Українська Федерація Інформатики

Інститут кібернетики імені В. М. Глушкова НАН України

Вищий навчальний заклад Укоопспілки
«ПОЛТАВСЬКИЙ УНІВЕРСИТЕТ ЕКОНОМІКИ І ТОРГІВЛІ»
(ПУЕТ)

ІНФОРМАТИКА ТА СИСТЕМНІ НАУКИ (ІСН-2015)

МАТЕРІАЛИ
VI ВСЕУКРАЇНСЬКОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ ЗА МІЖНАРОДНОЮ УЧАСТЮ

(м. Полтава, 19–21 березня 2015 року)

За редакцією професора О. О. Ємця

Полтава
ПУЕТ
2015

УДК 004.855:519.216

БАЗОВАНІ НА НЕЗАЛЕЖНОСТІ МЕТОДИ ІНДУКЦІЇ КАУЗАЛЬНИХ МЕРЕЖ І СЕПАРАЦІЯ В ОРГРАФАХ

***О. С. Балабанов**, д.ф.-м.н., пров.н.с.*
Інститут програмних систем НАН України
bas@isofts.kiev.ua

Проблема і суть підходу. Задача полягає у відтворенні моделі зв'язків та впливів між змінними об'єкту, виходячи з статистичних даних спостережень (без апріорних знань про структуру моделі). Популярний традиційний метод – регресійний аналіз, – неспроможний розв'язати цю задачу через те, що невизначеним є вже сам формат постановки регресії (невідомо, які змінні вважати факторами для яких цільових змінних). Невідомим може бути й темпоральний порядок змінних, і тоді кількість можливих варіантів порядку змінних є факторіально (експоненційно) великою. До того ж модель, виведена регресійним методом, часто «викривлює» каузальну картину зв'язків внаслідок наявності прихованих змінних. Протягом двох останніх десятиріч була розвинена методологія виведення каузальних моделей з даних, яка долає вказані проблеми завдяки системному підходу до аналізу залежностей. Моделі виводяться у формі каузальних мереж (точніше, їх модифікації, що враховує невизначеність орієнтацій) [1, 2]. Структура моделі задається ациклонним орієнтованих графом (АОГ). АОГ-модель залежностей визначається як $(G,\boldsymbol{\Theta})$, де G – АОГ, а $\boldsymbol{\Theta}$ – сукупність локально заданих параметрів у формі умовних розподілень ймовірностей (або щільності) $p(X \mid \mathbf{B}(X))$, де $\mathbf{B}(X)$ – множина всіх батьків вершини X. (Батько відповідає безпосередній «причині».) Найбільш відомі різновиди АОГ-моделей: 1) баєсові мережі, тобто моделі з категорними (дискретними) змінними; 2) гаусові мережі, тобто лінійні моделі з неперервними змінними та нормальними дистурбаціями.

Нехай $\mathbf{U}$ – множина всіх змінних моделі; $\mathbf{A}$ – множина всіх дуг орграфу G, а $|\mathbf{A}|$ – їх кількість. Тоді в теоретичній постановці задача формулюється так: задано розподілення ймовірностей $p^*(\mathbf{U})$; знайти таку (G,Θ), що $|\mathbf{A}| \mapsto \min$ та $p(\mathbf{U}|G,\Theta) = p^*(\mathbf{U})$.

Оскільки на практиці задається вибірковий розподіл $\tilde{p}(\mathbf{U})$, який випадково відхиляється від генеративного $p^*(\mathbf{U})$, то вказана постановка задачі – неприйнятна. Зазвичай обирають наступну постановку: задано розподіл $\tilde{p}(\mathbf{U})$ (власне, вибірка даних); знайти модель (G,Θ) з максимумом критерію, наприклад, критерію BIC. (BIC обчислює правдоподібність моделі й «штрафує» за складність.) Така задача дуже важка, бо кількість структур моделі може бути астрономічною.

Винайдено інший, **базований на незалежності, підхід** до розв'язання задачі. Він оснований на марковських властивостях моделі. Структура АОГ-моделі накладає на розподіл $p(\mathbf{U}|G,\Theta)$ обмеження (типу рівність), інваріантні до параметризації моделі. А саме, за будь-якої форми параметризації в розподілі $p(\mathbf{U}|G,\Theta)$ будуть виконуватися умовні незалежності відповідного формату. Всі ці умовні незалежності можуть бути зчитані з графу моделі G за допомогою критерію d-сепарації [1, 3]. Нагадаємо, що дуга (ребро) графу G моделі репрезентує безпосередній зв'язок. Ребро – це дуга без специфікації напрямку. Якщо маємо умовну незалежність $\mathrm{Ind}(X,Y|\mathbf{S})$, то $\mathbf{S}$ зветься сепаратором для (X,Y).

Постановка задачі розкладається за етапами:

1. Відтворити сукупність ребер, тобто верифікувати ребро для кожної пари змінних, знаходячи сепаратор і трактуючи умовну незалежність як факт d-сепарації (відсутність ребра).

2. Ідентифікувати напрямки ребер (які стають дугами), спираючись на аналіз сусідніх зв'язків [2, 4].

3. Обчислити всі $p(X|\mathbf{B}(X))$, виходячи з $\tilde{p}(\mathbf{U})$ та встановлених $\mathbf{B}(X)$.

Таким чином, здійснено концептуальну декомпозицію задачі.

Замість перебору цілих моделей виконується перебір сепараторів. Це забезпечує зниження розмірності статистик.

Ключ до рішення. Якщо розв'язувати задачу на етапі «1» самостійно для кожної пари змінних, то у випадку існування ребра $(X - Y)$ доведеться виконати $\sum_{i=0}^{n-2}\binom{n-2}{i}$ перевірок незалежності для (X, Y), де n – кількість змінних. В такому разі переваги підходу будуть майже втрачені. Найвідоміший (базований на незалежності) алгоритм PC [2] прискорює виконання етапу «1» за рахунок звуження дерева перевірок для кожної пари змінних (X, Y), використовуючи встановлені на поточний момент факти відсутності ребер між X та іншими змінними (аналогічно – для Y). Для подальшого підвищення ефективності пошуку було запропоновано відсікати цілі сектори простору сепараторів, використовуючи «глибокі» властивості критерію d-сепарації [3, 4]. Доцільно відшукувати тільки локально-мінімальні сепаратори. Встановлено необхідні вимоги до члена локально-мінімального сепаратора. Ці вимоги імплікують правила відсіювання змінних зі списку кандидатів до складу сепараторів. Найбільш корисним є наступне правило.

Правило «відсторонення» кандидатів у сепаратор ('placing aside'): якщо в орграфі G вершина X d-сепарує Z та Y, то вершина Z не є членом жодного локально-мінімального сепаратора для пари (X, Y).

Зрозуміло, що для виведення моделі з даних застосовується емпіричний (статистичний) «зліпок» (counterpart) вказаного правила. Розроблено цілий комплект правил оптимізації пошуку сепараторів [3, 4]. Завдяки застосуванню таких правил індуктивне виведення моделі суттєво прискорюється, залишаючись асимптотично-коректним [4, 5].

Випробування методу і верифікація індуктивного потенціалу. Для реалістичної оцінки обчислювальної трудомісткості розроблених алгоритмів та перевірки їх здатності відтворювати адекватні моделі необхідно виконати експерименти. Маючи на меті повно дослідити можливості метода відтворювати адекватні моделі, то провідний принцип

експериментів – модель виводиться виключно з статистичних даних (без будь-яких апріорних знань про структуру). Генеративна модель має бути відома досліднику (аналітику), але невідома методу. Схема таких експериментів показана на рис. 1.

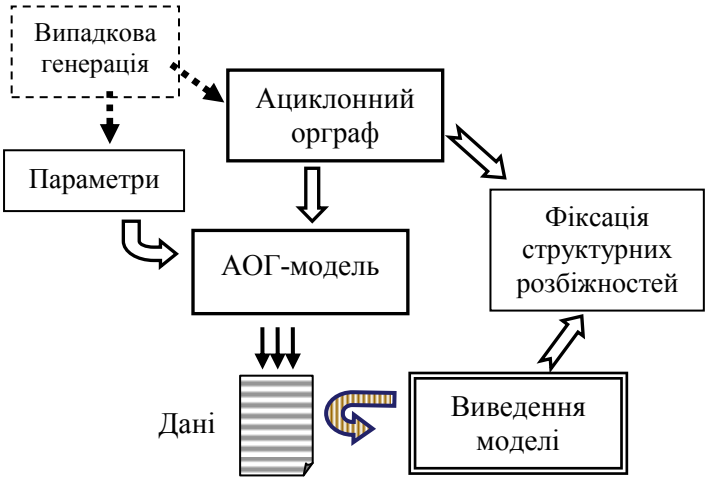

Рис. 1. Схема експериментів.

Аби уникнути підозри, що в метод закладено якісь знання про модель, аналітик мусить дотримуватися специфічної «рандомізації», коли й структура, й параметри моделі генеруються випадково. Для генерації АОГ задають тільки кількість вершин та кількість ребер. Структура моделі визначає номенклатуру параметрів. В разі гаусової мережі (і взагалі, при аналітичному описі фрагментів моделі) значення параметрів (тобто коефіцієнти відповідних рівнянь) беруться згідно заданого закону розподілення (наприклад, рівномірного) в заданому діапазоні. Для баєсової мережі задається значність змінних (наприклад, бінарні, тризначні, чотиризначні). Для них параметри генеруються наступним чином. Для кожної змінної генерується таблиця умовних розподілень. (Тут не може бути адитивного гамору.) Для кожного набору значень батьків існує своє умовне розподілення, яке генерується суто випадковим

чином, незалежно одне від інших. Генерується перша компонента розподілення для рядка таблиці (з рівномірного розподілення). Одиниця мінус отримане значення – це залишок на інші компоненти. З цього залишку генерується наступна компонента розподілення, і так далі. Звертаємо увагу, що такий суто випадковий механізм генерації параметрів створює дуже складну «картину» взаємодій факторів й «хаотичну» поведінку залежностей. Це ускладнює виведення моделі.

Інакший спосіб генерації параметрів баєсової мережі (який забезпечує певну змістовність й «прозорість» їх поведінки) полягає в наступному. Береться підходяща аналітична формула для функції залежності $Y = g(\mathbf{X})$ для кожної родини графу. Кожна функція має на один аргумент більше, ніж потрібно для моделі (ці додаткові аргументи позиціонуються як незалежні змінні). Для функцій генеруються значення параметрів. Генеруються дані. Додаткові аргументи відкидаються, дані дискретизуються.

В тезах та у доповіді подано принципи відтворення каузальних моделей з даних спостережень. Висвітлено методику випробувань і оцінки методів індуктивного виведення моделей.

Література

1. Pearl J. Causality: models, reasoning, and inference / J. Pearl. – Cambridge: Cambridge Univ. Press, 2000. – 526 p.
2. Spirtes P. Causation, prediction and search / Spirtes P., Glymour C., Scheines R. – New York: MIT Press, 2001. – 543 p.
3. Балабанов А. С. Формирование минимальных d-сепараторов в системе зависимостей / А.С. Балабанов // Кибернетика и системный анализ. – 2009. – № 5. – С. 38–50.
4. Балабанов О.С. Каузальні мережі: аналіз, синтез та виведення з статистичних даних / О.С. Балабанов / Автореферат дисертації на здобуття наук. ступеня док. фіз.-мат. наук. – Ін-т кібернетики ім. В.М. Глушкова НАНУ. – Київ, 2014.
5. Balabanov O.S. On perspectives of causal networks reconstruction by independence-based methods // Proceedings of the 4th Intern. Conf. on Inductive Modelling (ICIM'2013). – Kyiv, September 16–20, 2013. – Kyiv, Ukraine – P.139–142.