Доповіді
НАЦІОНАЛЬНОЇ
АКАДЕМІЇ НАУК
УКРАЇНИ

9 • 2008

ІНФОРМАТИКА ТА КІБЕРНЕТИКА

УДК 007:681.3.00

Академік НАН України **П. І. Андон, О. С. Балабанов**

До відкриття латентного бінарного фактора в статистичних даних категорного типу

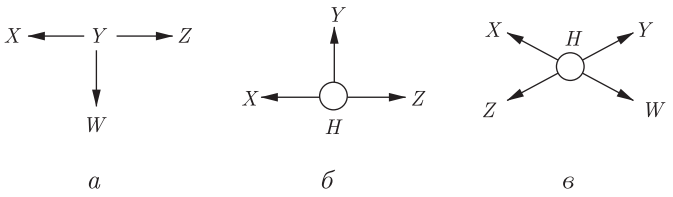
For a discrete model with tree-like structure, we demonstrate that if a separating variable (a root vertex) is binary, then constraints (like "tetrad difference") hold. Specifically, when there are four or three manifest variables in a model, the "ditetrad-constraint" or "triad-constraint", respectively, applies. So, these constraints facilitate the discovery of a hidden binary variable (latent class) which is responsible for associations among discrete manifest variables.

1. Моделі з латентними змінними (факторами) доволі поширені в соціометриці, економетриці, психометриці, медико-біологічних дослідженнях тощо. Практика аналізу емпіричних даних потребує методів швидкого виявлення латентних факторів. Для введення (включення) в модель прихованої змінної треба знайти відповідні свідчення в аналізованих даних. Структура моделі вважається адекватною, якщо вона сумісна з даними і не складніша за альтернативні структури, сумісні з даними. Модель M^* з прихованою змінною можна назвати обгрунтованою, якщо вона простіша за всі інші моделі M в заданому класі (як з прихованими змінними, так і без) і не поступається їм за узгодженістю з даними. Складність моделі в нашому випадку доцільно визначати кількістю незалежних параметрів.

Для виявлення прихованої змінної в лінійних моделях застосовують інструмент "тетрад-різниць" (тетрад-стримування), який був винайдений ще на початку 20-го століття [1]. Більш сучасний виклад можна знайти в [2–5]. Лінійні системи залежностей мають цікаву властивість — мультиплікативність (факторизованість) коефіцієнта кореляції на ланцюгових (деревовидних) структурах залежностей. Тобто для ланцюгів у структурі лінійних моделей коефіцієнт кореляції для транзитивної залежності дорівнює добутку коефіцієнтів кореляції для складових ланок, які утворюють ланцюг [3, 4]. Наприклад, у моделі зі структурою рис. 1, *а* змінні X, Z, W асоційовані лише через "центральну" змінну Y, отже кондиціонування змінної Y робить інші змінні взаємно незалежними. Зокрема, є чинною умовна незалежність змінної Z від змінної X за кондиціонування змінної Y, звідки маємо

$$r_{XZ} = r_{XY} \times r_{YZ}. \tag{1}$$

Завдяки такій властивості в моделі вигляду "4-зірка" з чотирма індикаторними змінними, які асоційовані тільки через єдину центральну змінну, виконуються "тетрад"-рівності.



а *б* *в*

Рис. 1

Наприклад, для лінійної моделі, яка зображена на рис. 1, *в*, маємо:

$$r_{XY} \times r_{ZW} = r_{XZ} \times r_{YW}, \qquad r_{XZ} \times r_{YW} = r_{XW} \times r_{YZ}, \tag{2}$$

а також третє рівняння, що випливає з цих двох.

Крім того, в лінійній моделі вигляду "зірка" є чинними обмеження типу нерівність, відомі як "співвідношення трикутника". Зокрема, для моделі рис. 1, *б* та 1, *в* виконується

$$|r_{XZ}| \geqslant |r_{XY} \times r_{YZ}|, \tag{3}$$

та аналогічні співвідношення для інших пар змінних.

Стосовно моделей для дискретних даних, переважна частина результатів обмежується бінарними змінними. Відомо [6], що у моделях вигляду "зірка" з бінарними змінними, де є чотири індикаторні змінні (рис. 1, *в*), теж виконується система рівнянь тетрад (2). А коли розглядаються лише три індикаторні змінні в моделі "зірка", виконується нерівність "трикутника" (3). І це зрозуміло, оскільки в моделях з бінарними змінними є чинною мультиплікативність міри залежності на ланцюгах, аналогічно лінійним моделям [7].

Аналогічний результат показано в [8, 9] в апараті категорних змінних. Для цього замість коефіцієнта кореляції застосовано коефіцієнт номінальної стохастичної детермінації $d_{\mathrm{nom}}(*)$ (див. далі). Тобто для ланцюгової структури X–Y–Z з бінарними змінними є чинним співвідношення

$$d_{\mathrm{nom}}(Z \leftarrow X) = d_{\mathrm{nom}}(Y \leftarrow X) \times d_{\mathrm{nom}}(Z \leftarrow Y).$$

Будемо називати цю властивість квазілінеарністю моделі, або квазілінеарністю залежності на ланцюгу.

Для загального випадку мультиноміальних (категорних) даних відомо мало результатів, які можна порівняти з такими для лінійних моделей. Лише нещодавно нами показано [10], що властивість квазілінеарності залежності поширюється на випадок, де "крайні" змінні ланцюга є мультиноміальними. Таким чином, пропонований нами аналітичний інструментарій заповнює певну "прогалину" і підсилює можливості аналізу даних.

2. Квазілінеарність на ланцюгах залежностей. Запропонований в [11, 12] коефіцієнт номінальної стохастичної детермінації (NSD-коефіцієнт) вимірює силу залежності між двома випадковими змінними мультиноміального (категорного) типу. Цей коефіцієнт визначається як

$$d_{\mathrm{nom}}(Z \leftarrow X) = C_N \sum_x \sum_y \left(p(Y|X) - \frac{1}{\|X\|} \sum_x p(Y|X) \right)^2, \tag{4}$$

де C_N — нормалізаційний коефіцієнт; $\|X\|$ — значність змінної X.

Трохи згодом була пропонована інша форма (версія) NSD-коефіцієнта, яка відрізняється використанням модуля замість квадрата [9, 10]. Модуль-форма коефіцієнта номінальної детермінації, яку будемо називати індексом детермінації, визначається як

$$di(Y(X)) = C_{Na} \sum_x \sum_y \left| p(Y|X) - \frac{1}{\|X\|} \sum_x p(Y|X) \right|. \tag{5}$$

Для бінарних змінних індекс детермінації $di(*)$ спрощується до вигляду

$$di(Y(X)) = |p(y_1|x_1) - p(y_1|x_2)| = |p(y_1, x_1) - p(x_1)p(y_1)|/p(x_1)p(x_2). \tag{6}$$

(Нормалізаційний коефіцієнт C_{Na} в цьому випадку дорівнює $1/2$.)

Нехай маємо ланцюгову структуру залежностей у вигляді X–Y–Z, тобто змінна Z умовно не залежна від змінної X за кондиціонування змінної Y; позначаємо це як $\Pr(X \perp Y \perp Z)$. Покажемо, що у випадку, коли медіаторна змінна Y — бінарна ($Y := \{y_1, y_2\}$), а змінні X, Z — дискретні (категорні), виконується квазілінеарність залежності.

Для зручності у подальшому будемо застосовувати ненормалізовану форму індекса детермінації (відкинемо коефіцієнт C_{Na}) і позначимо ненормалізований індекс $d(Y(X))$. Тоді визначаємо

$$d(Y(X)) = \sum_{x,y} \left| p(Y|X) - \frac{1}{\|X\|} \sum_x p(Y|X) \right|. \tag{7}$$

Тепер формулюємо базовий результат.

Твердження. *Якщо змінна Z умовно не залежна від X за кондиціонування змінної Y, змінна Y — бінарна, а змінні X та Z — дискретні довільної значності, то*

$$d(Z(X)) = \frac{d(Y(X))d(Z(Y))}{2}. \tag{8}$$

Доведення. Спростимо вираз для $d(Y(X))$, коли медіаторна змінна Y — бінарна, а змінна X — q-значна. Тривіальні маніпуляції дають

$$d(Y(X)) = \sum_{x,y} \left| p(Y|X) - \frac{1}{q} \sum_x p(Y|X) \right| = 2 \sum_x \left| p(y_1|X) - \frac{1}{q} \sum_x p(y_1|X) \right|. \tag{9}$$

Запишемо вираз для індексу детермінації $d(Z(Y))$ з бінарною Y:

$$d(Z(Y)) = \sum_{z,y} \left| p(Z|Y) - \frac{1}{2} \sum_y p(Z|Y) \right| = \sum_z |p(Z|y_1) - p(Z|y_2)|. \tag{10}$$

Згідно з визначенням (7), маємо

$$d(Z(X)) = \sum_z \sum_x \left| p(Z|X) - \frac{1}{q} \sum_x p(Z|X) \right|. \tag{11}$$

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

$$p(Z|X) = \sum_y [p(Z|Y)p(Y|X)] = p(y_1|X)[p(Z|y_1) - p(Z|y_2)] + p(Z|y_2). \tag{12}$$

Підставляючи (12) до (11), отримуємо

$$d(Z(X)) = \sum_{z,x} \left| p(y_1|X)[p(Z|y_1) - p(Z|y_2)] - [p(Z|y_1) - p(Z|y_2)] \frac{1}{q} \sum_x p(y_1|X) \right| =$$

$$= \sum_z |p(Z|y_1) - p(Z|y_2)| \sum_x \left| p(y_1|X) - \frac{1}{q} \sum_x p(y_1|X) \right|. \tag{13}$$

З огляду на (9) та (10), останнє (13) дає потрібне (8).

Співвідношення (8) може також виконуватись в окремих випадках, коли медіаторна змінна Y не є бінарною, але залежність має спеціальну форму [9, 10].

3. Характеристика структури моделі. Подальші результати стосуються дискретних моделей з бінарною центральною (вузловою, або сепараторною) змінною та кількома "індикаторними" змінними навколо неї.

Спочатку розглянемо випадок з чотирма "індикаторними" змінними ("4-зірка"). У цій структурі (рис. 1, *в*) кондиціонування центральної змінної робить всі інші змінні взаємно незалежними. Отож, чинні відношення умовної незалежності $\Pr(X \perp H \perp Y)$, $\Pr(X \perp H \perp Z)$, $\Pr(X \perp H \perp W)$, $\Pr(Y \perp H \perp Z)$ і т. д. В цій моделі можна нарахувати шість ланцюгових фрагментів. Кожний ланцюг задовольняє умови твердження, і можемо записати рівняння (8) для кожного з відповідною підстановкою змінних (для обох напрямків).

Для стислості будемо писати $Y(X)$ замість $d(Y(X))$. Отже, отримуємо систему 12-ти рівнянь:

$$\begin{aligned} &X(Y) = \frac{X(H)H(Y)}{2}, \qquad Y(X) = \frac{Y(H)H(X)}{2}, \qquad Y(Z) = \frac{Y(H)H(Z)}{2}, \\ &Z(Y) = \frac{Z(H)H(Y)}{2}, \qquad Z(X) = \frac{Z(H)H(X)}{2}, \qquad X(Z) = \frac{X(H)H(Z)}{2}, \quad \dots \end{aligned} \tag{14}$$

і так далі.

Виключаючи з системи (14) члени зі змінною H, одержуємо такі шість рівнянь для чотирьох змінних:

$$X(Y)Z(W) = X(W)Z(Y), \tag{15}$$

$$X(Y)W(Z) = X(Z)W(Y), \tag{16}$$

$$Y(X)Z(W) = Y(W)Z(X), \tag{17}$$

$$Y(X)W(Z) = Y(Z)W(X), \tag{18}$$

$$Z(X)W(Y) = Z(Y)W(X), \tag{19}$$

$$X(Z)Y(W) = X(W)Y(Z). \tag{20}$$

Серед цих шістьох рівнянь незалежними (алгебраїчно) є п'ять. Тобто, будь-яке рівняння імплікується п'ятьма іншими.

Як бачимо, ми отримали обмеження, подібне до "тетрад", але кількість рівнянь тут маємо іншу: шість замість трьох (або, рахуючи незалежні, п'ять замість двох). Називатимемо обмеження (15)–(20) "дітетрад-стримуванням".

Тепер розглянемо випадок з трьома "індикаторними" змінними. Ця модель із структурою "3-зірка" показана на рис. 1, *б*. Модель містить три ланцюгових фрагменти, кожен з яких задовольняє умови твердження (коли H — бінарна). Відтак можемо записати рівняння (8) з відповідною підстановкою змінних. Знов пишемо $Y(X)$ замість $d(Y(X))$. Отже, отримуємо систему рівнянь:

$$\begin{aligned} X(Y) &= \frac{X(H)H(Y)}{2}, \qquad Y(X) = \frac{Y(H)H(X)}{2}, \qquad Z(X) = \frac{Z(H)H(X)}{2}, \\ X(Z) &= \frac{X(H)H(Z)}{2}, \qquad Z(Y) = \frac{Z(H)H(Y)}{2}, \qquad Y(Z) = \frac{Y(H)H(Z)}{2}. \end{aligned} \tag{21}$$

Особливість цієї системи рівнянь полягає в тому, що члени зі змінною H входять до системи парами, як відношення (пропорції). Таких членів маємо шість, як і рівнянь системи. Тривіальні алгебраїчні перетворення дозволяють позбавлятися цих членів. В кінцевому результаті члени зі змінною H взаємознищуються, тож система редукується до єдиного рівняння, вираженого лише через X, Y, Z:

$$X(Y)Y(Z)Z(X) = Y(X)Z(Y)X(Z). \tag{22}$$

Рівняння (22) називатимемо "тріад-стримуванням". Воно характеризує модель вигляду "3-зірка" з бінарним вузлом і констатує інваріантність добутку трьох парних залежностей до їх рівночасного реверсу.

Підкреслюємо, що лінійні моделі з аналогічною структурою не дають можливостей, що ми їх зараз показали, оскільки коефіцієнт кореляції є симетрична міра залежності, а індекс детермінації — ні. Тому наш результат не має прямого аналогу серед відомих.

Кількість вільних параметрів у розглядуваних моделях можна підрахувати за схемою, за якою параметри визначаються в баєсівських мережах [4, 13], тобто у формі умовних розподілів ймовірностей $p(Y|F(Y))$, де $F(Y)$ — "батьки" (безпосередні причини) змінної Y. Нехай кардинальність (значність) індикаторних змінних X, Y, Z, W буде відповідно q, r, s, t. Тоді модель "4-зірка" (рис. 1, *в*) з бінарною центральною змінною має $2(q+r+s+t)-7$ параметрів. Відповідно, модель "3-зірка" (рис. 1, *б*) з бінарною центральною змінною має $2(q+r+s)-5$ вільних параметрів.

4. Виявлення прихованої змінної. Розглянуті структури моделі (вигляду "зірка") належать до класу дерев. Але якщо приховати центральну змінну H, то спостережувана структура моделі постане як модель, насичена зв'язками. Так, модель "4-зірка" (рис. 1, *в*), якщо приховати центральну змінну H, буде виглядати як "кліка" з чотирьох вершин. Свідчення про "справжню" структуру моделі заховані у співвідношеннях між значеннями "видимих" парних залежностей індикаторних змінних і описані як дітетрад-стримування.

Для відкриття латентної змінної в лінійних моделях інструмент "тетрад" потребує принаймні чотирьох індикаторних змінних (ефектів). Контрастуючи з цим, запропонований нами інструмент (тріад-стримування) задовольняється трьома ефектами. Але, в разі бінарних ефектів, "тріад" вироджується в тотожність безвідносно до структури моделі [10]. (Це легко перевірити з огляду на формулу (6).) Відтак, для верифікації існування латентного фактора трьох бінарних ефектів залишається звернутися до нерівностей "трикутника". Але в більш загальних випадках критерій тріад-стримування працює. Зокрема (як було показано в [10] на числовому прикладі), коли індикаторні змінні — тризначні, тріад-стримування помітно відрізняє випадок з тризначним спільним фактором.

Пропоновані обмеження — дітетрад-стримування та тріад-стримування — є необхідними умовами адекватності вказаних моделей ("4-зірка" та "3-зірка"). І хоча ці умови можуть бути досить жорсткими, та все ж вони не є достатніми (втім, як і інші відомі умови).

Дійсно, кількість вільних параметрів у насиченій (необмеженій) моделі з чотирма змінними дорівнює $qrst - 1$. Це набагато більше, ніж в моделі "4-зірка" (попри те, що "зірка" має на одну змінну більше). Наприклад, коли змінні X, Y, Z, W — тризначні, необмежена модель має 80 параметрів, а "4-зірка" — 17. Ясно, що таку віддаль у складності неможливо знівелювати за допомогою п'яти рівнянь дітетрад-стримування. Навіть у випадку бінарних індикаторних змінних складність зіставлюваних моделей буде 15 проти 9. Аналогічне є вірним для випадку трьох індикаторних змінних. Необмежена модель з трьох тризначних змінних має 26 вільних параметрів, а модель "3-зірка" з тризначними індикаторами — 13. Вказану різницю у складності неможливо знівелювати за допомогою одного рівняння тріад-стримування. Отже, можна припустити, що існують альтернативні моделі, які теж задовольняють дітетрад-стримування або тріад-стримування відповідно. Відтак, маючи на меті застосувати (22) або (15)–(20) для верифікації моделі, треба розвідати альтернативні моделі.

Одна з альтернатив — це модель, де виконується умовна незалежність деяких двох змінних при кондиціонуванні третьої (бінарної) змінної. Наприклад, нехай є чинною умовна незалежність $\Pr(X \perp Y \perp Z)$, причому змінна Y — бінарна. Тоді з формули (8) випливає:

$$Z(X) = \frac{Y(X)Z(Y)}{2}, \tag{23}$$

$$\frac{Y(Z)X(Y)}{2} = X(Z). \tag{24}$$

Добуток (23) та (24) дає тріад (22). До речі, (23) та (24) є більш жорстким обмеженням, ніж (22), отже, ланцюг та "зірку" можна розрізнити.

Відомі також альтернативні моделі, які накладають спеціальні обмеження на параметри моделі, кардинальність змінних та форму залежностей [9, 10]. Перелік альтернатив є відкритий. Проте, якщо альтернативне пояснення емпіричних свідчень потребує не зумовлених структурою обмежень на значення параметрів, то таке пояснення треба розглядати як штучне і непереконливе. Тож виконання співвідношень (15)–(20) або (22) можна вважати вагомим свідченням того, що асоціації поміж спостережуваними дискретними змінними пояснюються впливом прихованої бінарної змінної, яка виступає їх спільним фактором. Надійність ідентифікації латентної змінної підвищуватиметься із зростанням кількості індикаторних змінних та їх значності.

При виборі моделі з-поміж альтернатив береться до уваги і емпірична точність, і складність. Отже, коли правдоподібність моделі з прихованим фактором лише незначно гірша, ніж необмеженої моделі, то перша отримає перевагу завдяки тому, що вона багаторазово простіша. За наявності досить великої відбірки даних дітетрад-стримування та тріад-стримування можуть бути інструментом емпіричного виявлення прихованої бінарної змінної (спільного фактору).

Таким чином, на ланцюгах (фрагментах деревовидних структур) залежностей, де вузлова (сепараторна) змінна є бінарною, а інші змінні є дискретними з довільною значністю, можна факторизувати (декомпозувати) транзитивну залежність згідно з відтинками ланцюга. Ми показали, що, завдяки цій властивості, для структури моделі у формі "зірка з бінарним вузлом" виконуються спеціальні обмеження типу рівність на добуток парних

залежностей індикаторних змінних. (Дані про вузлову змінну при цьому не використовуються.) Завдяки цим обмеженням — дітетрад-стримування та тріад-стримування — можна ідентифікувати приховану бінарну змінну, відповідальну за асоціацію трьох або більше дискретних змінних.

1. *Spearman C.* General intelligence, objectively determined and measured // Americ. J. of Psychology. – 1904. – **15**. – P. 201–293.
2. *Bollen K. A., K. Ting.* A tetrad test for causal indicators // Psychological Methods. – 2000. – **5**. – P. 3–22.
3. *Прикладная* статистика: Исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – Москва: Финансы и статистика, 1985. – 487 с.
4. *Scheines R., P. Spirtes, C. Glymour et al.* The TETRAD project: constraint based aids to causal model specification // Multivariate Behavioral Research. – 1998. – **33**, No 1. – P. 65–118.
5. *Scheines R., Spirtes P.* Finding latent variable models in large databases // Intern. J. of Intelligent Systems. – 1992. – **7**, No 7. – P. 609–621.
6. *Pearl J.* Fusion, propagation and structuring in belief networks // Artificial Intelligence. – 1986. – **29**, No 3. – P. 241–288.
7. *Danks D., Glymour C.* Linearity properties of bayes nets with binary variables / In: Breese, J., Koller D. eds. Uncertainty in Artificial Intelligence: Proceedings of the 17th Conf. (UAI-2001). Morgan Kaufmann, San Francisco. – P. 98–104.
8. *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" // Пробл. программирования. – 2000. – № 1–2. – С. 527–535.
9. *Балабанов О. С.* Індуктивне відтворення деревовидних структур систем залежностей // Там же. – 2001. – № 1–2. – С. 95–108.
10. *Балабанов О. С.* Квазілінеарність у дискретних моделях залежностей та відкриття латентного фактору трьох ефектів // Пробл. програмування. – 2006. – № 4. – С. 28–36.
11. *Балабанов А. С.* Мера для обнаружения зависимостей между переменными в данных в условиях случайных возмущений // Пробл. программирования. – 1999. – № 2. – С. 63–69.
12. *Балабанов О. С.* Критерії ідентифікації ймовірнісних залежностей в базах даних // Праці 1-ї міжнар. наук.-практ. конф. з програмування (УкрПрог'98). – Київ, 1998. – 2–4 вересня. – С. 380–382.
13. *Neapolitan R. E.* Learning bayesian networks. – Englewood Cliffs: Prentice-Hall, 2003. – 686 p.

Інститут програмних систем НАН України, Київ

Надійшло до редакції 19.03.2008