* форму *індексу детермінації*. Тобто просто відкинемо коефіцієнт $C_{N\text{-}a}$ і позначимо ненормалізований індекс детермінації $d(Y(X))$.

Спочатку спростимо вираз для $d(Y(X))$, коли медіаторна змінна Y – бінарна, а змінна X – дискретна r-значна. Тривіальні маніпуляції дають

$$d(Y(X)) = \sum_x\sum_y |p(y|x) - (\sum_x p(y|x))/r| = 2\sum_x \{|\,p(y_1|x) - (\sum_x p(y_1|x))/r)\,|\}. \quad (7)$$

Запишемо вираження для індексу детермінації $d(Z(Y))$, коли змінна Z є q-значна, а змінна Y – бінарна:

$$d(Z(Y)) = \sum_y\sum_z |p(z|y) - (\sum_y p(z|y))/2| = \sum_z \{|p(z|y_1) - p(z|y_2)|\}. \quad (8)$$

Зауважте, що формула (8) не містить q в тілі свого ядра, тобто значність "ефекторної" змінної Z не відбивається на вигляді формули. Тепер формулюємо базовий результат.

Твердження 1. Якщо змінна Z умовно незалежна від X за кондиціонування змінної Y, де змінна Y – бінарна, змінна X – r-значна, а Z – q-значна, то:

$$d(Z(X)) = d(Y(X)) \cdot d(Z(Y))/2\,. \quad (9)$$

Доведення. Маємо

$$d(Z(X)) = \sum_x\sum_z |p(z|x) - (\sum_x p(z|x))/r|. \quad (10)$$

Використовуючи умовну незалежність Z від X при фіксованому Y, а також бінарність змінної Y, запишемо

$$p(z|x)=\sum_y [p(z|y) \cdot p(y|x)] = p(y_1|x)[p(z|y_1) - p(z|y_2)] + p(z|y_2). \quad (11)$$

Підставляючи (11) до (10), отримуємо

$$d(Z(X)) = \sum_x\sum_z |p(y_1|x)[p(z|y_1) - p(z|y_2)] - [p(z|y_1) - p(z|y_2)] \cdot (1/r) \cdot \sum_x p(y_1|x)| = \sum_z |p(z|y_1) - p(z|y_2)| \cdot \sum_x |p(y_1|x) - (1/r) \cdot \sum_x p(y_1|x)|. \quad (12)$$

З огляду на (7) та (8), останнє (12) дає потрібне (9).☑

Варто зауважити, що емпіричне дотримання квазі-лінеарності, тобто приблизне виконання рівності (9) у скінченій відбірці даних, можна використовувати як свідчення умовної незалежності $\Pr(X \perp Y \perp Z)$, ко-

ли Y – бінарна. Звісно, (9) є лише необхідною, але не достатньою умовою. Проте достоїнством такого засобу верифікації умовної незалежності є використання лише парних статистик. У практичних ситуаціях можуть бути відсутні дані у формі таблиці з трьома змінними, що унеможливлює оцінювання сумісного розподілення трьох змінних.

Тепер подаємо цікавий випадок квазілінеарності, який не потребує бінарності сепараторної змінної. Однак він нав'язує спеціальну форму залежності, яка була названа в [5, 6] взаємно-однозначним відображенням з рівномірно-розподіленим розсіянням. Ця залежність можлива за передумови $||Z||=||Y||=r$ та описується так:

$$p(z_j \mid y_i) = \begin{cases} 1-e*(r-1), if\ (i=j), \\ e, if\ (i \neq j). \end{cases} \quad (13)$$

Таку залежність можна вважати r-значним симетричним каналом передачі інформації з гамором, тому будемо називати її *SCC-залежністю*. Її поведінка має дещо спільне з лінійною залежністю за нормальних розподілень. Легко з'ясувати, що для цієї залежності буде

$$d(Z(Y)) = 2 \cdot (r-1) \cdot |1 - e \cdot r|. \quad (14)$$

Твердження 2. Якщо змінна Z умовно незалежна від X при кондиціонуванні змінної Y, $||X||=||Y||=||Z||=r$ та залежність Z від Y є SCC-залежністю (13), то чинне:

$$d(Z(X)) = d(Y(X)) \cdot d(Z(Y))/2 \cdot (r-1). \quad (15)$$

Доведення. Оскільки чинна умовна незалежність $\Pr(X \perp Y \perp Z)$, і маємо (13), то отримуємо

$$\begin{aligned} p(z_j|x_i) &= \textstyle\sum_k p(z_j|y_k) \cdot p(y_k|x_i) = \\ &= p(z_j|y_j) \cdot p(y_j|x_i) + \textstyle\sum_{k,k\neq j} p(z_j|y_k) \cdot p(y_k|x_i) = \\ &= (1 - e \cdot (r-1)) \cdot p(y_j|x_i) + e \cdot \textstyle\sum_{k,k\neq j} p(y_k|x_i). \end{aligned}$$

З огляду на $\sum_{k,k\neq j} p(y_k|x_i) = 1 - p(y_j|x_i)$ здобуваємо

$$p(z_j|x_i) = e + (1 - e \cdot r) \cdot p(y_j|x_i). \quad (16)$$

Тепер підставляємо (16) до $d(Z(X))$ і отримуємо

$$\begin{aligned} d(Z(X)) &= \\ &= \textstyle\sum_{i,j} |p(z_j|x_i) - (1/r) \sum_i p(z_j|x_i)| = \\ &= |1 - e \cdot r| \cdot \textstyle\sum_{i,j} |p(y_j|x_i) - (1/r) \sum_i p(y_j|x_i)|. \end{aligned} \quad (17)$$

Зіставляючи (17) з (14) констатуємо потрібне (15).☑

Зверніть увагу, що чинність (15) не пов'язується зі значенням параметра e. Така сама квазі-лінеарність чинна у разі, коли SCC-залежність посідає першу, а не другу позицію в "ланцюзі" [5]. (Зрозуміло, всі подібні рівності практично слід тлумачити в асимптотичному сенсі, коли йдеться про відбіркові оцінки.)

Вкладений випадок. Квазі-лінеарність можна знайти в інших (складніших) моделях. Структура моделі для реальних задач може не містити фрагментів типу "Марковський ланцюг", а натомість мати фрагменти з дещо складнішою структурою. Наприклад, як на рис.1. Треба зазначити, що структури рис.1, в та 1, г принципово не відрізняються

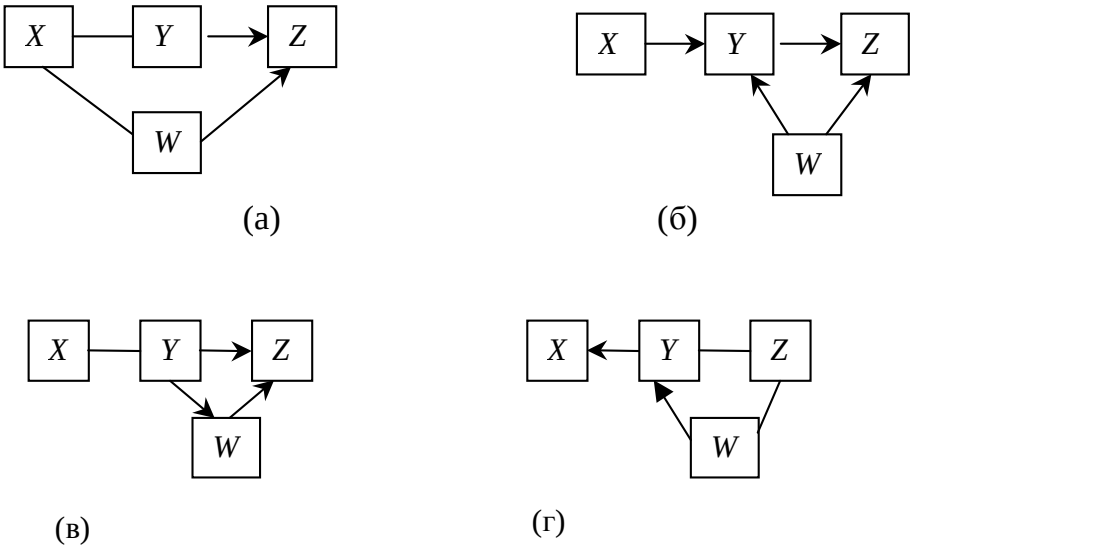


Рис. 1

від звичайної ланцюгової структури залежностей вигляду *X–Y–Z*, оскільки для них теж чинна умовна незалежність Pr(*X*⊥*Y*⊥*Z*). А от у структурах рис.1,а та 1, б умовна незалежність Pr(*X*⊥*Y*⊥*Z*) не виконується [9], однак виконується Pr(*X*⊥*YW*⊥*Z*). Легко бачити, що за блокування змінної *W* ми отримаємо "вкладений ланцюг" *X–Y–Z*. Інтуїтивно має бути зрозуміло, що за умови блокування змінної *W* має виконуватися квазі-лінеарність залежності. Однак для формалізації цього треба дати визначення "умовного індексу детермінації", тобто узагальнити визначення нашої міри *di*(*). Для зручності (спрощення технічних викладок) обмежимось розглядом випадку бінарних змінних *X*, *Y* та *Z*.

Визначимо "константно-умовний індекс детермінації" (у нормалізованій формі) для бінарних змінних *X* та *Y* і дискретної змінної *W* (довільної значності), яка набуває значення w_i, так:

$$di(Y(X)|_{w\text{-}i}) = |p(y_1|x_1, w_i) - p(y_1|x_2, w_i)|. \quad (18)$$

Покажемо, що для так визначеного умовного індексу детермінації в моделі, яка зображена на рис.1, а, має виконуватись квазі-лінеарність залежності на ланцюгу. Зауважимо, що для моделі рис.1, а завдяки відношенню умовної незалежності змінних *Y* та *W* при кондиціонуванні *X* маємо $di(Y(X)|_{w\text{-}i}) = di(Y(X))$, тобто вираження умовного індексу детермінації співпадає зі стандартним (звичайним, безумовним). Відтак, доведення чинності квазі-лінеарності для умовного індексу детермінації залежності нашої моделі можна розпочати з дещо спрощеного вираження. Отже, трансформуємо

$$\begin{aligned} di(Y(X)) \cdot di(Z(Y)|_{w\text{-}i}) &= |p(y_1|x_1) - \\ &- p(y_1|x_2)| \cdot |p(z_1|y_1, w_i) - p(z_1|y_2, w_i)| = \\ &= |p(z_1|y_1, w_i) \cdot [p(y_1|x_1) - p(y_1|x_2)] - \\ &- p(z_1|y_2, w_i) \cdot [p(y_2|x_2) - p(y_2|x_1)]|. \quad (19) \end{aligned}$$

Враховуючи умовну незалежність Pr(*Z*⊥*YW*⊥*X*) та Pr(*Y*⊥*X*⊥*W*), маємо

$$p(z|y,w_i) = p(z|y,x,w_i),$$
$$p(y|x) = p(y|x,w_i).$$

Згортаємо добуток останніх згідно байєсівської формули

$$p(z|y,w_i) \cdot p(y|x) = p(z|y,x,w_i) \cdot p(y|x,w_i) = p(zy|x,w_i).$$

З урахуванням цього формула (19) набуває вигляду

$$\begin{aligned} di(Y(X)) \cdot di(Z(Y)|_{w\text{-}i}) &= \\ = |p(z_1, y_1|x_1, w_i) - p(z_1y_1|x_2,w_i) &- \\ - p(z_1, y_2|x_2, w_i) + p(z_1, y_2|x_1,w_i)| &= \quad (20) \\ = |p(z_1|x_1, w_i) - p(z_1|x_2,w_i)| &= \\ = di(Z(X)|_{w\text{-}i}). & \end{aligned}$$

Отже, на "вкладеному ланцюгу" (рис.1, а) виконується квазі-лінеарність залежності для константно-умовного індексу детермінації.

Легко показати аналогічне для моделі рис.1, б, з тією відмінністю, що тоді рівність $di(Y(X)|_{w\text{-}i})=di(Y(X))$ не буде чинна, а відтак замість (20) отримаємо аналогічну рівність з трьома константно-умовними індексами детермінації:

$$di(Y(X)|_{w\text{-}i}) \cdot di(Z(Y)|_{w\text{-}i}) = di(Z(X)|_{w\text{-}i}). \quad (21)$$

Продемонстрована властивість може здатися тривіальною, оскільки маємо Pr(*X*⊥*YW*⊥*Z*), тобто за кондиціонування змінної *W* виникає "вкладений ланцюг" *X–Y–Z*, для якого формально є чинна квазі-лінеарність залежності. До того ж квазі-лінеарність залежності для константно-умовного індексу детермінації має меншу цінність, ніж основне (9), додання умови автоматично означає роздрібнення відбірки даних [9], і, як наслідок, – посилення ефекту відбіркового ухилення (від строгої рівності). Хоча як деяку "компенсацію" цього ефекту можна зарахувати збільшення кількості приблизних рівностей. Якщо з метою відновлення статистичної робастності звести в суму всі рівності у формі (21) для всіх *i*, то отримаємо формулу, яка за своєю структурою не є мультиплікативною (факторизованою), тож феномену квазі-лінеарності не буде.

Цікавіше було б дати таке узагальнене визначення умовного індексу детермінації, яке б охоплювало (підсумовувало) усі значення змінної *W*, аналогічно тому, як це є у формулі умовної взаємної інформації. А вже для такого умовного індексу детермінації було б корисно встановити відповідну форму квазі-лінеарності залежності.

3. Аналіз структури "зірка з бінарним вузлом"

Нехай маємо модель, що зображена на рис. 2, де H – бінарна змінна. (Можна зауважити, що всі наші висновки збережуться, якщо дуги замінити на неорієнтовані ребра.) Подібні моделі мають застосування, зокрема, у соціометриці, психометриці та медико-біологічних дослідженнях. У такій структурі три дискретні змінні є асоційовані через одну бінарну "центральну" (вузлову, медіаторну) змінну, тому кондиціонування центральної змінної робить всі інші змінні взаємно незалежними [9]. Тобто в цій моделі виконуються твердження умовної незалежності $\Pr(X \perp H \perp Y)$, $\Pr(X \perp H \perp Z)$, $\Pr(Y \perp H \perp Z)$, та симетричні варіанти.

Отже, ця модель задовольняє умовам твердження 1, і можемо записати рівняння (9) з відповідною підстановкою змінних. Для спрощення викладок надалі будемо писати $Y(X)$ замість $d(Y(X))$. Отже, отримуємо систему рівнянь:

$$\begin{aligned} X(Y) &= X(H) \cdot H(Y)/2, \\ Y(X) &= Y(H) \cdot H(X)/2, \\ Z(X) &= Z(H) \cdot H(X)/2, \\ X(Z) &= X(H) \cdot H(Z)/2, \\ Z(Y) &= Z(H) \cdot H(Y)/2, \\ Y(Z) &= Y(H) \cdot H(Z)/2. \end{aligned} \qquad (22)$$

Особливість цієї системи рівнянь полягає в тім, що члени зі змінною H входять до системи парами, як відношення (пропорції). Таких членів маємо 6, як і рівнянь системи. Тривіальні алгебраїчні перетворення дозволяють позбавлятися цих членів. А оскільки ці члени входять до системи тільки як відношення, то в процесі перетворень останні з цих членів взаємознищуються. Тож система редукується до єдиного рівняння, яке не містить членів зі змінною H:

$$X(Y) \cdot Y(Z) \cdot Z(X) = Y(X) \cdot Z(Y) \cdot X(Z), \quad (23)$$

яке будемо називати "тріад-стримуванням". Воно характеризує модель вигляду "зірка з бінарним вузлом" і констатує інваріантність добутку трьох парних залежностей до реверсу всіх цих залежностей (рівночасної їх заміни на "каузально-обернені").

Маючи на меті застосувати (23) як свідчення для ідентифікації моделі, слід розвідати альтернативні моделі, які теж додержуються тріад-стримування. Звісно, деградований випадок рівності 0= 0 відкидається.

4. Альтернативні дискретні моделі, які підтримують "тріад-стримування"

4.1. Серед альтернатив, де виконується тріад-стримування (23), найбільш варта уваги модель з *бінарними змінними*. Дійсно, нехай маємо три асоційовані бінарні змінні X, Y, Z. Візьмемо ненормалізований індекс детермінації для бінарних змінних (див. (4)) $Y(X) = 2 \cdot |\mathrm{p}(y_1|x_1) - \mathrm{p}(y_1|x_2)|$ та перетворимо його так:

$$\mathrm{p}(x_1) \cdot \mathrm{p}(x_2) \cdot Y(X) = 2 \cdot |\mathrm{p}(x_2) \cdot \mathrm{p}(y_1,x_1) - \mathrm{p}(x_1) \cdot \mathrm{p}(y_1,x_2)| =$$

$$= 2 \cdot |\mathrm{p}(y_1,x_1) - \mathrm{p}(x_1) \cdot \mathrm{p}(y_1,x_1) - \mathrm{p}(x_1) \cdot \mathrm{p}(y_1,x_2)| = 2 \cdot |\mathrm{p}(y_1,x_1) - \mathrm{p}(x_1)\mathrm{p}(y_1)|.$$

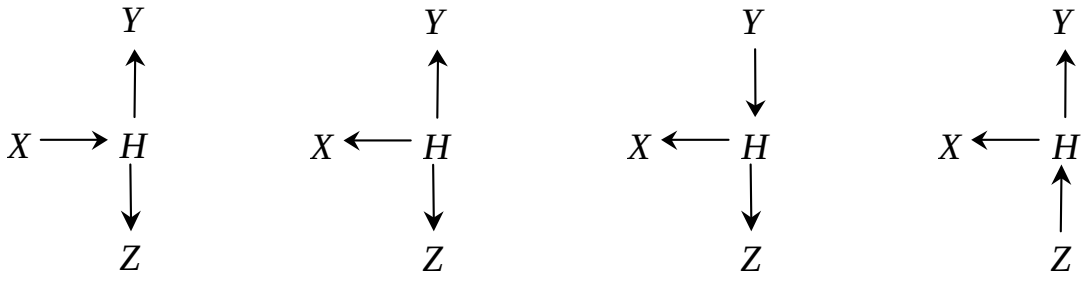
Отже, маємо

$$Y(X) = 2|\mathrm{p}(y_1,x_1) - \mathrm{p}(x_1)\mathrm{p}(y_1)|/\mathrm{p}(x_1) \cdot \mathrm{p}(x_2). \quad (24)$$

Аналогічно буде

$$X(Y) = 2|\mathrm{p}(y_1,x_1) - \mathrm{p}(x_1)\mathrm{p}(y_1)|/\mathrm{p}(y_1) \cdot \mathrm{p}(y_2). \quad (25)$$

Очевидно, так само буде і для решти пар



(а) (б) (в) (г)

Рис. 2

бінарних змінних:

$$Y(Z) = 2 \cdot |p(z_1,y_1) - p(y_1)p(z_1)|/p(z_1) \cdot p(z_2); \quad (26)$$

$$X(Z) = 2 \cdot |p(z_1,x_1) - p(x_1)p(z_1)|/p(z_1) \cdot p(z_2); \quad (27)$$

$$Z(X) = 2 \cdot |p(z_1,x_1) - p(x_1)p(z_1)|/p(x_1) \cdot p(x_2); \quad (28)$$

$$X(Z) = 2 \cdot |p(z_1,x_1) - p(x_1)p(z_1)|/p(z_1) \cdot p(z_2); \quad (29)$$

$$Z(Y) = 2 \cdot |p(z_1,y_1) - p(y_1)p(z_1)|/p(y_1) \cdot p(y_2). \quad (30)$$

Тепер ліву частину тріад-стримування (23) виражаємо через (25), (26) та (28):

$$X(Y) \cdot Y(Z) \cdot Z(X) = \\ = |p(y_1,x_1) - p(x_1)p(y_1)| \cdot |p(z_1,y_1) - \\ - p(y_1)p(z_1)| \cdot |p(z_1,x_1) - p(x_1)p(z_1)|/A, \quad (31)$$

де $A = p(x_1) \cdot p(x_2) \cdot p(y_1) \cdot p(y_2) \cdot p(z_1) \cdot p(z_2)/8$.

Легко перевірити, що, виразивши праву частину тріад-стримування (23) через (24), (30) та (29), ми отримаємо таку саму формулу, що і в (31).

Таким чином, в моделі з бінарними змінними тріад-стримування (23) – це тотожність, яка виконується завжди, безвідносно до структури моделі.

4.2. Інший альтернативний випадок – це модель, де виконується *умовна незалежність* якихось двох змінних при кондиціонуванні третьої (бінарної) змінної. Нехай модель – це ланцюгова структура *X–Y–Z*, яка, зокрема, має структуру рис.1, б або 1, г, де змінна *Y* – бінарна. Тоді чинна умовна незалежність $Pr(X \perp H \perp Y)$, і відтак виконується квазі-лінеарність залежності (9), і симетрична формула. Відтак тривіально слідує (23).

Припустимо, наприклад, модель є "ланцюгом" *Z–X–Y*, (де *X* – бінарна). Тоді отримуємо

$$Y(Z) = Y(X) \cdot X(Z), \quad (32)$$

$$Z(X) \cdot X(Y) = Z(Y). \quad (33)$$

Добуток (32) та (33) дає (23). Таку саму ситуацію отримаємо в разі, коли в моделі рис. 2 одна з "реберних" залежностей (наприклад $Z(H)$) буде детерміністична.

Всі інші (відомі автору) альтернативи, де виконується тріад-стримування, є спеціальними випадками, тобто потребують співвідношень між параметрами моделі.

4.3. Тривіальний випадок – коли існує *детерміністична залежність між двома* змінними. Тобто, наприклад, нехай між змінними *X* та *Y* існують взаємно однозначні відношення. Легко бачити, в такому разі буде $X(Y) = Y(X)$, $Y(Z) = X(Z)$, $Z(X) = Z(Y)$, звідки слідує (23).

Існують також спеціальні випадки зореподібних дискретних моделей, коли вузлова змінна не є бінарною. Спочатку розглянемо випадок, де обмеження стосуються усіх трьох залежностей моделі. А в подальших двох випадках обмежуються дві з трьох залежностей.

4.4. Нехай модель має структуру рис. 2, б, із тим, що три залежності "індикаторних" змінних *X*, *Y*, *Z* від "центральної" є *SCC-залежностями* (див. (13)). В такому разі теж виконується квазі-лінеарність залежності (твердження 2), і отримуємо систему рівнянь, яка відрізняється від (22) лише коефіцієнтами пропорційності. Ці коефіцієнти взаємно скорочуються, і відтак слідує (23).

4.5. Інший випадок – коли "центральна" змінна *H* поводить себе щодо двох індикаторних змінних як бінарна, і лише для третьої змінної демонструє справді більшу значність. Нехай, наприклад, змінна *H* буде тризначна. Тоді цей випадок буде мати місце, коли для двох значень змінної *H* – *i*-го та *j*-го – спостерігається співпадіння розподілів, наприклад $p(X|h_i) = p(X|h_j)$, $p(Y|h_i) = p(Y|h_j)$. Зрозуміло, в такому разі можна уявити, що між змінною *H* та парою змінних *X*, *Y* розташована бінарна змінна *V*, так що виконуються незалежності $Pr(X \perp V \perp Y)$, $Pr(X \perp V \perp H)$, $Pr(Y \perp V \perp H)$, $Pr(X \perp V \perp Z)$ та $Pr(Y \perp V \perp Z)$. Легко бачити, що така модель, по-суті, є "зіркою з бінарним вузлом", де роль "центрального" вузла виконує змінна *V*. Тоді чинне (23).

4.6. Ще один випадок – ізоморфізм двох залежностей, тобто співпадіння, з точністю до перейменування значень, умовних (за умови на "центральну" змінну) розподілів двох індикаторних змінних. Достатньо одного з трьох варіантів ізоморфізму розподілів: $P(X|H) \square P(Y|H)$, або $P(X|H) \square P(Z|H)$, або $P(Y|H) \square P(Z|H)$. Нехай буде $P(X|H) \square P(Y|H)$. Звідси негайно висновуємо $X(Y) = Y(X)$,

$Z(X) = Z(Y)$, $Y(Z) = X(Z)$. Подібну ситуацію, було розглянуто вище.

5. Виявлення прихованої бінарної змінної в дискретних моделях

Як відомо [2, 3, 10], факторизація коефіцієнта кореляції дає змогу виявляти латентні змінні (дійсного типу). Аналогічно, інструмент квазі-лінеарності залежностей стає засобом виявлення прихованих бінарних змінних у дискретних доменах, зокрема, коли ця змінна є центральною у структурі "зірка з трьома променями". Подібні моделі відомі під назвою latent class model. Звертаємо увагу, що відомі методи відкриття латентної змінної (дійсного типу [3, 10]) потребують принаймні чотирьох індикаторних змінних (ефектів), натомість як запропонована тут техніка задовольняється трьома ефектами.

Автентична (генеративна) структура моделі, яку розглядаємо (рис. 2), належить до класу дерев, але якщо приховати центральну (вузлову) змінну H, то спостережувана структура моделі стане як "трикутник" ребер. (Зауважимо, що "трикутник" виходить за рамки класу монопотокових моделей [9, 11]. А взагалі автентична структура моделі може бути "полі-деревом", і тоді після приховування змінної H можемо отримати двобічно-орієнтовані ребра [3].) "Видимі" парні залежності між індикаторними змінними утворюють співвідношення (23). Тож виконання тріад-стримування є свідченням і необхідною умовою для того, аби стверджувати, що попарні асоціації поміж трьома дискретними змінними пояснюються існуванням прихованої бінарної змінної, яка виступає спільною причиною індикаторних змінних. Проте перш ніж робити такий висновок, необхідно попередньо перевірити і спростувати альтернативні пояснення цього свідчення (див. розділ 4). Зокрема, у випадку бінарності всіх змінних тріад-стримування (23) виконується незалежно від структури моделі, отже, воно в такому разі не свідчить про якусь певну структуру (з латентною змінною чи без). Отже треба переконатися, що:

- індикаторні змінні не є бінарні;
- не чинна умовна незалежність двох змінних за кондиціонування третьої (бінарної);
- немає статистичної тотожності двох змінних.

За цих застережень тріад-стримування як критерій оснащує відкривання автентичної структури моделі у формі "зірки", де вузлова бінарна змінна є прихована. Звісно, цей критерій не є достатнім, проте вичерпно достатні умови такого відкриття знайти, мабуть, взагалі неможливо.

Нехтування можливими спеціальними випадками чинності тріад-стримування (моделями зі зв'язаними параметрами – див. розділ 4) заради моделі, стабільної до коливань значень параметрів, тримається загальновизнаної традиції статистичного моделювання. Це цілком відповідає припущенню необманливості, яке прийнято в методології ідентифікації графових статистичних моделей залежностей [2, 9].

Порівнюючи статистичні моделі, беруть до уваги не лише емпіричну точність, але і складність моделі, яку можна оцінити кількістю параметрів. Нехай кардинальність індикаторних змінних буде відповідно r, q, s. Тоді модель "зірка з прихованим бінарним вузлом" (рис. 2) має $2(r + q + s – 3) + 1$ вільних параметрів. А у моделі без прихованої змінної, де три змінні попарно поєднані ребрами (модель-"трикутник"), кількість вільних параметрів дорівнює $rq(s – - 1) + r(q – 1) + r – 1$. Тож, відносна складність цих моделей оцінюється приблизно як $2(r + q + s) – 5$ проти $rqs – 1$. Модель з прихованим вузлом є багаторазово простіша.

У деяких ситуаціях можна розрізнити випадки з бінарним прихованим вузлом та випадки з небінарним (мультіноміальним) вузлом. Дійсно, коли дві змінні пов'язані лише через бінарну змінну, то величина їхньої асоціації обмежена зверху і не може досягати абсолютного максимуму. Наприклад, для тризначних індикаторних змінних, поєднаних через бінарну змінну, максимум ненормалізованого індексу детермінації дорівнює 8/3, тоді як детерміністична залежність дає величину 4. Отже, якщо тріад-стримування виконується, і до того ж принаймні для однієї пари тризначних змінних індекс детермінації перевищує 8/3, то зро-

зуміло, що прихована вузлова змінна (якщо вона існує) не є бінарною.

У загальному випадку критерій тріад-стримування є придатним для виявлення прихованої бінарної змінної. Покажемо ефективність критерію на прикладі моделі з тризначними змінними. А саме, покажемо, що коли прихована змінна є тризначною, то в загальному випадку тріад-стримування не виконується, навіть коли структура моделі є "зіркою", як на рис. 2. Нехай маємо модель рис. 2, б з тризначними змінними *X*, *Y*, *Z* та *H*, з такими значеннями параметрів:

$$p(H) = (0.7; 0.2; 0.1),$$
$$p(X|h_1) = (0.6; 0.2; 0.2),$$
$$p(X|h_2) = (0.2; 0.6; 0.2),$$
$$p(X|h_3) = (0.6; 0.2; 0.2),$$
$$p(Y|h_1) = (0.6; 0.2; 0.2),$$
$$p(Y|h_2) = (0.2; 0.6; 0.2),$$
$$p(Y|h_3) = (0.2; 0.6; 0.2),$$
$$p(Z|h_1) = (0.8; 0.2; 0.0),$$
$$p(Z|h_2) = (0.2; 0.6; 0.2),$$
$$p(Z|h_3) = (0.2; 0.2; 0.6).$$

Тоді ліва частина тріад-стримування (23) буде дорівнювати 0.076, а права частина (23) – відповідно 0.057. Розбіг лівої та правої частин тріад-стримування сягає 25%. Таким чином, кардинальність прихованої змінної є суттєвою для чинності тріад-стримування.

Отже, інструмент тріад-стримування достатньо помітно дискримінує моделі і, , здатен виявляти приховану бінарну причину кількох індикаторних змінних (трьох, двох з трьох, трьох з чотирьох і т. д.). Звичайно, для цього потрібно мати відбірку даних відносно великого обсягу. (Відбірка даних збирається за стандартного припущення i.i.d. За наявності великої відбірки даних навіть припустима селекція даних, звісно, коли змінна селекції не є нащадком двох індикаторних змінних [2, 9]). Дані практично задовольняють критерій тріадстримування, якщо рівність (23) виконується з точністю до статистичної похибки.

Практично питання вибору моделі зводиться, з одного боку, до статистичної правдоподібності емпіричного відхилення обчислених оцінок від тріад-стримування (23), а з іншого, – до статистичної правдоподібності випадкового (не зумовленого структурою моделі) приблизного виконання тріад-стримування в даних, коли модель не є "зіркою з прихованим бінарним вузлом". Надійність ідентифікації латентної змінної зростатиме, коли збільшується кількість індикаторних змінних та їх значність.

Доречно зазначити, що опис моделі у вигляді "зірки з прихованим вузлом" (навіть коли автентична модель має іншу структуру) є зручною формою репрезентації, бо підвищує обчислювальну ефективність використання моделі в задачах висновків від свідчень [12].

Висновки

1. Показано, що на ланцюгових (деревовидних) структурах залежностей, де проміжна (сепараторна) змінна є бінарною, а інші змінні є дискретними з довільною значністю, виконується квазі-лінеарність залежностей. Це співвідношення є аналогом факторизації (декомпозиції) коефіцієнта кореляції в лінійно-нормальних системах залежностей. Але для цього залежність в дискретних моделях треба вимірювати за допомогою індексу детермінації (замість коефіцієнта кореляції).

Квазі-лінеарність залежностей може бути використана як емпіричне свідчення умовної незалежності в моделі (тобто як свідчення певної структури), коли відповідна медіаторна змінна є бінарною. Перевагою такого способу перевірки умовної незалежності є використання лише парних статистик (спосіб можна застосовувати навіть коли оцінка сумісного розподілення усіх змінних є недоступна).

2. Застосування індексу детермінації для аналізу структури у формі "зірка з бінарним вузлом" дозволило отримати характерне співвідношення парних залежностей, назване "тріад-стримуванням" (яке не містить параметрів центрального вузла). Показано, що (за наявності достатньо великого обсягу даних) тріад-стримування може слугувати інструментом емпіричного виявлення прихованої бінарної змінної, яка відіграє роль посередника взаємозв'язків між трьома (чи більше) індикаторними змінними, зокрема, є їхнім спільним фактором ("причи-

ною"). Проте, для виправдання такого висновку необхідно перевірити, верифікувати чи спростувати декілька варіантів альтернативного пояснення поведінки даних, зокрема, що індикаторні змінні не є бінарними, не чинна умовна незалежність двох змінних за кондиціонування третьої (бінарної), немає статистичної тотожності двох змінних тощо.

1. *Прикладная статистика*: Исследование зависимостей / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1985. – 487 с.
2. *Scheines R., Spirtes P., Glymour C., Meek C., Richardson T.* The TETRAD Project: Constraint Based Aids to Causal Model Specification // Multivariate Behavioral Research, (1998), **33**. – N 1. – P. 65 - 118.
3. *Scheines R., Spirtes P., Glymour C., C.Meek, T. Richardson.* TETRAD 3: Tools for Causal Modeling. User's Manual. – CMU, dep. Philosophy, Pittsburgh, PA, 2000.
4. *Балабанов А.С.* К проблеме вывода знаний о структуре зависимостей между переменными из данных большого объема в условиях помех // Материалы 2-й Междунар. конф. "УкрПРОГ'2000". – "Проблемы программирования", 2000. – № 1-2. – С. 527 - 535.
5. *Балабанов О.С.* Індуктивне відтворення деревовидних структур систем залежностей // Проблемы программирования, 2001.– № 1-2. – С. 95 - 108.
6. *Балабанов А.С.* Мера для обнаружения зависимостей между переменными в данных в условиях случайных возмущений // Проблемы программирования. – 1999. – № 2. – С. 63 - 69.
7. *Балабанов О.С.* Критерії ідентифікації ймовірнісних залежностей в базах даних // Праці 1-ї Міжнар. наук.-практ. конф. з программування (УкрПрог'98). – К.: 1998, 2-4 вересня, – С. 380 - 382.
8. *Кендалл М., Стьюарт А.* Статистические выводы и связи. – М.: Наука, 1973. – 900 с.
9. *Балабанов А.С.* К выводу структур моделей вероятностных зависимостей из статистических данных // Кибернетика и системный анализ, 2005. – № 6. – С. 19 - 31.
10. *Scheines, R., Spirtes, P.* Finding latent variable models in large databases // Intern. J. of Intelligent Systems. – (1992), **7**, – № 7, P. 609 - 621.
11. *Балабанов А.С.* Индуктивный метод восстановления монопотоковых вероятностных графовых моделей зависимостей // Проблемы управления и информатики. – 2003. – № 5. – С. 75 - 84.
12. *Pearl J.* Fusion, Propagation and Structuring in Belief Networks // Artificial Intelligence. – 1986. – **29,** N 3. – P. 241 - 288.

Отримано 22.02.2006

Про автора:

Балабанов Олександр Степанович,
старший науковий співробітник,
канд. техн. наук.

Місце роботи автора:

Інститут програмних систем НАН України,
проспект Академіка Глушкова, 40.
03680, Київ-187, Україна.
Тел. (044) 526 6249
E-mail: bas@isofts.kiev.ua
факс: (38044)526 6263