УДК 004.855:519.216 https://doi.org/10.15407/pp2019.03.058

О.С. Балабанов

ЗАДАЧІ ТА МЕТОДИ АНАЛІЗУ ВЕЛИКИХ ДАНИХ (ОГЛЯД)

Розглянуто основні задачі та методи глибокого аналізу великих даних. У викладі зроблено акцент на «фізичному» сенсі задач і методів, без математичних деталей. Спектр аналізу й використання великих даних охоплює чотири концептуальні класи завдань: «інтелектуальний» пошук інформації; масовану (конвеєрну) переробку даних; індукцію моделі об'єкту (середовища) та екстракцію знань з даних (відкриття закономірностей). Висвітлено суть типових класів задач великої аналітики: групування випадків (кластеризація даних); виведення ціле-визначених моделей (класифікація, регресія); виведення генеративних моделей; відкриття структур і закономірностей. Розглянуто ключові методи кластеризації, регресії та класифікації (включаючи глибоке навчання), а також виведення генеративних моделей. Методи розв'язання ціле-визначених задач поділяються на ті, що виводять модель у явному вигляді (модель «відокремлюється» від даних) та методи, «прив'язані до даних». Охарактеризовано особливості задач аналізу темпоральних даних (сегментація, виявлення точок зміни і т. д.). Детальніше викладено індуктивне виведення каузальних мереж методами, основаними на незалежності. Вказано особливості виведення динамічних каузальних мереж. Окремо підсумовано загальні особливості застосування статистичних методів у аналізі великих даних.

Ключові слова: великі дані, аналіз даних, виведення генеративної моделі, статистичні методи, кластеризація, регресія, прогноз, виявлення закономірностей, темпоральні дані, каузальні мережі.

Вступ

Стаття є другою частиною огляду аналітики великих даних. В першій частині [1] було розглянуто сфери застосувань великих даних, режими їх використання, основні напрямки, принципи, задачі глибокого аналізу великих даних та організацію циклу робіт з аналізу даних. У другій частині детальніше викладено задачі й методи глибокого аналізу даних.

Великі дані (ВД) характеризуються як масовані (сьогодні – це порядку 10^{21} байтів), різноманітні («строкаті»), неоднорідні, неструктуровані (чи погано структуровані), мінливі та «швидкі» (тобто такі, що швидко оновлюються чи поповнюються) [2–9]. Про глибокий аналіз доречно говорити тільки коли даних багато і вони багатовимірні. Основними сферами застосування аналітики великих даних є бізнес та наукові дослідження. Застосування у державному секторі подібні до бізнесових або наукових. Застосування у бізнесі зосереджуються переважно на задачах предикції (прогнозування), розпізнавання, виявлення трендів (тенденцій) та аномалій, сумарізації даних тощо. Застосування у наукових дослідженнях в першу чергу спрямовані на інтегративну (синтетичну) переробку експериментальних даних, виявлення закономірностей та зв’язків, генерацію та перевірку гіпотез, виведення моделей, які допомагають зрозуміти об’єкт дослідження. Образно кажучи, науковці шукають пояснень, узагальнень та знань, а бізнес цікавить, що відбувається, що буде і «що станеться, якщо ми зробимо так».

Основний режим використання ВД – глибокий аналіз даних, коли величезний масив сирої інформації «перетравлюється» і перетворюється на концентровану й цінну інформацію кінцевого споживання. Як кажуть, з даних «висмоктується» (екстрагується) їх цінний сенс. Впровадження великих даних неодмінно і безальтернативно («автоматично») передбачає застосування великої аналітики. Оскільки сфера великих даних та сфера великої аналітики взаємно доповнюють одна другу, доречно вести мову про формування єдиного річища, до якого зіллються дослідження і розробки, що охоплюють цикл діяльності від збору даних до вироблення інформаційного продукту кінцевого споживання. Тобто формується «наука» (в широкому розумінні) «Великі дані плюс Велика Аналітика» [3, 7–20].

В роботі [1] виділено наступні типи (концептуальні режими) використання ВД:

1) «інтелектуальний» пошук потрібної інформації (фактів) або запису, файлу, що містить ту інформацію;

2) масована (конвеєрна) переробка даних («відпрацювання» даних за один-два сканування);

3) індукція моделі об'єкту (джерела даних);

4) екстракція знань з даних (відкриття закономірностей та структур).

В процесі «інтелектуального» пошуку також застосовується аналіз даних, але результат пошуку по рівню абстракції є той самий, у якому дані зберігаються. (По-суті, результат є компіляцією фрагментів даних.) Гіпотетичний приклад запиту на «інтелектуальний» пошук інформації наведено в [1]. Режими використання даних «3» та «4» єдині в тому, що вони здійснюють узагальнення даних, тобто результатом є концентрована («кристалізована») інформація. Далі розглядаються методи розв'язання саме задач глибокого аналізу даних. Вважається, що дані вже підготовлені для аналізу. Сучасні методи аналізу даних та їх застосування викладено в [4, 10, 21–29]. Багато стандартних методів аналізу даних імплементовано у середовищах програмування **R**, Python, SAS, Matlab, Apache Mahout, Apache Spark і т. д.

Спектр задач аналізу даних та відкриття знань

Як аргументовано в [1], на виході інформаційної технології можна отримати тільки те, що містилося в масі даних перед переробкою (в «розчиненій», розпорошеній формі), а також було задано як апріорна інформація. Коли даних багато, внесок даних переважає. Успішність аналізу ВД визначається гармонійним взаємним доповненням даних, апріорної інформації та вдалою специфікацією завдання. Для розв'язання задач аналізу використовуються переважно статистичні методи. Протягом останніх кількох десятиліть карколомне зростання швидкодії комп'ютерів та обчислювальних систем стимулювало бурхливий розвиток статистичних методів обробки даних [23–29]. На передній план висунулися обчислювально-інтенсивні та комбінаторні методи – непараметричні, ітеративні та апроксимаційні, а також методи, вільні від припущень і форм розподілень. Широко застосовується техніка бутстрепінгу, згладжування кернелом, генерація вибірки за Гіббсом, максимізація очікування і методи MCMC [25–28]. Великі дані спричинили наступний поштовх розвитку статистичних методів.

Загальна спрощена система задач аналізу ВД показана на рис. 1. (Деякі з задач, представлених на рис. 1, можуть виступати етапом вирішення інших задач.) Типовий і лаконічний спосіб визначити акцент й мету завдання аналізу – вказати цільову змінну (характеристику, атрибут) y. В такому разі отримуємо ціле-визначену («націлену») задачу. Виведення ціле-визначених моделей – один з найпоширеніших класів задач аналізу даних. Задачі, в яких не задано цільової змінної, часто називають unsupervised learning [25, 26]. До таких задач відносять виведення генеративних моделей, кластеризацію і багато іншого. Коли інтерес аналітика не акцентовано на певній змінній, підходяща «загальна» задача – вивести генеративну модель для певного набору змінних X, тобто модель, яка відображає сумісне розподілення ймовірностей $p(X)$. Опис сукупності даних називаємо моделлю тоді, коли він компактний, поданий у зручній (наочній) формі й відображає головний вміст даних (відкидаючи випадкове і несуттєве). Подібні моделі $p(X)$ є генеративними моделями в слабкому сенсі. Іноді більш компактний і аналітичний опис розподілення даних $p(X)$ можна отримати через гіпотетичні змінні Z, так що $p(X) = \Omega(p_Z(Z))$, де $\Omega(\cdot)$ – певне стандартне перетворення. Подібну репрезентацію надають факторний аналіз, аналіз головних компонент тощо.

Серед методів виведення ціле-визначених моделей доцільно розрізняти методи, призначені безпосередньо для прогнозування (оцінки) значення цільової змінної, і методи, призначені для побудови самої моделі як синтетичного змістовного результату. Крім того, аналітика може цікавити навіть не вся модель, а тільки

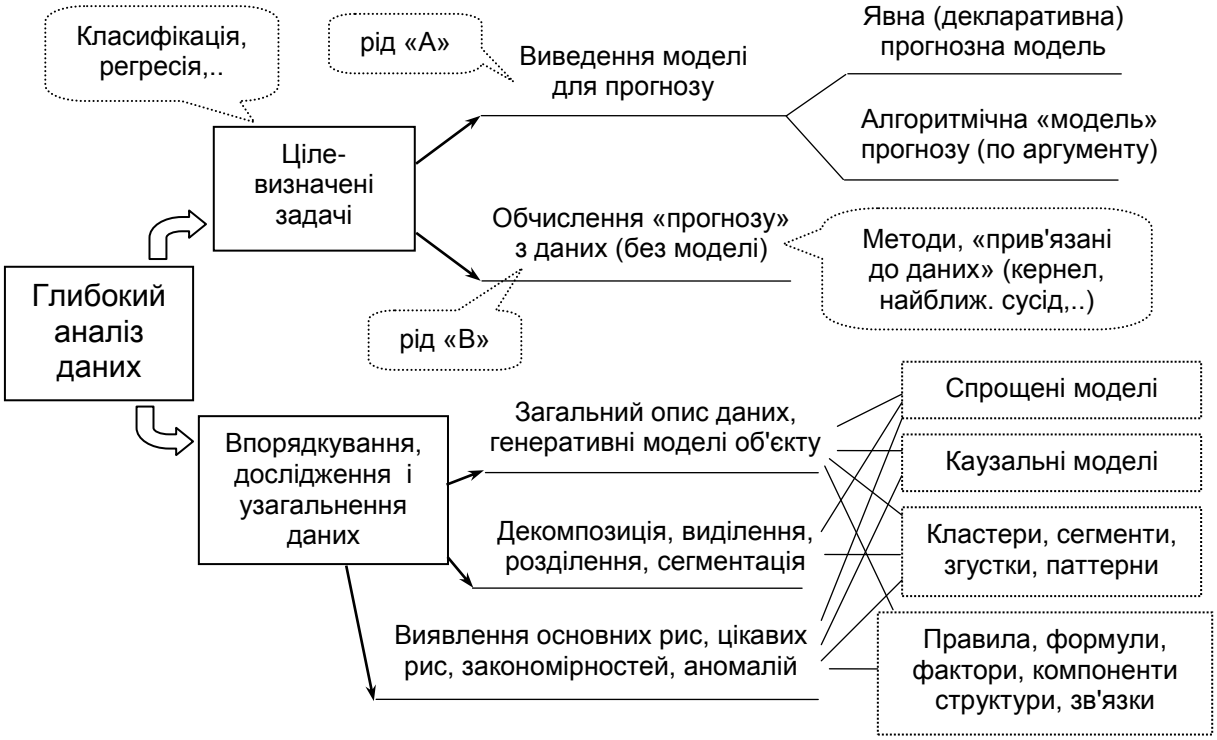

Рис. 1. Система задач великої аналітики

головні фактори (причини), які об'єктивно визначають значення вказаної змінної. Тут пролягає нечітка межа між задачами виведення моделей та задачами відкриття знань в даних. Якщо разом з отриманим описом моделі $p(X)$ виявлено цікаві, несподівані й статистично значущі «риси» моделі, доречно казати про відкриття знань. Навіть результати кластерного аналізу іноді можна кваліфікувати як відкриття знань, але за умови, що виявлені кластери є статистично значущими та чітко визначеними.

В результаті неакцентованого аналізу (дослідження) даних можна отримати закономірності в таких формах: послідовності, що повторюються (motifs); часто повторювані набори елементів (асоціації); структури залежностей; імплікативні зв'язки подій (можливо, нечіткі); періодичність коливань індикаторів у часі; інваріанти на сукупності значень характеристик (сталі співвідношення) тощо.

Традиційно на вхід статистичних методів подається статистична вибірка $\mathbf{X}$, тобто плаский масив, утворений зі записів-випадків (прецедентів) однакового формату. Зазвичай постулюється, що ці дані є статистичною вибіркою за схемою I.I.D. Але реальні дані можуть походити з різних «моделей». Розділити компоненти суміші в загальній ситуації важко. Класичним підходом до розділення (групування) випадків (записів) є кластерний аналіз. Сучасні практичні задачі стикаються з масивами даних, які не є класичною вибіркою. В багатьох ситуаціях між окремими записами даних є залежності у часі («післядія»). «Сирі» темпоральні дані з природного об'єкту не є I.I.D.-вибіркою. (Неможливо повернутися в минуле і повторити вимірювання.) Аналіз темпоральних даних охоплює зв'язки між різними записами даних.

Методи кластеризації

Кластерний аналіз має давню історію і продовжує розвиватися. Мета кластеризації – розбити множину прикладів (точок, записів даних) на кілька кластерів (груп, підмножин) так, щоб точки одного й того самого кластеру були значно подібніші (ближче) одна до одної, ніж точки, належні різним кластерам [23, 25, 26, 30]. (Зрозуміло, що розглядаються нетривіальні ситуації, коли задача не розв'язується сортуванням даних за значеннями одної

змінної чи за простим критерієм. Неодмінно треба розглядати одночасно кілька (багато) змінних, між якими існують залежності невідомого характеру.) Накопичено багатий арсенал методів кластеризації. Серед розмаїття цих методів можна виділити три підходи, або принципи: 1) кластеризація, основана на сукупній близькості прикладів; 2) кластеризація за принципом локальної близькості і множинного сусідства (зв'язності); 3) кластеризація на основі статистичної моделі розподілення даних [31].

Модель даних – це, зазвичай, суміш компонент, де кожна компонента описана параметрично-заданим розподіленням щільності ймовірності. Як правило, використовується нормальне розподілення. Критерієм вибору моделі є правдоподібність моделі. Відомий метод кластеризації, оснований на суміші моделей – AutoClass. Варіант систематизації методів кластеризації показано на рис. 2. Іншу (розгорнуту) класифікацію методів можна знайти в [30].

Традиційні методи кластеризації спираються на «середню» (або сукупну) відстань між точками. (Якщо точки розташовано у єдиному просторі, сукупна відстань обчислюється відносно центру кластеру.) Відстань є оберненою мірою щодо близькості; аналогічно, розбіжність обернена до подібності. У випадку категорних змінних аналітик задає матрицю відстані (розбіжності) для пар точок. Мабуть найбільш популярним методом, базованим на середній відстані, є відомий K-means. Цей метод, отримавши «ззовні» кількість кластерів, ітеративно повторює корекцію кластерів, чергуючи два кроки: 1) кожна точка даних присвоюється (приписується) тому кластеру, центр якого розташований найближче до точки; 2) для кожного кластеру обчислюється новий центр, використовуючи сукупність точок цього кластеру. Робота завершується, коли припиняється перерозподіл точок між кластерами. Достоїнство методу K-means –економічність обчислень (не треба обчислювати відстані для пар точок даних). Проте цей метод не дає задовільного результату у складних ситуаціях. Можна узагальнити принцип роботи методу K-means і розширити сферу застосування, відмовившись від обчислення метричних центрів кластерів, а замість центру використовувати один з членів кластеру (так, як це робиться у методі K-medoids). Тоді можна взагалі працювати без метричного простору прикладів і спиратися на попарно задану величину розбіжності прикладів.

Одні методи потребують, щоб кількість кластерів була задана на вході. Інші, більш гнучкі методи мають розбити дані на стільки кластерів, скільки їх «об'єктивно викристалізовується» згідно розташування точок. Зрозуміло, що аналітику ціка-

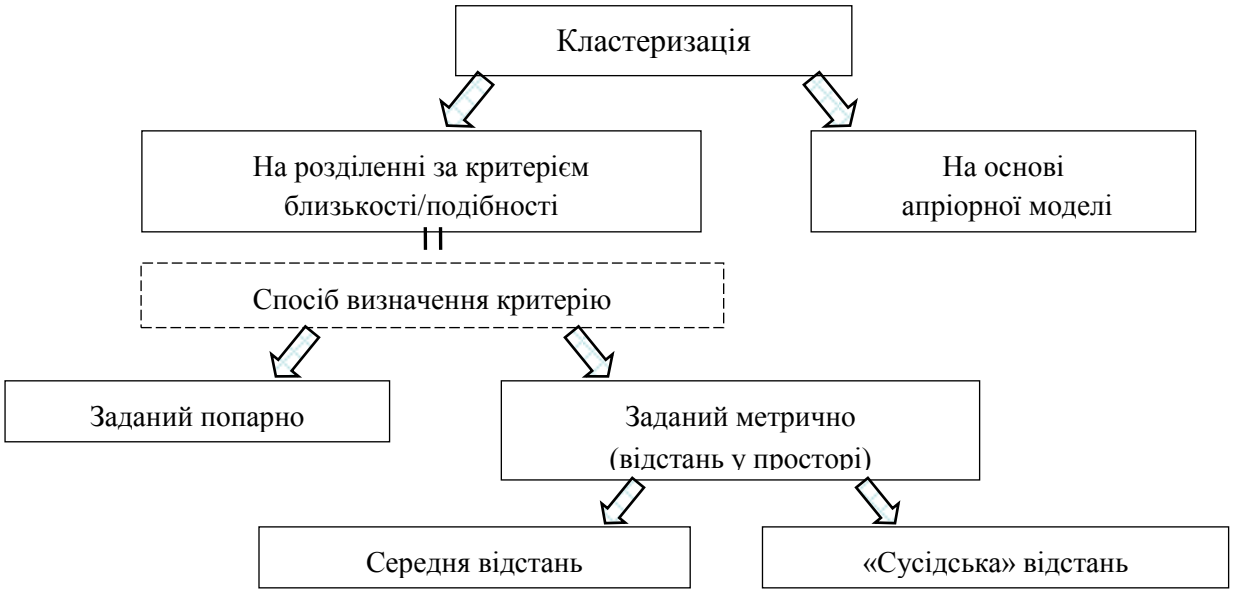

Рис. 2. Принципи кластеризації

ві такі кластери, які відображають об'єктивне групування (стратифікацію) прикладів, тобто відображають «контури» джерел даних (об'єктів). Відтак, аналітик зацікавлений у методах, які автоматично й обґрунтовано визначають кількість кластерів в даних. Можна уникнути довільного визначення кількості кластерів, звернувшись до ієрархічної кластеризації. Тотальне застосування ієрархічного принципу («до кінця») породжує дендрограму даних, де на нижньому «поверсі» кожний кластер представлено однією точкою, а кожний наступний поверх утворений злиттям (об'єднанням) двох найближчих кластерів.

Для вибору найкращого варіанту кластеризації потрібно визначити критерій якості. Універсального ефективного критерію не створено. Оцінка валідності результатів кластеризації залишається відкритим питанням. Хоча задача кластеризації інтуїтивно зрозуміла, в загальній постановці ця задача неоднозначна і не має єдиної теорії. Без додаткових уточнень і обмежень задача кластеризації погано поставлена. Наприклад, невизначеність задачі випливає з невизначеності міри близькості (відстані), коли різні змінні мають різні одиниці виміру або навіть відносяться до різних типів. На результати кластеризації критично впливає вибір (зміна) масштабів для змінних. Якщо маємо змінні різних типів, то взагалі немає єдиного простору змінних. Задавати величину відстані (розбіжності) для всіх пар точок може бути практично неприйнятно (особливо для великих даних).

Класичні алгоритми кластеризації стикаються з труднощами у випадках, коли кластери не є опуклими і не розмежовуються лінійно (тим більше – коли одні кластери оточують інші). В таких ситуаціях потрібні методи, основані на локальній близькості і зв'язності. Особливо складна ситуація – коли кластери перетинаються (частково суміщені). Слабкість традиційних методів в таких ситуаціях випливає з використання сферичної або еліптичної метрики. Один з підходів, що позбавлений цих обмежень – спектральна кластеризація. Суть цього підходу полягає в тому, що ті й тільки ті пари точок, які достатньо подібні (близькі) одна до одної, поєднуються ребром. Тоді задача ставиться як розбиття графу. Для розв'язання прикладних задач (головним чином для ситуацій з кластерами нестандартних форм) фахівці розробили багато евристичних алгоритмів кластеризації. Більшість тих методів спирається на відношення сусідства. Критерій об'єднання точок у кластер спирається на колективну локальну близькість та нерозривність (гущину) кожного кластера.

Достатню гнучкість у важких умовах (викривлених, не-опуклих кластерів) показав метод само-організуючих відображень, або мап (self-organizing map) [26]. Ідея само-організуючих мап (СОМ) полягає в тому, що у багатовимірному просторі, утвореному даними, будується двовимірна «різноманітність», яка апроксимує дані. Ця різноманітність формується як грати (у відповідних координатах) з $k \times m$ точок-прототипів. Дані скануються послідовно, для кожної точки даних знаходять найближчий прототип і зсувають його, щоб наблизити до точки даних. Але зсув стримується, щоб зберігати просторову гладкість двовимірної різноманітності і відношення сусідства прототипів. Для цього в СОМ підтримується певна кількість сусідів кожного прототипу. Сусідство визначається умовою, що відстань (у координатах гратів) не перевищує заданого порогу.

Для того, щоб результати кластеризації можна було сприймати як цікаві знахідки та знання, потрібно компактно описати кластери і показати статистичну значущість виділення саме таких кластерів. Кластеризацію можна розглядати як засіб розділення суміші даних на компоненти. Але зробити це буває важко або неможливо. Функції щільності ймовірностей компонент можуть значною мірою перетинатися. Тоді для ідентифікації компонентів (кластерів) потрібні методи, основані на моделі. Але ці методи спираються на апріорні відомості про характер компонент. Навіть після вірної ідентифікації компонент суміші однозначно розділити самі дані неможливо. Тобто кожний запис (приклад) можна віднести (приписати) до кількох компонент (з різною ймовірністю).

Кластеризація на основі суміші моделей нагадує виведення моделей з прихованими (латентними) змінними. Для кластеризації даних з пропусками застосовуються техніка максимізації очікування (EM). Коли обсяг даних надто зростає, відомі методи стають практично неприйнятними. Пом'якшити проблему можна, наприклад, за рахунок фокусування ітеративного процесу обчислень, уникаючи сканування частини даних [32].

Виведення ціле-визначених моделей

В задачах цього типу аналітик вказує одну цільову змінну y (відгук) серед змінних обраного масиву. В задачах регресії цільова змінна неперервна, а в задачах класифікації та розпізнавання – дискретна. Ціле-визначена задача виводить результат (модель) у формі $y = \Phi(X)$. (Строго кажучи, оскільки маємо $y \in X$, треба писати $y = \Phi(Z)$, де $Z \subseteq X \setminus \{y\}$.) Коли цільова змінна y дискретна, модель вигляду $y = \Phi(Z)$ називають «дискримінативною» (на противагу «генеративній» моделі $p(X)$). В разі дискретної цільової змінної модель може мати форму $p(y \mid X)$. Тоді зазвичай додається ще правило рішення $\hat{y} = \arg\max\{p(y)\}$.

Якщо метою є тільки вироблення прогнозу значення цільової змінної, то опис $\Phi(\cdot)$ може бути алгоритмом чи процедурою (і не мати декларативної або аналітичної форми). Результат у вигляді $y = \Phi(Z)$ або $p(y \mid Z)$ є предиктивною моделлю по формі, тобто у слабкому сенсі. Натомість предиктивними моделями у строгому сенсі є такі, які добре узагальнюють залежності і зберігають адекватність «на відстані» від точок оброблених даних, тобто достатньо точно екстраполюють залежності. Вимога до таких моделей – точні прогнози в усьому просторі (на всьому діапазоні) застосування.

Оскільки йдеться про аналіз великих даних, то є сенс систематизувати методи виведення ціле-визначених моделей згідно режиму використання даних. З цієї точки зору всі методи розділяємо на два роди: 1) рід «А» – «модельні» методи; 2) рід «Б» – «відкриті процедури» (методи, «прив'язані до даних»). Методи роду «А» надають завершений компактний загальний опис (модель у явному вигляді) $y = \Phi(Z)$. Загальна модель виводиться з даних один раз і надалі зберігається та застосовується «окремо» від тих даних. Для отримання прогнозу за допомогою такої моделі потрібно задати на вхід опису $\hat{y} = \Phi(Z)$ тільки значення «аргументу» (значення предикторів) Z_0. Натомість для отримання прогнозу із застосуванням методу роду «Б» використовується процедура вигляду $\hat{y} = \Phi(\mathbf{yZ}, Z_0)$, і потрібно задавати не тільки значення «аргументу» Z_0, але й кожний раз використовувати всі дані $\mathbf{yZ}$, на які спирається метод. Кожний раз дані обробляються заново. Методи роду «Б» не продукують моделі, а тільки обчислюють окремі значення $\hat{y}$ прямо з даних.

Методи роду «А» поділяються на ті, що виводять явну компактну модель $y = \mathrm{F}(Z)$ у декларативній або аналітичній формі (родина «А1»), та на ті, що надають лише алгоритмічний засіб $\hat{y} := \Phi(Z)$ для обчислення прогнозу, виходячи з значення «аргументу» (родина «А2»). Між методами родин «А1» і «А2» позиціонуються проміжні варіанти, коли маємо кілька явних моделей відповідно для кількох секторів (сегментів) простору значень «аргументу». В свою чергу, методи роду «Б» можуть використовувати спільну процедуру «налаштування», яка виконується один раз, так що її результати (параметри) потім застосовуються багатократно для обчислення прогнозу. Спільна процедура «налаштування» також може включати відбір необхідних (значущих) факторів (предикторів, ознак). Відбір значущих предикторів (регресорів) виконується майже всіма методами. Для багатьох прикладних задач (класифікації, розпізнавання) перед власне побудовою моделі формуються «ознаки», тобто нові змінні (підвищеного рівня порівняно з заданими на вході).

Створено великий арсенал методів виведення ціле-визначених моделей (зокрема, класифікації та регресії) [215-27].

Багато методів регресійного аналізу намагаються відтворити функцію $y = \mathrm{F}(X)$, незважаючи на те, що емпірична залежність не є однозначною функцією (внаслідок недоступності деяких факторів та завдяки домішку «гамору» в даних). Втім, іноді можна було би точно описати дані однозначною функцією $y = \mathrm{F}(X)$, але це означало би відтворювати випадковий гамір. Така модель не буде адекватною. Історично першим методом була лінійна регресія, яка виводить модель вигляду $\mathrm{E}(y \mid X) = b_0 + \mathbf{B}^{\mathrm{T}} \cdot \mathbf{X}$. Критерієм якості моделі (для налаштування коефіцієнтів $\mathbf{B}$) є мінімум суми квадратів відхилень (помилок). Зрозуміло, що у більшості практичних задач лінійна модель не дасть задовільного результату. Було пропоноване багато варіантів нелінійної регресії. Залучення все більш складних і гнучких форм залежності (наприклад, поліномів високого ступеня) дозволяє мінімізувати відхилення даних від прогнозу моделі. Але висока гнучкість (адаптивність) моделі веде до синдрому гіпер-специфікації, тобто «натяжки» (overfitting), з яким треба боротися. Наприклад, нехай для відгуку y маємо лише один регресор x. В такому разі будь-які дані **xy** (коли немає двох записів з однаковими значеннями змінної x) можна абсолютно точно описати моделлю $y = = c \cdot \sin(a + bx)$. Для цього налаштовуються три параметри a, b, c (завважте, значення b може бути великим). Але ясно, що така модель (будучи «точною» з точки зору оброблених даних) буде катастрофічно неадекватною для нових даних, навіть з того самого джерела. (Цей приклад також показує, що номінальна кількість параметрів далеко не завжди вірно характеризує складність моделі.)

Найбільш популярними способами стримування гіпер-специфікації є кросвалідація, регуляризація та процедури на основі бутстрепінгу [26, 27]. З точки зору аналітики критичним питанням виведення ціле-визначених моделей є підбір значущих предикторів (коваріат). Мабуть найбільш робастна процедура підбору предикторів (серед традиційних) – це двохетапна процедура: на першому етапі предиктори послідовно включаються в модель, а на другому – виключаються. Проте повного розв'язання ця задача знаходить тільки в апараті каузальних мереж, оскільки предиктори часто є умовно-інформативними [1]. Найбільш популярними критеріями якості моделі (для підбору складу предикторів) є *AIC* (або C_p) та *BIC* [27–29]. Інший підхід полягає у тому, щоб обмежувати величину коефіцієнтів регресії («стягувати» їх до нуля). Гребенева регресія та «Лассо» [26, 27, 29] включають в постановку задачі оптимізації моделі штрафи на розмір коефіцієнтів. Але гребенева регресія має тенденцію зменшувати коефіцієнти, але не видаляти терми з моделі. Натомість постановка задачі «Лассо» стимулює жорсткий відбір (відсів) підмножини предикторів. Тому «Лассо» дає простішу модель.

Зазвичай немає підстав розраховувати на те, що «істинна» модель має якусь аналітичну форму. Взагалі, апріорне визначення класу моделі часто виглядає волюнтаристським й непереконливим. Тому цілком закономірно виникла ідея іншого підходу до розв'язання ціле-визначених задач. Замість побудови єдиного математичного опису $\hat{y} = \mathrm{F}(X)$ відтворюють локальну залежність в межах ареалів (сегментів, ділянок) простору значень предикторів. До таких методів (родина «A1-Loc»), належать сплайни регресії. Аби локальні функції регресії з'єднувалися без розривів і зламів, достатньо застосувати кубічні сплайни [26, 29].

Протягом останніх 20-30-ти років було розроблено багато адаптивних методів відтворення ціле-визначених моделей, які вдаються до нових способів формування моделі. Виникли методи, що структурують (сегментують) простір предикторів, тобто розбивають простір змінних на ареали (квазі-прямокутники, квазі-паралепіпеди) відповідно до локальної поведінки залежності. Популярним способом адаптивної сегментації простору стала побудова дерев (дерева регресії, методи MARS, boosting tree, bagging, випадкові дерева) [26, 29]. Ці методи ієрархічно та ітеративно обирають оптимальне розбиття

підмножини даних на кожному кроці так, щоб сегменти простору охоплювали дані з близькими значеннями цільової змінної. Такі прості методи, як CART, C4.5, C5.0, утворюють розгалуження дерева, обравши чергову змінну та її певне значення. В кожному листі дерева змінна y визначається просто (наприклад, це відповідна константа). Для задачі регресії таке рішення породжує проблему – відсутність гладкості. Ця проблема дерев долається більш витонченими методами, наприклад, MARS (багатовимірні адаптивні регресійні сплайни). Метод MARS формує модель з «однобічних» лінійних базових функцій, які мають вигляд $b \cdot \max\{0,(X_j - t)\}$ та $b \cdot \max\{0,(t - X_j)\}$. Модель виводиться як сума обраних базових функцій з налаштованими параметрами t та b. Навіть такі прості базові функції дозволяють відтворювати нелінійні залежності. В разі недостатньої точності в модель додаються добутки базових функцій. За такою процедурою можна отримати дуже складну модель, тому на заключному етапі виведення виконується зворотній процес «підрізання» (спрощення) дерева. Деякі терми моделі видаляються, але так, щоб відхилення моделі від даних зростало на мінімальну величину. Оптимальний «розмір» моделі визначається за допомогою узагальненої крос-валідації [26, 27, 29]. Вдосконаленням методу CART є «ієрархічна суміш експертів». Додаткова адаптивність досягається за рахунок того, що розгалуження дерева утворюють згідно (лінійних) комбінацій змінних (тому «нарізка» простору – не прямокутна).

Хоча виведення дерева не породжує єдиної компактної моделі (у традиційному сенсі), але обґрунтована сегментація простору змінних придатна для інтерпретації та візуалізації і надає певне знання.

Якщо пріоритетом аналітика є точність предикції чи класифікації, то застосовують ансамблі моделей. Виводять набір «моделей-дерев» з квазі-вибірок, отриманих за допомогою бутстрепінгу. Для розгалужень в кожному дереві черговий предиктор обирають стохастично (щоб зменшити кореляцію та ухил). Прогноз обчислюється як середнє значення для всіх дерев (в задачах регресії) або обирається більшістю «голосів» (в задачах класифікації). Це дозволяє знизити дисперсію прогнозу. Але перехід від однієї моделі до ансамблю означає втрату наочності та «зрозумілості».

Протягом останніх 15 років серед методів класифікації значної популярності набув напрямок support vector machine (SVM) [25, 29]. Ці методи є розвитком ідеї класифікатора з максимальними полями. Такі методи спрямовані на побудову подільної (дискримінативної, сепараторної) поверхні або лінії, такої, щоб вона вірно розділяла всі точки на класи і, крім того, щоб відстань від сепараторної поверхні (лінії) до найближчих точок була якнайбільшою. Обчислювальна перевага методів виведення класифікаторів за критерієм максимальних полів випливає з того, що в задачі оптимізації враховуються тільки точки, що лежать на полях (а не всі). Для моделі support vector classifier вимога коректної класифікації всіх точок (даних) послаблюється (тому їх звуть класифікаторами з «м'якими» полями). Тобто достатньо вірно класифікувати не всі, а лише переважну більшість точок, але натомість треба додатково забезпечити робастність класифікації шляхом збільшення полів та спрощення (згладжування) поверхні сепарації. Точність класифікації моделями SVM підвищується за рахунок побудови нелінійної сепараторної поверхні з використанням кернелу (ядра).

Переглянемо методи, «прив'язані до даних», тобто методи роду «Б». В літературі їх іноді називають «базованими на пам'яті» (“memory-based”). Зрозуміло, що прогноз має спиратися на ближчі точки даних. Найпростіший спосіб використання принципу локальності – метод найближчого сусіда: для оцінки відгуку $\hat{y}$ в заданій точці X_0 береться значення y з того запису, у якому значення X є найближчим до заданого ($X \approx X_0$). Завважимо, що хоча цей метод не передбачає обчислень, у разі використання великих обсягів даних пошук ближчих сусідів може потребувати значних витрат. Уявімо, що

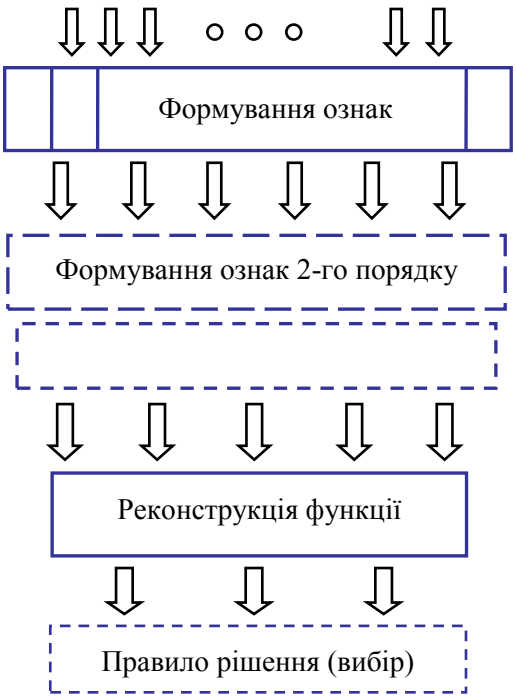

Рис. 3. Схема нейронної мережі для глибокого навчання

користувача цікавить прогноз в точці X_0, для якої є відповідний (i–й) запис даних (тобто маємо $X_i = X_0$). З огляду на загамованість даних значення цільової змінної з i–го запису даних далеко не завжди є кращим прогнозом. Вдосконалити метод найближчого сусіда (підвищити робастність) можна за рахунок того, що прогноз обчислюється як середнє значення y на основі кількох ближчих точок даних (з вагами). Ще гнучкіший метод – непараметрична регресія, – використовує згладжування залежності навколо X_0 за допомогою кернела. Явну модель не виводять (але на попередньому етапі виконується настройка параметрів, зокрема, «вікна» кернелу). Кожна чергова предикція потребує значних обчислень на основ даних (модель існує віртуально).

Локальна лінійна регресія поєднує традиційну регресію з принципом локального непараметричного згладжування. При цьому замість єдиної моделі настроюється функція для кожної цільової точки. Задача ставиться наступним чином [27]:

$$\hat{\beta}(x_0) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} K(x_0, x_i)(y_i - x_i\beta)^2 ,$$

де $K(\cdot,\cdot)$ – функція кернелу, а n – кількість записів даних. Хоча така постановка схожа на звичайну регресію, але модель (в строгому сенсі) не будується. Техніка лінійної регресії використовується тут для того, щоб уникнути зміщення (викривлення) прогнозів на краях діапазону даних. (Таке зміщення є типовим синдромом звичайного непараметричного згладжування.) Для того, щоб обчислення прогнозу не вимагало використання всієї вибірки даних **yZ**, можна використати функцію кернелу без «хвостів» або попередньо настроювати параметри «вікна» кернелу для окремих секторів простору. Але ефект такого вдосконалення може бути незначний, бо все одно треба переглядати всі дані, щоб відібрати частину даних, яка буде оброблена.

Техніку локальної лінійної регресії можна застосувати також в режимі виведення явної аналітичної моделі, але в локальному варіанті (родина «A1-Loc»). Тоді можна не повторювати обробку всіх даних кожний раз, а натомість зберігати набір локальний (частинних) моделей. В кожному ареалі предикторів обирається «представницька» точка, і модель, виведена для цієї точки, застосовується для всього ареалу. Для втілення такого режиму потрібно виконати розбивку простору предикторів на ареали (сегменти) і підтримувати засіб вибору відповідної локальної моделі.

Для розв'язання ціле-визначених задач, призначених, перш за все, для розпізнавання та класифікації, спільнота комп'ютерників давно розвиває техніку так званих нейронних мереж. Методи «навчання» нейронних мереж спочатку розробляли радше евристично, за аналогією до функціонування мозкових структур, яким його уявляли комп'ютерники. Тому ці методи традиційно позиціонували як один з напрямків «штучного інтелекту». Логічно віднести цей розділ досліджень і розробок до напрямку «самонавчання алгоритмів» [1]. Навчання нейронних мереж належить до методів родини «A2» (див. вище). Нейронні мережі продукують радше вміння, а не знання. Нейронні мережі є багато-параметричними статистичними

моделями, які мають кілька рівнів, де застосовуються лінійні перетворення, нелінійні функції та порогові функції [25–27]. З вхідних змінних формують інформативніші (синтетичні) «ознаки», з яких вже будується модель. Модель може включати багато рівнів. По-суті, техніка навчання нейронної мережі є варіантом виконання градієнтної оптимізації. Параметри конструкції «моделі» (кількість рівнів, блоків, ознак,..) іноді задаються, але можуть налаштовуватися за допомогою кросвалідації або підбиратися за допомогою експериментів.

Схематично конструкція нейронної мережі показана на рис. 3 (але треба мати на увазі, що сусідні рівні поєднані перехресними зв'язками). Кожний наступний рівень конструкції підвищує рівень абстракції (узагальнення) інформації. Навчання здійснюється як налаштування багатьох (до мільйонів) вагових коефіцієнтів моделі.

Варіант нейронної мережі може будуватися наступним чином. Кожна ознака формується за допомогою функції вигляду

$$z_i = \sigma(\alpha_{i,0} + \alpha_{i,1} \cdot x_1 + ... + \alpha_{i,m} \cdot x_m),$$

де $x_1, x_2, ..$ – вхідні сигнали, $\sigma(\cdot)$ – функція активації, наприклад, порогова або сигмоїдна (згладжена версія).

Вихідний результат виробляється за допомогою функції вигляду

$$y_k = g_k(\beta_{k,0} + \beta_{k,1} \cdot z_1 + \beta_{k,2} \cdot z_2 + ...),$$

де $g_k(\cdot)$– нелінійна функція, наприклад, подібна до $\sigma(\cdot)$, але з вільними параметрами, які налаштовуються. Для моделей класифікації в ролі $g_k(\cdot)$ може використовуватися функція softmax.

Коли мережа має кілька виходів, додається ще правило рішення (обрання). Навчання полягає у підборі коефіцієнтів $\alpha_{i,j}$ та $\beta_{k,j}$ з метою мінімізувати середньоквадратичне відхилення прогнозу від фактичних значень y_i.

Впродовж останнього десятиріччя у руслі нейронних мереж виокремилися нова хвиля розробок під назвою «глибоке навчання» [25–27, 33]. Ці засоби відрізняються від попередніх поколінь нейронних мереж більшою кількості рівнів конструкції та ускладненням форм перетворень. На старті процесу глибокого навчання задано «каркас» моделі, який розрахований на певний клас прикладних задач. Розрізняють кілька різновидів моделей глибокого навчання. «Конволюційні» мережі пристосовані для обробки образів, відео і мовлення, а «рекурентні» мережі – для послідовних даних [33]. В «конволюційних» мережах чергуються шари (рівні) згортки та злиття [27].

Фактори, що правдоподібно забезпечують успішність глибокого навчання (у відповідних впровадженнях), названо в [1]. В задачах глибокого навчання кількість вхідних змінних (предикторів) велика, але вони «дрібні». Сусідні змінні взаємно тісно корельовані.

Аналіз темпоральних даних

Зібрання даних з темпоральною прив'язкою прикладів та характеристик інтенсивно накопичуються. Спектр застосувань методів обробки й аналізу темпоральних та просторово-темпоральних типів даних надзвичайно широкий. Він охоплює такі різноманітні сфери, як біологія, медицина, астрофізика, економетрика, інтелектуалізація роботів і багато іншого. В формі темпоральних даних фіксуються фінансові та біржові індекси, аудіо- та відеозаписи, транзакції через систему тощо. Розв'язання прикладних проблем може потребувати цілого комплексу методів і технологій. Методи аналізу темпоральних даних все більше диференціюються й спеціалізуються відповідно до прикладних задач. Тут доцільно розглянути тільки базові задачі й методи.

Взагалі, в багатьох ситуаціях аналітик стикається з даними, об'єктивно вбудованими в певну просторово-часову структуру. Дані мають невід'ємні координати виміру. Спеціальним (одновимірним) випадком просторової структури є послідовність. Відмінність впорядкованості у часі від просторової послідовності можна пояснити за допомогою понять інерції (післядії) та спрямованості в одну сторону (у майбутнє). Завдяки часовому виміру ста-

ють практично важливими такі поняття, як форма сигналу, швидкість зміни, періодичність, спектр і т. д. З'являються вагомі підстави розділяти «корисний сигнал» та «гамір» у єдиному потоці даних.

Статистична методологія вже давно озброєна процедурами роботи з даними у формі часових рядів. (Відомі трансформації дозволяються привести такі дані до схеми I.I.D.-вибірки.) Дискретизація (у часі) неперервних процесів відбувається вже на етапі вимірювання. Багато методів розв'язання типових задач аналізу, розроблених для традиційних даних (зокрема, класифікація, кластеризація), були поширені на темпоральні дані за допомогою трансформацій та спрощення даних. Наприклад, виділяють фрагменти часового ряду і трактують їх як ознаки в традиційних задачах. Традиційні підходи до аналізу сигналів та неперервних процесів використовують різні варіанти згортки і фільтрації. В епоху великих даних попередня обробка темпоральних даних залучає широкий набір простих й практичних процедур – «нарізка» послідовності на фрагменти, масштабування, агрегація (спрощення, зменшення точності і деталізації). «Нарізка» є типовим засобом сформувати вибірку випадків з єдиного потоку даних. Іноді вибір варіанта нарізки на фрагменти має вирішальний вплив на результати подальшого аналізу, причому невідомо, як треба «вірно» нарізати дані. Задачі аналізу часових рядів залучають спеціальну відповідну техніку, зокрема, авторегресію (AR), moving average (MA), приховані марковські ланцюги (HMM), DTW [25, 34] і т. д.

Задача сегментації темпоральних даних та часових рядів широко розповсюджена і має варіанти з дуже різними постановками відповідно до прикладних цілей. Якщо йдеться про аналіз об'єктивного процесу (фізичного, економічного і т. д.), то мета може полягати в тому, щоб знайти фрагменти (попередньо невідомої довжини), які характеризуються стаціонарністю або певною динамікою зміни форми сигналу у часі. В роботі [34] головною метою сегментації часових рядів названо апроксимацію й зниження розмірності. Тому сегменти мають бути типовими (а їх номенклатура – мінімальною), щоб кожний тип сегменту описувався простою моделлю. Якщо маємо задачу розпізнавання аудіо мови, то такі показники, як швидкість, інерція, стаціонарність і тренд будуть несуттєвими. Але спрямованість послідовності даних у часі не можна відкинути. Для розпізнавання аудіо мови потрібно розділити ряд даних на фрагменти, які відповідають словам. Тоді ряд даних трактується як ланцюжок варіантів стандартних паттернів (фрагменти є репрезентантами класів еквівалентності). Така задача сегментації даних переплітається з класифікацією. Натомість якщо поставити мету розшифрувати аудіозапис невідомої мови, то спочатку, мабуть, треба застосувати кластеризацію та пошук motifs.

Взагалі, можна розрізняти типи темпоральних даних згідно природи генераторів даних. Типовими генераторами даних є: 1) природні процеси, 2) ергатичні системи (або цілеспрямовані об'єкти), 3) бібліотеки «паттернів» та записів. Деякі генератори даних останнього типу породжують такі послідовності символів, що для них вибір початку і кінця послідовності є питанням домовленості. Такі дані можна трактувати як одновимірний варіант просторових даних. Методи аналізу символьних послідовностей і виявлення аномалій в них оглянуто в [35].

Відомо кілька постановок задачі виявлення аномалій в послідовностях. Перша: виявити аномальний фрагмент в довгій суцільній послідовності. Друга: серед багатьох послідовностей виявити таку, яка значно відрізняється від решти. Третя: знайти послідовність, що містить деякий фрагмент (паттерн), який повторюється в цій послідовності значно частіше від очікуваного («нормального» чи середнього) рівня. В разі, коли розглядаються дані, генеровані природними процесами чи ергатичними системами, актуальною є задача прогнозу, тобто передбачення кількох найближчих майбутніх значень. Для цього застосовуються такі традиційні методи, як авторегресія, приховані марковські ланцюги тощо.

Задача виявлення точок змін є характерною саме для аналізу темпоральних даних (перш за все аналізу даних з природних процесів). Ця задача є різновидом задачі сегментації, сформульованим в інших поняттях. Точка зміни трактується як грубий, раптовий, різкий (швидкий), крутий перехід процесу від одної поведінки до іншої. Наприклад, зміну можна розуміти як перехід від одного стаціонарного режиму до іншого («переключення»). Створено багато методів виділення точок зміни [36–39]. В економетричному аналізі так звані структурні зміни («злами») розуміються як моменти, коли різко змінюються коефіцієнти лінійної залежності цільового часового ряду від інших (паралельних) часових рядів. Проблему виявлення точок зміни іноді формулюють як задачу вибору кращої статистичної моделі даних [36]. При цьому кожний фрагмент даних між двома точками зміни має бути достатньо однорідним, щоб добре описуватися простою локальною моделлю. Мета цієї задачі – оптимізація «сукупної» моделі за трьома вимогами: максимізація точності відтворення даних, спрощення всіх локальних моделей для фрагментів даних і зменшення кількості точок зміни.

Виявлення точок зміни є однією з задач відкриття знань. На перший погляд може здатися, що результат у формі точок зміни відображає лише окремі (поодинокі) події, тобто не має характеру регулярності, повторюваності, закономірності і узагальнення, що зазвичай асоціюється з поняттям знань. Насправді ця начебто суперечливість є оманливою, бо точка зміни – це демаркатор (кордон) між двома виявленими й локалізованими регулярними послідовностями (режимами). Тобто за «точкою» стоять дві регулярності (закономірності).

Дещо спрощуючи, можна поділити базові методи аналізу темпоральних даних на «феноменологічний» аналіз та «пізнавальний» аналіз. «Феноменологічний» аналіз виявляє, розпізнає, порівнює, підраховує, групує та типізує різні паттерни в даних. Такий аналіз оперує з великими зібраннями окремих послідовностей. Виявлені паттерни можуть бути використані як ознаки для задач класифікації, кластеризації, розпізнавання та пошуку. Задача виявлення типових фрагментів (motifs) полягає в тому, щоб в довгій послідовності знайти короткі послідовності, які найчастіше повторюються. Якщо аналізуються дані у формі суцільного ряду спостережень за об'єктом (середовищем), до поверхового аналізу можна віднести такі задачі, як виявлення періодичності, трендів, динамічних аномалій тощо [34, 40]. До «пізнавального» аналізу відносимо задачі та методи, які допомагають аналітику зрозуміти механізм розвитку процесу, знайти «неявні» об'єктивні зв'язки і закони, що керують процесом. Для цього треба аналізувати процес разом з його оточенням, тобто системно аналізувати багатовимірні ряди даних. Підходящим апаратом для цього, зокрема, є динамічні каузальні моделі. Методи їх виведення з даних узагальнюють такі традиційні підходи, як авторегресія. Виявлення каузальної структури зв'язків дає змогу прогнозувати наслідки втручання в об'єкт (ефект керування). Якщо є підстави вважати, що кілька рядів даних формуються на основі гармонічного сигналу з єдиного джерела, то перед тим, як приступити до виявлення зв'язків між рядами, бажано виділити ту спільну гармоніку й «очистити дані». Припустимо, частота (або період) гіпотетичної гармоніки відома, але невідомий зсув фази. Тоді для кожного ряду можна ідентифікувати модель за допомогою лінійної регресії, включивши в набір предикторів дві фіктивні змінні – $\sin(t \cdot x)$ та $\cos(t \cdot x)$. Отримавши коефіцієнти регресії, можна перетворити дві такі гармоніки на одну (з відповідним зсувом фази).

Із задачами й методами масованої переробки просторово-часових даних можна ознайомитися в [41]. Розширюється практика використання даних з просторовою структурою. Аналіз тривимірних та двовимірних даних застосовується у біології (протеоміка), медичній діагностиці, геофізиці тощо. Для деяких задач більш зручно працювати з не-метричними структурами (графовими, топологічними).

Виведення генеративних моделей

До цього роду задач і методів відносимо виведення (з даних) таких моделей, конструкцій і описів, які компактно описують сукупність даних, а іноді описують механізм породження (генерації) цих даних. Для такої задачі аналітик визначає набір змінних, без акценту на жодній з них. Виведення генеративних моделей охоплює вельми різноманітний арсенал методів і форм моделей. Результати (моделі) бувають різної точності та різного ступеня спрощення. Механізм генерації даних можна розуміти як зручний алгоритмічний опис процесу породження даних, еквівалентних (за статистичними характеристиками) фактичним даним, поданим на вхід методу. Натомість виведення генеративної моделі у сильному сенсі має більш амбітну мету – вивести таку модель, яка відображає структуру реального механізму генерації даних в об'єкті. До таких «структурно-адекватних» моделей належать каузальні мережі.

До методів виведення генеративних моделей (в їх широкому розумінні) можна віднести аналіз головних компонент (PCA), аналіз незалежних компонент (ICA), факторний аналіз, виявлення асоціацій і навіть (з деякими умовами) кластеризацію. (До речі, задачу характеристики сумісного розподілення ймовірностей $p(X)$ іноді редукують до пошуку «піків» й інтервалів великих значень ймовірності $p(X)$, а це нагадує кластеризацію.) В той час як PCA та класичний факторний аналіз обмежуються припущеннями лінійності та нормальності [25, 26, 29], метод ICA працює з прихованими факторами, розподіленими довільно. Відомий також нелінійний варіант ICA [42]. Не обмежені припущенням лінійності методи принципових кривих та принципових поверхонь, які шукають апроксимацію для розподілу даних у просторі [26]. Принципова крива проходить через «середні» точки прилеглих скупчень точок даних. Таких кривих можна знайти безліч, але задача розв'язується завдяки вимогам гладкості кривої. Розв'язок шукається ітеративною процедурою, яка уточнює криву та перерозподіляє точки даних між «відповідальними» за них точками на кривій. Аналогічно, мета побудови принципової поверхні – знайти двовимірну апроксимацію для даних. Побудова принципової поверхні дуже подібна до само-організуючих мап (COM). Різниця в тому, що метод COM знаходить невелику кількість точок-прототипів, а метод принципових поверхонь використовує окремий прототип для кожної точки даних [26].

Задача виявлення правил асоціації для дискретних даних була відповіддю на практичні потреби (виявлення типових кошиків покупок). Найвідоміший метод – алгоритм Apriori; він застосовує ідею покрокового збільшення довжини k асоціацій [15, 23, 26]. На старті список пропозицій складається з окремих змінних ($k=1$), а далі ітеративна процедура працює за наступною схемою:

– зібрати пропозиції довжиною $k+1$, такі, що кожний фрагмент пропозиції входить у список знайдених (прийнятих) асоціацій.

– відібрати ті пропозиції, які повторюються в даних частіше заданого порогу, і додати їх списку асоціацій.

З метою розширити можливості аналізу даних та підвищити рівень експресивності результатів аналізу дослідники шукають варіанти інтеграції статистичних методів з логікою [43–45]. Один з нових варіантів поєднання ймовірнісних і логічних моделей з каузальною семантикою – реляційна логістична регресія [45]. Моделі цього виду, подібно до каузальних мереж, використовують орієнтовані зв'язки.

На роль генеративних моделей в сильному сенсі претендують каузальні мережі [1, 46–49]. Каузальні мережі структурно (ізоморфно) описують процес генерації змінних моделі, відображаючи процес розгортання відповідних характеристик в об'єкті. Якщо цей опис дійсно адекватний, то каузальна мережа придатна для прогнозування наслідків планованих дій (наприклад, виконання рішень менеджера). Водночас ці моделі є багатоцільовими, оскільки дозволяють оперативно обирати цільову змінну і адаптуватися до будь-якого формату запиту без потреби повторно виводити модель.

Каузальна мережа (КМ) описується як пара (G, Θ), де G – граф, що специфікує структуру моделі, Θ – кількісні параметри, прив'язані до G. Граф структури G зазвичай є ациклонним орграфом. Якщо модель формується з одно-орієнтованих ребер, то параметри моделі специфікуються у формі $p(y|X)$ або $y = \mathrm{F}(X)$, де X – набір безпосередніх причин для y. Згідно форми параметризації моделі виділяються кілька різновидів КМ. Баєсові мережі побудовані на основі ординарних орграфів та дискретних (категорних) змінних. (Залежності в них описуються таблицями.) Каузальні Гаусові мережі використовують лінійні залежності та нормальні розподілення змінних. Гібридні мережі використовують неперервні та дискретні змінні в рамках одної моделі. Такі мережі покликані практичними потребами аналізу біологічних та медико-біологічних даних (і взагалі, великих даних). В разі гібридних мереж необхідно обробляти суміш категорних та неперервних змінних [50, 51].

Каузальні мережі підтримують два відмінні типи задач прогнозування – «пасивну» предикцію та каузальний прогноз («активну предикцію») [46, 52]. «Пасивна» предикції – це задача, подібна до тої, яку розв'язують задачі регресії та класифікації, вона формулюється як $p(C|a,b,..)$. (Але, на відміну від моделей регресії та класифікації, змінні $C, A, B, ..$ можна оперативно обирати, причому вони можуть займати будь-які позиції в моделі.) Каузальний прогноз виражається як $p(C|do(a),b,d..)$ і дає оцінку ефекту після втручання на змінну A (тобто ефект при мусового присвоєння $A = a$). Такий прогноз обчислюється на основі локально зміненої моделі [46, 48, 53].

Для опису таких процесів у часі розроблено динамічні каузальні мережі.Особливість динамічних КМ полягає в тому, що замість кожної окремої змінної в моделі представлено серію однойменних змінних (рис. 4), які репрезентують послідовні стани відповідного показника (характеристики). Функціонування динамічної каузальної мережі (генерація даних) може продовжуватися необмежено довго, послідовно просуваючись у часі дискретними інтервалами. Тобто черговий крок функціонування динамічної КМ (просування у часі) виконується таким чином, що, наприклад, вектору змінних з індексом часу $i-2$ присвоюється значення вектору змінних з індексом часу $i-1$. Тобто всі вектора значень пересуваються на один інтервал «назад» (у минуле). Значення вектору змінних з найбільшим індексом (тобто для поточного часу) генерується згідно залежностей моделі, як у звичайних каузальних мережах.

Динамічна каузальна мережа представляє регулярну структуру залежностей багатовимірного процесу у часі (рис. 4). Такі моделі описують генерацію багатовимірних часових рядів. Динамічна каузальна мережа передбачає стаціонарність

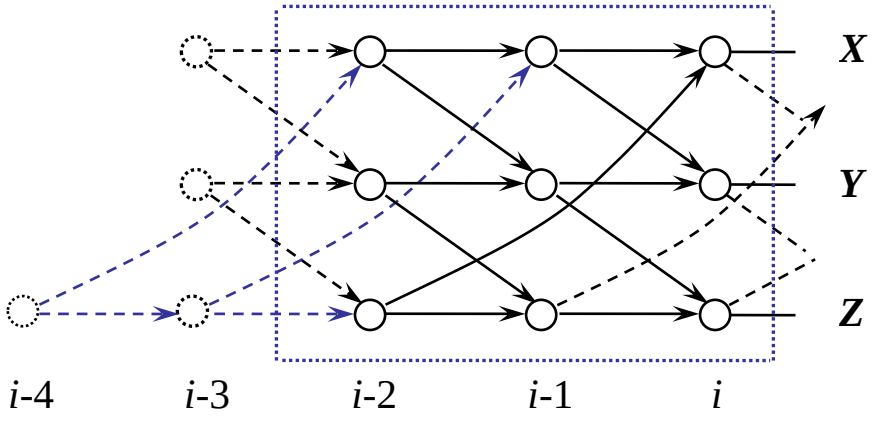

Рис. 4. Формування динамічної каузальної мережі

процесу і використовує форму векторної авторегресійної моделі [54, 55]. Модель виглядає як фрагмент часового ряду з довжиною, достатньою для явного відображення «найдовших» зв'язків. (Тобто в моделі мусить бути описано впливи з найбільшим лагом). На відміну від звичайної («статичної») каузальної мережі, в динамічний моделі структура виглядає як «клонована», оскільки зв'язки (ребра), як правило, повторюються на кожному кроці часового ряду.

Виведення каузальних мереж з емпіричних даних

Методи індуктивного виведення каузальних мереж націлені на виявлення каузальних зв'язків в ситуації, коли дані було зібрано як пасивні спостереження. (Більш того, зазвичай на вході не задається апріорних знань про структуру моделі.) Багато сучасних методів розраховані на існування прихованих (не присутніх в даних) спільних причин двох (чи більше) змінних. Далеко не для всіх зв'язків можливо розпізнати автентичний напрямок впливу. Через вказані обставини модель ідентифікується тільки як клас еквівалентності моделей, де використовуються ребра (зв'язки) різних типів, в тому числі з невизначеною орієнтацією (напрямком впливу). В типовій ситуації виведена модель використовує безпосередні зв'язки (ребра) чотирьох типів. Каузальне ребро (дуга) вигляду $X \to Y$ відображає каузальний вплив X на Y. Асоціативне ребро $U \leftrightarrow W$ позначає існування прихованої змінної (причини), що впливає рівночасно (паралельно) на U та W. Субкаузальне ребро $V \circ\!\!\to Z$ відображає два можливих варіанти: каузальний вплив або існування прихованої змінної (спільної причини). Неорієнтоване ребро $Q \circ\!\!-\!\!\circ R$ означає, що каузальний характер цього зв'язку зовсім не визначений. Кінці орієнтованого ребро $X \to Y$ називають «хвіст» та «вістря» відповідно.

Існуючі методи відтворення КМ з даних являються за своєю природою статистичними методами. Тому із зростанням довжини вибірки даних має підвищуватися точність виведеної моделі (за умови коректності методу), але ускладнюються обчислення. Збільшення ширини (багатовимірності) даних по-своєму ускладнює аналіз, але водночас створює можливість підвищити адекватність та точність моделі.

Стисло оглянемо та охарактеризуймо методи й алгоритми відтворення КМ з даних. За принципом роботи методи поділяються на «оптимізаційні» та «базовані на обмеженнях» (constraint-based) [47, 48, 50, 56]. «Оптимізаційні» методи спираються на показник «якості» моделі, який характеризує точність апроксимації даних (або предикції значень змінних) із штрафом за складність (розмірність) моделі. Зазвичай показником точності є правдоподібність моделі. Найбільш відомим «оптимізаційним» алгоритмом є GES. (Він використовує критерій BIC.) Серед методів, «базованих на обмеженнях», провідне місце посідають методи, що спираються на виявлення фактів умовної незалежності змінних (коротко – основані на незалежності). Методи, основані на незалежності, більш адаптовані до роботи з прихованими змінними, часто виграють у швидкості та мають тенденцію задовольнятися статистиками меншого формату. Відомі також гібридні методи відтворення моделей. Якщо дані розсічені («розщеплені») на окремі вибірки із скороченою номенклатурою змінних, класичні методи відтворення моделі стають не результативними. В таких ситуаціях можуть допомогти спеціальні співвідношення показників залежностей, які характерні для певних структур моделі. Приклади таких співвідношень можна знайти в [57].

Методи, основані на незалежності, виходять з припущення, що умовна незалежність (з відповідною умовою) буде чинна в даних тоді і тільки тоді, коли ця незалежність випливає зі структури генеративної моделі. В такому разі знайдена умовна незалежність для пари змінних X, Y свідчить про відсутність ребра між X та Y. Схема роботи цих методів показана на рис. 5. Найбільш витратний етап – ідентифікація безпосередніх зв'язків. Технічно, на цьому етапі для кожної пари змінних шукається умова («сепаратор»), яка забезпечує умовну взаємну незалеж-

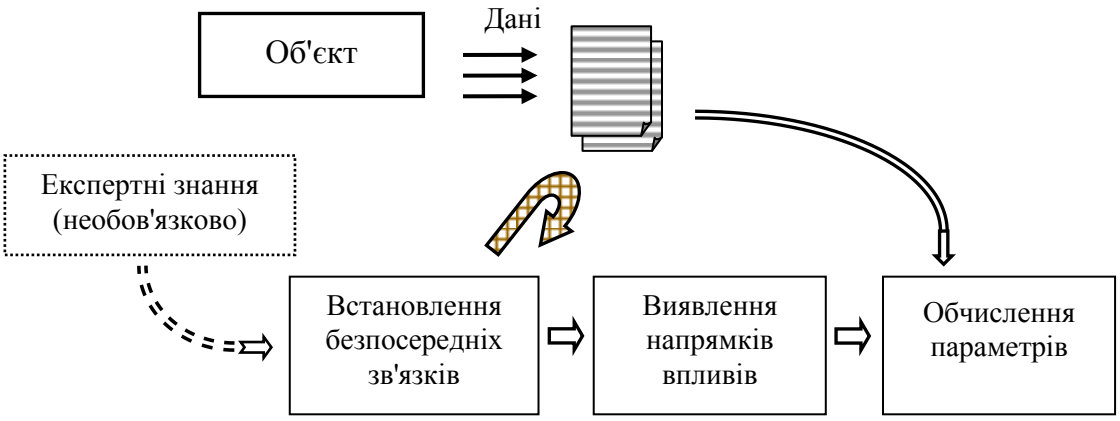

Рис. 5. Схема виведення каузальних мереж з даних

ність змінних цієї пари. Кількість тестів умовної незалежності, що виконуються на цьому етапі, може сягати десятків чи сотень тисяч.

Вказані методи здатні ідентифікувати каузальне відношення з даних тільки за наявності достатніх свідчень. Зрозуміло, що сама змінна ефекту (наслідку) та її причина мусять бути присутні в даних. Остаточно каузальний зв'язок (ребро) ідентифікується на підставі того, що знайдено квазі-інструментальну змінну, а також супутню «автономну» змінну, яка допомагає розпізнати першу змінну як квазі-інструментальну [46, 47, 58]. Тому зростання «широти» даних має сприяти виявленню каузальних зв'язків. Якщо структура мережі наближається до насиченої, то каузальні зв'язки не ідентифікуються.

Еталонними інструментами цього напрямку стали алгоритми PC та FCI [47, 50, 59]. Протягом останніх років було запропоновано низку вдосконалень цих алгоритмів та інструментів. Алгоритм FCI-Stable [60] наразі постає стандартом замість FCI. Цей алгоритм – більш робастний, в ньому усунута чутливість до порядку змінних. В алгоритмі FCI-Stable ребра видаляються з кістяка моделі тільки після виконання всіх тестів незалежності поточного рангу. Алгоритм залучає техніку тестування для мішаних даних й також використовує розпаралелювання. Показано, що в стратегію роботи алгоритмів відтворення (таких, як PC) доцільно залучити принцип пошуку, оснований на максимумі ймовірності. Перший варіант алгоритму з цим принципом (не розрахований на латентні змінні) – PC-Max. Нова версія (FCI-MAX) залучає пошук за максимумом ймовірності для орієнтації ребер. У [51] окреслено чотири нові стратегії для відтворення каузальних моделей (з допуском латентних змінних), на основі мішаних даних: FCI-MAX, MGM-FCI-MAX, MGM-FCI та MGM-CFCI. Версія FCI-MAX для орієнтації ребер перевіряє цілий набір підходящих сепараторів. Техніка тестування незалежності основана на лінійній та логістичній регресії. Найбільш універсальний підхід до тестування умовної незалежності спирається на техніку репродуктивного кернелу [25, 26, 61]. Оскільки ця техніка потребує дуже витратних обчислень, багато досліджень присвячено її вдосконаленню.

Процес виведення моделі часто стикається з обчислювальними проблемами, бо кількість можливих структур моделі є факторіально (експоненційно) великою, і кількість можливих сепараторів – також. З метою стримування комбінаторної складності алгоритмів виведення моделі були знайдені можливості фокусування пошуку сепараторів. Для цього було створено апарат локально-мінімальної сепарації та розроблено набір резолюцій індуктивного виведення [49, 62, 63]. (Ці засоби мають логічний характер і придатні для будь-яких форм параметризації моделі.) Тим самим, запропоновано систематичний підхід до пошуку сепараторів, базований на використанні імплікацій марковських властивос-

тей каузальних мереж. Теоретичний ґрунт новацій – необхідні вимоги до членів локально-мінімального d-сепаратора. Розроблені засоби вбудовано в алгоритми серії Razor і дозволяють адаптивно оптимізувати і звужувати пошук складних мінімальних сепараторів, виходячи з знання вже знайдених «сусідніх» простих сепараторів та паттернів залежностей. Відсікаються цілі сектори простору пошуку сепараторів. Результати випробувань показали перевагу алгоритмів Razor над базовим аналогом PC за кількістю тестів (а відтак, за швидкодією) і за адекватністю відтворення каузальних зв'язків [49, 63].

Для демонстрації можливостей відтворення каузальних мереж методами, основаними на незалежності, наведемо приклад аналізу реальних даних соціального характеру (ці дані не є великими). Мета – ідентифікація факторів, які визначають вік матері при народженні першої дитини в США (за даними 70-х років 20-го століття). Спеціалісти зібрали вибірку даних з наступним набором змінних (можливих факторів): 1) професія батьків; 2) раса; 3) відсутність братів (сестер); 4) жила (чи ні) матір на фермі; 5) регіон США; 6) наявність двох дорослих у родині, де росла матір; 7) релігія; 8) паління сигарет; 9) був чи ні викидень; 10) освітній рівень матері (на час виходу заміж); 11) вік матері при народженні першої дитини. Ця задача раніше була розв'язана за допомогою алгоритму PC [47]. Тепер для цієї задачі застосовано розроблений автором алгоритм Razor-1.3. Подібно до постановки в [47], також було гіпотезовано лінійність залежностей. Але на відміну від вищезгаданої постановки не було задано жодних апріорних обмежень на структуру (і не вказувалась роль змінних). Алгоритм Razor-1.3 виводить модель в більш загальному класі – класі не-рекурсивних каузальних мереж, де розрізняються каузальні, суб-каузальні, асоціативні та неорієнтовані ребра (зв'язки) [47, 49, 58, 63]. Результат роботи (структура моделі) показано на рис. 6. «Кільця» позначають невизначений кінець ребра (він може виявитися «вістрям» або «хвостом»). Для візуальної наочності подвійними лініями зображено ті ребра, які ідентифіковано як каузальні зв'язки (тобто такі, що відображають спрямований вплив одної змінної на іншу). Отже, виявлено п'ять каузальних зв'язків. Зокрема, на вік матері при народженні першої дитини впливають освітній рівень матері та регіон проживання (‘$5 \to 11$’ та ‘$10 \to 11$’). Більшість виведених ребер є суб-каузальними та неорієнтованими, наприклад, ‘$2 \circ \to 11$’ та ‘$2 \circ - \circ 10$’ відповідно. Для отримання інформативніших результатів потрібно залучити до аналізу ширший комплект релевантних характеристик.

Завдяки гіпотезі лінійності залежностей тестування незалежності виконується просто (через частинні кореляції, які обчислюються з матриці парних коваріацій.) Але поза лінійністю (наприклад, коли змінні – номінального типу) кожний тест незалежності потребує нового сканування даних, що призводить до зростання тривалості виведення моделі.

Стисло охарактеризуємо методи відтворення динамічних каузальних мереж. Ці методи ґрунтуються на припущенні, що задані на вході часові ряди даних відображають стаціонарний векторний авторегресійний процес. Виведення каузальних моделей з багатовимірних часових рядів має тривалу передісторію. Подібні за характером моделі й методи розроблялися в руслі економетричних досліджень. Концепція каузальності для часових рядів була запропонована К. Грейнжером (Granger C.W.J.) [64, 65]. Сучасну постановку ця проблема отримала в термінах каузальних мереж [54, 56, 66]. Адаптація звичайних (статичних) методів відтворення каузальних мереж до ситуації динамічних даних ґрунтується на припущенні стаціонарності та на наступних домовленостях. Структура зв'язків повторюється на кожному інтервалі вимірювання даних. Модель виглядає як фрагмент багатовимірного ряду, але компактно презентує увесь довгий ряд (в згорнутому вигляді). Особливістю цієї задачі є регулярна присутність авторегресійних зв'язків, а також заданий темпоральний порядок в даних. Розмір (часова глибина) моделі має визначатися довжиною (глибиною) безпосереднього зв'язку з

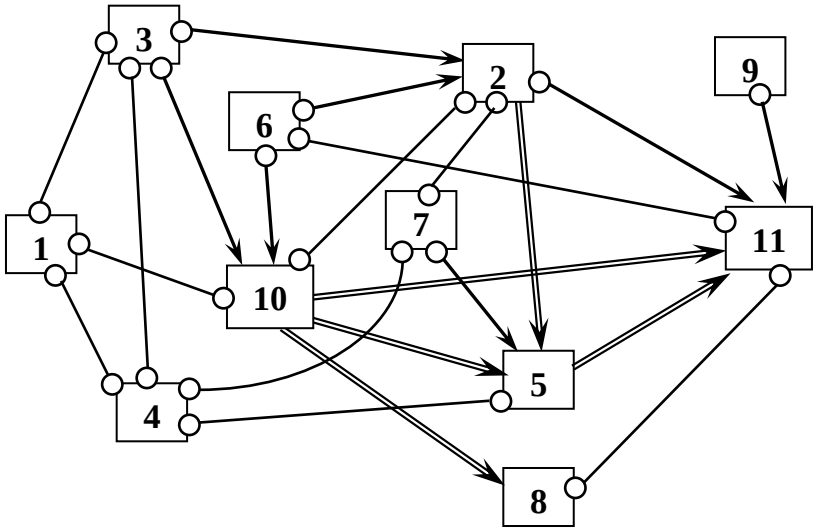

Рис. 6. Модель взаємодії гіпотетичних факторів, що впливають на вік матері при народженні першої дитини

найбільшим часовим лагом (післядією). В динамічних каузальних мережах завжди присутні додаткові «неявні» залежності, які є наслідком впливів з минулого. Наприклад, на рис. 4 ілюструється ситуація, коли найбільший лаг зв'язку дорівнює двом (зв'язок $Z_{i-2} \to X_i$). Модель охоплює три послідовні стани (обведено пунктирною рамкою) і включає зв'язки, показані суцільними дугами. Стоїть питання, як відображати впливи й залежності «з минулого». Вони не вкладаються в рамки моделі і тому показані на рис. 4 пунктирними дугами. Такі залежності розповсюджуються далі через ребра моделі і проявляються як транзитні залежності. Якщо названі впливи не відображати в моделі, то, згідно буквальної інтерпретації структури моделі, наприклад, змінні X_{i-2} та Y_{i-2} мають бути взаємонезалежні (а це не так). Якщо ж замість зв'язків, показаних на рис. 4 пунктиром, ввести в моделі додаткові дуги (ребра), які замінять впливи з минулого, то структура моделі стане нерегулярною (більш насиченою зліва), що суперечить стандартним припущенням. Але ці незручності мають синтаксичний характер. Можливі й більш суттєві проблеми. Існування прихованого часового ряду може створити феномен «нескінченно-довгого» часового лагу [54]. Це робить неможливим тестування марковських властивостей в ході виведення моделі.

Регулярність структури та відомий темпоральний порядок змінних сприяють спрощенню пошуку сепараторів і орієнтації ребер в ході виведення динамічних каузальних мереж (порівняно з звичайними мережами). Але задача ускладнюється через те, що зазвичай аналітик на знає довжини найбільшого лагу впливів.

Алгоритми SVAR-FCI та SVAR-GFCI [54] є результатом адаптації методу FCI до даних у формі багатовимірних часових рядів з прихованими змінними. Ці алгоритми припускають існування «швидких» зв'язків, тобто зв'язків між змінними в межах одного інтервалу вимірювання даних, але без утворення циклонів. Циклони не утворюються, якщо впливи протягом одного інтервалу вимірювання не встигають оббігти цикл й повернутися до стартової змінної. Якщо дані часових рядів фіксувалися з недостатньою частотою (тобто інтервал між вимірюваннями надто великий), то відтворити адекватну каузальну структуру проблематично [66]. Відомою проблемою аналізу часових рядів даних є нестаціонарність. Тоді вимірювання і інтерпретація залежностей стають суперечливими. В [67] пропонується метод виведення динамічних каузальних мереж в умовах

однієї з форм нестаціонарності (за присутності тренду).

В принципі, припущення про суцільну регулярність системи зв'язків крізь всі інтервали часу не є обов'язковим. Можна виводити модель навіть якщо зв'язки для сусідніх інтервалів відрізняються. Але структура зв'язків має бути періодичною, тобто повторюватися з зсувом на кожні k інтервалів. (Така ситуація нетипова.) Якщо величина періоду k невідома, виявити її важко. В такому разі статистики для тестування обчислюються інакше (по-суті, треба паралельно виводити кілька моделей).

Особливості застосування методів аналізу до великих даних

Достаток і рясніта доступних даних можуть спровокувати надмірний оптимізм і нехтування науковим підходом та настановами статистичного аналізу. Під враженням успіхів перших (поверхових) прикладів аналізу даних з Інтернету програмісти припустили легковажність у підході до аналізу даних і у тлумаченні результатів [68]. Номінально величезний обсяг даних не завжди означає подолання проблем скінченних вибірок даних. Дані можуть бути деформовані внаслідок селекції (особливо якщо вони зібрані за допомогою пошуку в Інтернеті). Якими би «повними» не були (не здавалися) дані, не припустимо плутати кореляцію з каузальністю.

Поява ВД впливає на вибір методів аналізу і стимулює їх розвиток [21, 22, 27, 28, 69]. З огляду на великий обсяг даних преференції надаються швидким методам (навіть якщо вони менш точні). Підсилюються стимули застосовувати розпаралелювання обчислень (особливо коли розпаралелювання узгоджується з схемою розміщення даних). ВД характерні також тим, що включають дані (змінні, атрибути) різних типів (дійсні, дискретні, ординальні, категорні), що ускладнює техніку обробки і побудови моделей. В методах, що спираються на оптимізацію квазіправдоподібності моделі, обчислювальна складність стає неприйнятною. Тому пропонують відмовитися від обчислення повного градієнту і послідовно просуватися вздовж окремих координат [12]. Коли йдеться про ВД, обчислювальну складність методів аналізу треба оцінювати дещо інакше. На відміну від традиційних застосувань, значним фактором складності стає кількість (кратність) сканувань даних. (Іноді це навіть важливіше за кількість абстрактних обчислювальних операцій.)

Уявімо, що дані є статистичною вибіркою у формі плаского масиву. Зростання тільки довжини даних (збільшення розміру вибірки) забезпечує один ефект – зростання надійності й точності результату статистичного аналізу, звуження довірчих інтервалів. Зрозуміло, що аналіз великих даних має значно амбітнішу мету – отримати повніші й змістовніші результати, глибше «зазирнути» в суть об'єкту. Для цього потрібні дані більшого формату. Зростання довжини («вишини») даних має супроводжуватися зростанням їх «ширини», тобто їх вимірності. (Дані мусять бути великими одночасно в обох «вимірах».) Отже, одна з важливих ознак великих даних (для глибокого аналізу) – багатовимірність. Багатовимірні дані, з одного боку, надають нові можливості й шанси, а з іншого – породжують проблеми для аналізу [12, 70]. Наприклад, для збереження ефективності методів найближчого сусіда необхідно, аби довжини даних зростала експоненційно швидше за вимірність даних [26, 27, 70]. Від зростання вимірності даних також потерпають методи, що спираються на техніку кернелу. Небажані ефекти особливо загострюються, коли ширина даних перевищує довжину (це характерно для даних експресії генів і взагалі біоінформатики). Коли багато змінних одночасно використовуються як ознаки, виникає синдром накопичення гамору для задач класифікації. Між «істинними» факторами та деякими не-релевантними змінними можуть випадково виникати обманні асоціації.

Коли дані вертикально- секціоновані («розщеплені») і не вдається їх синтезувати (ототожнити випадки), то аналітику залишається послаблений варіант багатовимірного аналізу. Замість аналізу на рівні випадків (прецедентів) доведеться перейти до аналізу на рівні нечітких «класів еквівалентності» прецедентів (які формуються

на підставі подібності властивостей). Зрозуміло, що зв'язки між такими класами еквівалентності напевно будуть нечіткими, «розмазаними» і малоінформативними.

Із зростанням довжини вибірки стає критичним питання вибору критерію якості моделі, бо збільшується розходження (відмінність) між вибором за критерієм BIC та критерієм AIC, а також результатом традиційного тестування гіпотез [27]. Штраф за складність моделі в BIC має додатковий (порівняно з AIC) множник $\log(n)$. Тому BIC обирає меншу (простішу) модель, ніж AIC.

Великі дані відкривають перспективи розробки нових методів. Покращуються умови для розв'язання «витончених» задач «не-прямими» методами, такими, як «навчання на чужому досвіді» [27, 28]. Ідею цього підходу можна пояснити наступним чином. Уявімо, що класичної I.I.D.-вибірки даних немає. Лише мала частка даних строго релевантна для задачі оцінки (прогнозу) характеристики. Тоді «пряму» оцінку характеристики (отриману з релевантних даних) беруть як орієнтовну, початкову. Потім вносять поправку згідно загальної закономірності, попри те, що закономірність була отримана з інших вибірок даних. Популярна нині стратегія масового тестування гіпотез за критерієм FDR також може розглядатися як втілення ідеї «навчання на чужому досвіді».

Кількість випадків (довжину масиву) не завжди припустимо тлумачити як розмір статистичної вибірки. Часто зростання масиву даних досягається шляхом об'єднання даних з різних джерел. Утворена таким чином «вибірка» даних буде неоднорідною і порушує стандартні припущення. З дещо послабленим варіантом неоднорідності даних аналітик стикається, якщо дані збиралися протягом тривалого часу, так що об'єкт змінювався (еволюціонував). Для ситуацій з можливою різнорідністю даних пропонується [21] ввести припущення інваріантності для підмножини коваріат. Цьому припущенню задовольняють каузальні коваріати. Різнорідність даних призводить до погіршення стабільності та якості результатів аналізу. Водночас різнорідність даних надає можливість вирішити корисну задачу виокремлення різних джерел даних (коли вони апріорі невідомі аналітику). Внаслідок різнорідності даних стандартна крос-валідація стає поганим інструментом оцінки предиктивної адекватності моделі (і адекватності каузального виведення). В такій ситуації краще звернутися до використання синтетичних даних та симуляції (але потрібні предметні знання).

Взагалі, якщо для розробки традиційних статистичних методів можна було прийняти зручні припущення про вибірку даних, то тепер, маючі справу з великими даними, необхідно сприймати як факт реальні механізми збору даних і відповідно обирати чи розробляти методи аналізу. Треба створювати нові методи аналізу, які будуть повніше враховувати механізм збору даних (і навіть «виправляли» і компенсували викривлення даних, наскільки це можливо). Методи аналізу даних мають вибиратися (налаштовуватися) у відповідності з особливостями збору даних. Наприклад, якщо дані були піддані селекційному зміщенню, то в залежності від обставин можливі такі рішення [53, 71]: 1) компенсувати (виправити) селекційне зміщення; 2) інтерпретувати результати аналізу інакше; 3) констатувати, що задачу неможливо розв'язати. Принципову можливість отримати коректний результат на основ «викривлених» даних можна проілюструвати на прикладі методу адекватної оцінки каузального ефекту на основі даних, що пройшли селекцію [71]. Налаштування методів аналізу на характер збору даних буде свідчити про інтеграцію комплексу досліджень «Великі дані плюс Велика Аналітика» у єдину науку.

Алгоритми й методи аналізу необхідно піддавати всебічному випробуванню у різних сценаріях, включно з «крайніми». Результати глибокого аналізу приймаються як основа для наукових узагальнень та практичних висновків тільки після кількісної оцінки невизначеності, статистичних помилок, стабільності і реплікативності результатів. (Реплікативність означає, що альтернативні аналітичні дослідження обраного об'єкту дають близькі результати [21, 69].)

Підсумки

Інтенсивний збір даних та накопичення великих даних у комп'ютерних сховищах стимулює зміни інформаційних технологій. На передній план висувається індуктивно-емпірична методологія та технології глибокого аналізу даних. Виник попит на методи, які швидко й результативно перетворять величезний масив «сирих» даних на цінну інформацію кінцевого споживання. Необхідна передумова глибокого аналізу даних – їх багатовимірність та великий обсяг. Оскільки зібрані дані зазвичай «сирі» й погано структуровані, перед власне застосуванням аналітичних методів необхідно виконати підготовку даних.

Увесь арсенал задач аналізу великих даних можна розбити на такі групи: 1) впорядкування даних (зокрема, кластеризацію); 2) виведення ціле-визначених (предиктивних) моделей; 3) дослідження та узагальнення даних (в тому числі відкриття структур, зв'язків, паттернів і закономірностей).

Спектр методів кластерного аналізу за принципом роботи можна розподілити на три підходи: кластеризація на основі сукупної близькості; кластеризація на основі локальної близькості та множинному сусідстві; кластеризація на основі статистичної моделі даних. Оскільки універсального критерію якості кластеризації не запропоновано, виділення кластерів часто є евристичним і залежить від вибору аналітика. Для здатності виділяти кластери нестандартних (не-опуклих) форм розроблено багато евристичних алгоритмів, які спираються на відношення сусідства, на колективну локальну близькість точок.

До ціле-визначених задач відноситься виведення предиктивних (дискримінативних) моделей (зокрема, регресії та класифікації), які описують цільову змінну через інші змінні (предиктори або ознаки). Використовуючи гнучкі й адаптивні форми залежності, можна мінімізувати відхилення моделі від вхідних даних. Але узгодженість моделі з вхідними даними далеко не завжди означає адекватність і здатність прогнозувати значення у впровадженнях. Втрата адекватності моделі відбувається через синдром гіпер-специфікації (over-fitting), до якого призводить зависока адаптивність і складність обраного класу моделей. Типовими засобами захисту від гіпер-специфікації є крос-валідація та регуляризація. Адекватність ціле-визначених моделей суттєво залежить від підбору значущих предикторів. Популярна тенденція – побудова моделі за допомогою дерева (дерева регресії, методи MARS, випадкові дерева). Висока адаптивність цих методів досягається завдяки тому, що простір предикторів раціонально розбивається на сегменти (ареали) відповідно до локальної поведінки залежності.

Задачі «глибокого навчання» створюють високо-адаптивні моделі для розпізнавання. Успішність «глибокого навчання» пояснюється тим, що задача високоспеціалізована, результат – лаконічний, а вхідні змінні – «пасивні» і взаємозамінні. Складність отриманих моделей виправдана тим, вхідні змінні «дрібні», а «сусідні» змінні тісно корельовані. Натомість задачі глибокого аналізу даних та відкриття знань (на відміну від «глибокого навчання») характеризуються більш невизначеною ситуацією на вході. Виявлені закономірності та «знання» є результатом «кристалізації» неявних (прихованих) відносин між різноплановими змінними. Ці два напрями різняться також характером переробки даних: перший має характер тренування, «підгонки» й оптимізації; другий має пошуково-дослідницький характер. Модель з високою предиктивною ефективністю не завжди корисна для розуміння (пояснення) об'єкту.

Розв'язання багатьох прикладних задач аналізу ВД все частіше потребує враховувати часову прив'язку записів даних та автокореляцію. Аналізуються зв'язки не тільки в межах записів, але й поміж записами. Задачі аналізу темпоральних даних вирізняються великим розмаїттям, потребують спеціалізації й комплексування методів та їх ієрархічного застосування. Ці задачі включають сегментацію, виявлення точок змін, виявлення трендів, аномалій, періодичності, спектру і т. д.

Традиційні ціле-визначені моделі прив'язані не тільки до цільової змінної,

але й до фіксованого набору предикторів. Тому неясно, як застосовувати модель, коли відомі не всі предиктори. Це питання знаходить коректне розв'язання в апараті каузальних моделей. Каузальні моделі (мережі) поєднують у собі переваги генеративних, ціле-визначених та багатоцільових моделей. Каузальні мережі здатні адекватно описати процес генерації даних в об'єкті. Каузальну мережу можна розглядати як інтегровану систему предиктивних та дискримінативних моделей. Головна перевага каузальних моделей над традиційними – вони підтримують прогнозування наслідків втручання в об'єкт (ефект керування).

Методи виведення каузальних мереж з даних за принципом роботи поділяються на «оптимізаційні» та основані на незалежності. Останні спираються на виявлення фактів умовної незалежності змінних. Ці методи більш адаптовані до роботи з латентними змінними, часто виграють у швидкості та потребують статистик меншого формату. Найбільш універсальна техніка тестування умовної незалежності – репродуктивний кернел (ядро); її недоліком є високі обчислювальні витрати. Зниження обчислювальних витрат в ході виведення каузальних мереж можна досягти, зокрема, обмежуючи комбінаторну складність виведення завдяки фокусуванню пошуку сепараторів. Для цього розроблено набір резолюцій, які дозволяють розв’язувати питання щодо присутності ребер раніше, які обґрунтовані на графовому рівні (через поняття локально-мінімального d-сепаратора та необхідні вимоги до членів локально-мінімального сепаратора).

Методи виведення каузальних мереж з емпіричних даних (які були зібрано як пасивні спостереження) мають обмеження у можливостях вичерпно розпізнавати каузальні зв'язки. Результатом виведення каузальної мережі з даних зазвичай є неповністю визначена модель, де характер (спрямування) багатьох зв'язків неоднозначний.

Можливості методів відтворення каузальних мереж продемонстровано на прикладі. Виведено модель взаємодії гіпотетичних факторів, що впливають на вік матері при народженні першої дитини. Більшість виведених зв'язків – суб-каузальні та неорієнтовані (тільки декілька каузальних). Для отримання більш інформативних результатів потрібні дані з розширеним набором релевантних характеристик.

Застосування методів аналізу до великих даних має низку особливостей. Важливим фактором обчислювальної складності методів стає кількість (кратність) сканування даних. Поява великих даних впливає на вибір методів аналізу (включаючи статистичні), і стимулює їх розвиток. Не припустимо нехтувати науковим підходом та застереженнями статистичного аналізу. Зокрема, не можна плутати кореляцію з каузальністю. Дані, зібрані за допомогою пошуку в Інтернеті, можуть бути деформовані внаслідок селекції. Серед засобів оцінки коректності методів та адекватності результатів зростає роль симуляції з використанням синтетичних даних, а також випробування у різних сценаріях. Щоб прийняти результати аналізу даних як основу для узагальнень та висновків, необхідно оцінити їх невизначеність, стабільність, реплікативність та статистичні помилки.

Єдина методологія для всього циклу життя великих даних дозволить вибирати і налаштовувати методи аналізу у відповідності з особливостями збору даних. Технології збору даних, збереження, менеджменту та пошуку будуть інтегруватися з методами глибокого аналізу даних, утворюючи нову науково-технологічну сферу «Великі дані плюс Велика Аналітика».

Література

1. Балабанов О.С. Аналітика великих даних: принципи, напрямки і задачі. *Проблеми програмування*. 2019. № 2. С. 47–68.
2. Bühlmann P., Drineas P., Kane M., van der Laan M. (eds.) Handbook of Big Data. Taylor and Francis, 2016. 456 p.
3. Mayer-Schönberger V., Cukier K. Big Data: A revolution that will transform how we live, work, and think. Boston, MA: Houghton Mifflin Harcourt, 2013. 256 p.
4. Chen C.L.P. and Zhang C.-Y. Data-intensive applications, challenges, techniques and

technologies: A survey on Big Data. *Information Sciences*. 2014. Vol. 275. P. 314–347.
5. Chen M., Mao S. and Liu Y. Big Data: A Survey. *Mobile Networks and Applications*. 2014. Vol. 19, Issue 2. P. 171–209.
6. Bhadani A. and Jothimani D. Big Data: Challenges, opportunities and realities / In.: M.K. Singh and D.G. Kumar (eds.). Effective Big Data management and opportunities for implementation. – IGI Global, Pennsylvania, USA, 2016. – [Електронний ресурс] Доступ: https://arxiv.org/pdf/1705.04928.
7. Oussous A., Benjelloun F.-Z., Lahcen A.A. and Belfkih S. Big Data technologies: A survey. *Journal of King Saud University. Computer and Information Sciences*. 2018. Vol. 30, Issue 4. P. 431–448.
8. Cao L. Data science: a comprehensive overview. *ACM Computing Surveys*. 2017. Vol. 50, N 3, Article 43, 42 p.
9. Gandomi A. and Haider M. Beyond the hype: Big data concepts, methods, and analytics. *Intern. Jour. of Information Management*. 2015. Vol. 35, N 2. P. 137–144.
10. Tsai C.-W., Lai C.-F., Chao H.-C. and Vasilakos A.V. Big data analytics: a survey. *Journal of Big Data*. 2015. Vol. 2, N 1. P. 1–32.
11. Watson H.J. Tutorial: Big Data analytics: Concepts, technologies, and applications. *Comm. of the Association for Information Systems*. 2014. Vol. 34, Article 65. P. 1247–1268.
12. Fan J., Han F. and Liu H. Challenges of Big Data analysis. *Nat. Scient. Rev*. 2014., Vol. 1, N 2. P. 293–314.
13. Franke B., Plante J.-F., Roscher R., Lee E.A., Smyth C., Hatefi A., Chen F., Gil E., Schwing A.G., Selvitella A., Hoffman M.M., Grosse R., Hendricks D. and Reid N. Statistical inference, learning and models in Big Data. *Intern. Statistical Review*. 2016. Vol. 84, N. 3. P. 371–389.
14. Zafarani R., Abbasi M.A. and Liu H. Social media mining. An introduction. Cambridge University Press, 2019. 380 p.
15. Андон Ф.И., Балабанов А.С. Выявление знаний и изыскания в базах данных: подходы, модели, методы и системы (обзор). *Проблемы программирования*. 2000. № 1–2, С. 513–526.
16. Балабанов А.С. Выделение знаний из баз данных – передовые компьютерные технологии интеллектуального анализа данных. *Математичні машини і системи*. 2001. № 1–2. С. 40–54.
17. Azzalini A. and Scarpa B. Data analysis and Data Mining: An introduction. – N.Y.: Oxford University Press, 2012. 288 p.
18. Swanson N.R. and Xiong W. Big Data analytics in economics: What have we learned so far, and where should we go from here? *Canadian J. of Economics*. 2018, Vol. 51, Issue 3. P. 695–746.
19. Graham E. and Timmermann A. Forecasting in Economics and Finance. *Annual Review of Economics*. (2016). Vol. 8. P. 81–110.
20. Weihs C. and Ickstadt K. Data Science: the impact of statistics. *Intern. J. of Data Science and Analytics*. 2018. Vol. 6. P. 189–194.
21. The role of statistics in the era of big data. Special issue of the journal: *Statistics and Probability Letters*. May 2018. Vol. 136.
22. Secchi P. On the role of statistics in the era of big data: A call for a debate. *Statistics and Probability Letters*. 2018. Vol. 136. P. 10–14.
23. Witten I.H., Eibe F., Hall M.A. (3rd ed.).Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2011. 629 p.
24. Maimon O., Rokach L. (Eds.) Data Mining and Knowledge Discovery Handbook. 2nd ed., Springer-Verlag New-York Inc., 2010. 1285 p.
25. Murphy K.P. Machine learning: a probabilistic perspective. MIT Press, Cambridge, Massachusetts, 2012. 1055 p.
26. Hastie T., Tibshirani R. and Friedman J. The elements of statistical learning. (2nd ed.). Springer. 2009. 745 p.
27. Efron B. and Hastie T. Computer age statistical inference. Cambridge University Press, 2016. 475 p.
28. Efron B. Large-scale inference. Stanford University Press, 2010. 263 p.
29. James G., Witten D., Hastie T. and Tibshirani R. An introduction to statistical learning with applications in R. Springer, N.Y., 2013. 426 p.
30. Berkhin P. A survey of clustering data mining techniques. In: Kogan J., Nicholas C., Teboulle M. (eds.). *Grouping multidimensional data*. Springer-Verlag: Berlin-Heidelberg, 2006. P. 25–71.
31. Bouveyron C., Brunet-Saumard C. Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*. 2014. Vol. 71. P. 52–78.
32. Kurban H., Jenne M. and Dalkilic M.M. Using data to build a better EM: EM* for big data. *Intern. J. of Data Science and Analytics*. 2017. Vol. 4, Issue 2. P. 83–97.
33. LeCun Y., Bengio Y., Hinton G. Deep learning. *Nature*. 2015. Vol. 521, P.436–444.

34. Esling P. and Agón C. Time-series data mining. *ACM Computing Surveys*. 2012. Vol. 45, Issue 1. P. 12–34.
35. Chandola V., Banerjee A. and Kumar V. Anomaly detection for discrete sequences: a survey. *IEEE Trans. on Knowledge and Data Eng*. (TKDE). 2012. Vol. 24, N 5. P. 823–839.
36. Truong C., Oudre L. and Vayatis N. Selective review of offline change point detection methods. [Electronic resource] URL: https://arxiv.org/abs/1801.00718.
37. Aminikhanghahi S. and Cook D.J. A survey of methods for time series change point detection. *Knowledge and Information Systems*. 2017. Vol. 51, Issue 2. P. 339–367.
38. Frick K., Munk A. and Sieling H. Multiscale change point inference. *J. Roy. Statist. Soc., ser. B*. 2014. Vol. 76, Pt. 3. P. 495–580.
39. Wang T. and Samworth R.J. High dimensional change point estimation via sparse projection. *J. Roy. Statist. Soc., ser. B*. 2018. Vol. 80, Pt. 1. P. 57–83.
40. Liao T.W. Clustering of time series data – a survey. *Pattern Recognition*. 2005. Vol. 38. P. 1857–1874.
41. Atluri G., Karpatne A. and Kumar V. Spatio-temporal Data Mining: a survey of problems and methods. *ACM Computing Surveys*. 2018. Vol. 51, Issue 4, Article N 83.
42. Lee T.-W., Girolami M., Bell A.J., Sejnowski T.J. A unifying information-theoretic framework for Independent Component Analysis. *Intern. J. Computers and Mathematics with Applications*. 2000. Vol. 39. P. 1–21.
43. Neville J. and Jensen D. Relational Dependency Networks. *Jour. of Machine Learning Res*. 2007. Vol. 8. P. 653–692.
44. De Raedt L., Kersting K., Natarajan S. and Poole D. Statistical relational artificial intelligence: Logic, probability, and computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*. 2016. Vol. 10, N 2. P.1–89.
45. Kazemi S.M., Buchman D., Kersting K., Natarajan S. and Poole D. Relational logistic regression: The directed analog of Markov logic networks. *Workshops at the Twenty-Eighth AAAI Conf. on Artificial Intelligence*. 2014. P. 41–43.
46. Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge Univ. Press, 2000. 526 p.
47. Spirtes P., Glymour C. and Scheines R. Causation, prediction and search. New York: MIT Press, 2001. 543 p.
48. Peters J., Janzing D. and Schölkopf B. Elements of Causal Inference. Foundations and Learning Algorithms. MIT Press, Cambridge, MA, USA, 2017. 265 p.
49. Балабанов О.С. Відкриття знань в даних та каузальні моделі в аналітичних інформаційних технологіях. *Проблеми програмування*. 2017. № 3. С. 96–112.
50. Raghu V.K., Ramsey J.D., Morris A., Manatakis D.V., Sprites P., Chrysanthis P.K., Glymour C., Benos P.V. Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *Intern. Jour. of Data Science and Analytics*. 2018. Vol. 6, Issue 1. P. 33–45.
51. Tsagris M., Borboudakis G., Lagani V., Tsamardinos I. Constraint-based causal discovery with mixed data. *Intern. Jour. of Data Science and Analytics*. 2018. Vol. 6, Issue 1. P. 19–30.
52. Pearl J. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*. 2019. Vol. 62, Issue 3. P. 54–60.
53. Pearl J. and Bareinboim E. External validity: From do-calculus to transportability across populations. *Statistical Science*. 2014. Vol. 29, N 4. P. 579–595.
54. Malinsky D. and Spirtes P. Causal structure learning from multivariate time series in settings with unmeasured confounding. *Proc. of 2018 ACM SIGKDD Workshop on Causal Discovery*, August 2018, London, UK. PMLR, Vol. 92. P. 23–47.
55. Entner D. and Hoyer P.O. On causal discovery from time series data using FCI. *Proc. of the 5th European Workshop on Probabilistic graphical models*. 2010, Helsinki, Finland. P. 121–128.
56. Runge J. Causal network reconstruction from timeseries: From theoretical assumptions to practical estimation. *Chaos*. 2018. Vol. 28, paper 075310. 20 p.
57. Балабанов А.С. Верхняя граница для суммы корреляций трех индикаторов в отсутствии общего фактора. *Кибернетика и системный анализ*. 2019. № 2. С. 10–21.
58. Балабанов О.С. Від коваріацій до каузальності. Відкриття структур залежностей в даних. *Системні дослідження та інформаційні технології*. 2011. № 4. С. 104–118.
59. Colombo D., Maathuis M.H., Kalisch M. and Richardson T.S. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*. 2012. Vol. 40, Issue 1. P. 294–321.

60. Colombo D., Maathuis M.H. Order-independent constraint-based causal structure learning. *Jour. of Machine Learning Research.* 2014. Vol.15. P. 3921−3962.
61. Kernel-based conditional independence test and application in causal discovery / K.Zhang, J. Peters, D. Janzing, B. Schölkopf. / *Proc. of the 27th Conf. on Uncertainty in Artificial Intelligence*, (UAI-2011). Corvallis, Oregon: AUAI Press, 2011. P. 804–813.
62. Балабанов А.С. Минимальные сепараторы в структурах зависимостей. Свойства и идентификация. *Кибернетика и системный анализ*. 2008. № 6. Р. 17–32.
63. Балабанов О.С. Відтворення каузальних мереж на основі аналізу марковських властивостей. *Математичні машини та системи*. 2016. № 1. С. 16–26.
64. Granger C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*. 1969. Vol. 37. P. 424–459.
65. Swanson N.R. and Granger C.W.J. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *J. of the American Statistical Association*. 1997. Vol. 92, N 437, P. 357–367.
66. Gong M., Zhang K., Schölkopf B., Tao D. and Geiger P. Discovering temporal causal relations from subsampled data. *Proc. of the 32nd Intern. Conf. on Machine Learning*, 2015. P. 1898–1906.
67. Malinsky D. and Spirtes P. Learning the structure of a nonstationary vector autoregression. The 22nd Intern. Conf. on Artificial Intelligence and Statistics. *Proc. of Machine Learning Research*, *PMLR*, 2019, Vol. 89. P. 2986–2994.
68. Harford T. Big data: A big mistake? *Significance*. 2014. Vol. 11, N 5. P. 14–19.
69. Bühlmann P. and van de Geer S. Statistics for high-dimensional data: Methods, theory and applications. Springer, 2011. 556 p.
70. Donoho D.L. High-dimensional data analysis: the curses and blessings of dimensionality – In: American Mathematical Society Conf. "*Math Challenges of the 21st Century*", 2000, Los Angeles. P. 1–32.
71. Bareinboim E., Tian J., Pearl J. Recovering from selection bias in causal and statistical inference. *Proc. of the 28th AAAI Conf. on Artificial Intelligence*. 2014. P. 2419–2416. (July 27–31, 2014, Québec Convention Center, Québec City, Québec, Canada).

References

1. Balabanov O.S. Big Data Analytics: principles, trends and tasks (a survey). *Problems in programming*. 2019. N 2. P. 47–68. (ISSN 1727–4907) [In Ukrainian].
2. Bühlmann P., Drineas P., Kane M., van der Laan M. (eds.) Handbook of Big Data. Taylor and Francis, 2016. 456 p.
3. Mayer-Schönberger V., Cukier K. Big Data: A revolution that will transform how we live, work, and think. Boston, MA: Houghton Mifflin Harcourt, 2013. 256 p.
4. Chen C.L.P. and Zhang C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*. 2014. Vol. 275. P. 314–347.
5. Chen M., Mao S. and Liu Y. Big Data: A Survey. *Mobile Networks and Applications*. 2014. Vol. 19, Issue 2. P. 171–209.
6. Bhadani A. and Jothimani D. Big Data: Challenges, opportunities and realities / In.: M.K. Singh and D.G. Kumar (eds.). Effective Big Data management and opportunities for implementation. – IGI Global, Pennsylvania, USA, 2016. – [Електронний ресурс] Доступ: https://arxiv.org/pdf/1705.04928.
7. Oussous A., Benjelloun F.-Z., Lahcen A.A. and Belfkih S. Big Data technologies: A survey. *Journal of King Saud University. Computer and Information Sciences*. 2018. Vol. 30, Issue 4. P. 431–448.
8. Cao L. Data science: a comprehensive overview. *ACM Computing Surveys*. 2017. Vol. 50, N 3, Article 43, 42 p.
9. Gandomi A. and Haider M. Beyond the hype: Big data concepts, methods, and analytics. *Intern. Jour. of Information Management*. 2015. Vol. 35, N 2. P. 137–144.
10. Tsai C.-W., Lai C.-F., Chao H.-C. and Vasilakos A.V. Big data analytics: a survey. *Journal of Big Data*. 2015. Vol. 2, N 1. P. 1–32.
11. Watson H.J. Tutorial: Big Data analytics: Concepts, technologies, and applications. *Comm. of the Association for Information Systems*. 2014. Vol. 34, Article 65. P. 1247–1268.
12. Fan J., Han F. and Liu H. Challenges of Big Data analysis. *Nat. Scient. Rev*. 2014., Vol. 1, N 2. P. 293–314.
13. Franke B., Plante J.-F., Roscher R., Lee E.A., Smyth C., Hatefi A., Chen F., Gil E., Schwing A.G., Selvitella A., Hoffman M.M., Grosse R., Hendricks D. and Reid N. Statistical inference, learning and models in Big Data.

Intern. Statistical Review. 2016. Vol. 84, N. 3. P. 371–389.
14. Zafarani R., Abbasi M.A. and Liu H. Social media mining. An introduction. Cambridge University Press, 2019. 380 p.
15. Andon P.I. and Balabanov O.S. Vyjavlenie znanij i izyskanija v bazah dannyh. Podhody, modeli, metody i sistemy. *Problems in programming*. 2000. N 1–2. P. 513–526. (Kyjv, UA). [In Russian].
16. Balabanov O.S. Knowledge extraction from databases – advanced computer technologies for intellectual data analysis. *Mathematical Machines and Systems*. 2001. N 1–2. P. 40–54. [In Russian].
17. Azzalini A. and Scarpa B. Data analysis and Data Mining: An introduction. – N.Y.: Oxford University Press, 2012. 288 p.
18. Swanson N.R. and Xiong W. Big Data analytics in economics: What have we learned so far, and where should we go from here? *Canadian J. of Economics*. 2018, Vol. 51, Issue 3. P. 695–746.
19. Graham E. and Timmermann A. Forecasting in Economics and Finance. *Annual Review of Economics*. (2016). Vol. 8. P. 81–110.
20. Weihs C. and Ickstadt K. Data Science: the impact of statistics. *Intern. J. of Data Science and Analytics*. 2018. Vol. 6. P. 189–194.
21. The role of statistics in the era of big data. Special issue of the journal: *Statistics and Probability Letters*. May 2018. Vol. 136.
22. Secchi P. On the role of statistics in the era of big data: A call for a debate. *Statistics and Probability Letters*. 2018. Vol. 136. P. 10–14.
23. Witten I.H., Eibe F., Hall M.A. (3rd ed.).Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2011. 629 p.
24. Maimon O., Rokach L. (Eds.) Data Mining and Knowledge Discovery Handbook. 2nd ed., Springer-Verlag New-York Inc., 2010. 1285 p.
25. Murphy K.P. Machine learning: a probabilistic perspective. MIT Press, Cambridge, Massachusetts, 2012. 1055 p.
26. Hastie T., Tibshirani R. and Friedman J. The elements of statistical learning. (2nd ed.). Springer. 2009. 745 p.
27. Efron B. and Hastie T. Computer age statistical inference. Cambridge University Press, 2016. 475 p.
28. Efron B. Large-scale inference. Stanford University Press, 2010. 263 p.
29. James G., Witten D., Hastie T. and Tibshirani R. An introduction to statistical learning with applications in R. Springer, N.Y., 2013. 426 p.
30. Berkhin P. A survey of clustering data mining techniques. In: Kogan J., Nicholas C., Teboulle M. (eds.). *Grouping multidimensional data*. Springer-Verlag: Berlin-Heidelberg, 2006. P. 25–71.
31. Bouveyron C., Brunet-Saumard C. Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*. 2014. Vol. 71. P. 52–78.
32. Kurban H., Jenne M. and Dalkilic M.M. Using data to build a better EM: EM* for big data. *Intern. J. of Data Science and Analytics*. 2017. Vol. 4, Issue 2. P. 83–97.
33. LeCun Y., Bengio Y., Hinton G. Deep learning. *Nature*. 2015. Vol. 521, P.436–444.
34. Esling P. and Agón C. Time-series data mining. *ACM Computing Surveys*. 2012. Vol. 45, Issue 1. P. 12–34.
35. Chandola V., Banerjee A. and Kumar V. Anomaly detection for discrete sequences: a survey. *IEEE Trans. on Knowledge and Data Eng*. (TKDE). 2012. Vol. 24, N 5. P. 823–839.
36. Truong C., Oudre L. and Vayatis N. Selective review of offline change point detection methods. [Electronic resource] URL: https://arxiv.org/abs/1801.00718.
37. Aminikhanghahi S. and Cook D.J. A survey of methods for time series change point detection. *Knowledge and Information Systems*. 2017. Vol. 51, Issue 2. P. 339–367.
38. Frick K., Munk A. and Sieling H. Multiscale change point inference. *J. Roy. Statist. Soc., ser. B*. 2014. Vol. 76, Pt. 3. P. 495–580.
39. Wang T. and Samworth R.J. High dimensional change point estimation via sparse projection. *J. Roy. Statist. Soc., ser. B*. 2018. Vol. 80, Pt. 1. P. 57–83.
40. Liao T.W. Clustering of time series data – a survey. *Pattern Recognition*. 2005. Vol. 38. P. 1857–1874.
41. Atluri G., Karpatne A. and Kumar V. Spatio-temporal Data Mining: a survey of problems and methods. *ACM Computing Surveys*. 2018. Vol. 51, Issue 4, Article N 83.
42. Lee T.-W., Girolami M., Bell A.J., Sejnowski T.J. A unifying information-theoretic framework for Independent Component Analysis. *Intern. J. Computers and Mathematics with Applications*. 2000. Vol. 39. P. 1–21.
43. Neville J. and Jensen D. Relational Dependency Networks. *Jour. of Machine Learning Res*. 2007. Vol. 8. P. 653–692.
44. De Raedt L., Kersting K., Natarajan S. and Poole D. Statistical relational artificial intelligence: Logic, probability, and computation. *Synthesis Lectures on Artificial*

Intelligence and Machine Learning. 2016. Vol. 10, N 2. P.1–89.
45. Kazemi S.M., Buchman D., Kersting K., Natarajan S. and Poole D. Relational logistic regression: The directed analog of Markov logic networks. *Workshops at the Twenty-Eighth AAAI Conf. on Artificial Intelligence*. 2014. P. 41–43.
46. Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge Univ. Press, 2000. 526 p.
47. Spirtes P., Glymour C. and Scheines R. Causation, prediction and search. New York: MIT Press, 2001. 543 p.
48. Peters J., Janzing D. and Schölkopf B. Elements of Causal Inference. Foundations and Learning Algorithms. MIT Press, Cambridge, MA, USA, 2017. 265 p.
49. Balabanov O.S. Knowledge discovery in data and causal models in analytical informatics. *Problems in programming*. 2017. N 3. P. 96–112. (ISSN 1727–4907). [in Ukrainian].)
50. Raghu V.K., Ramsey J.D., Morris A., Manatakis D.V., Sprites P., Chrysanthis P.K., Glymour C., Benos P.V. Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *Intern. Jour. of Data Science and Analytics*. 2018. Vol. 6, Issue 1. P. 33–45.
51. Tsagris M., Borboudakis G., Lagani V., Tsamardinos I. Constraint-based causal discovery with mixed data. *Intern. Jour. of Data Science and Analytics*. 2018. Vol. 6, Issue 1. P. 19–30.
52. Pearl J. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*. 2019. Vol. 62, Issue 3. P. 54–60.
53. Pearl J. and Bareinboim E. External validity: From do-calculus to transportability across populations. *Statistical Science*. 2014. Vol. 29, N 4. P. 579–595.
54. Malinsky D. and Spirtes P. Causal structure learning from multivariate time series in settings with unmeasured confounding. *Proc. of 2018 ACM SIGKDD Workshop on Causal Discovery*, August 2018, London, UK. PMLR, Vol. 92. P. 23–47.
55. Entner D. and Hoyer P.O. On causal discovery from time series data using FCI. *Proc. of the 5th European Workshop on Probabilistic graphical models*. 2010, Helsinki, Finland. P. 121–128.
56. Runge J. Causal network reconstruction from timeseries: From theoretical assumptions to practical estimation. *Chaos*. 2018. Vol. 28, paper 075310. 20 p.
57. Balabanov O.S. Upper bound on the sum of correlations of three indicators under the absence of a common factor. *Cybernetics and Systems Analysis*. 2019. Vol. 55, N 2. P. 174–185.
58. Balabanov O.S. From covariation to causation: Discovery of dependency structures in data. *System research and information technologies*. 2011. N 4, P. 104–118. [In Ukrainian]
59. Colombo D., Maathuis M.H., Kalisch M. and Richardson T.S. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*. 2012. Vol. 40, Issue 1. P. 294–321.
60. Colombo D., Maathuis M.H. Order-independent constraint-based causal structure learning. *Jour. of Machine Learning Research.* 2014. Vol.15. P. 3921−3962.
61. Kernel-based conditional independence test and application in causal discovery / K.Zhang, J. Peters, D. Janzing, B. Schölkopf. / *Proc. of the 27th Conf. on Uncertainty in Artificial Intelligence*, (UAI-2011). Corvallis, Oregon: AUAI Press, 2011. P. 804–813.
62. Balabanov A.S. Minimal separators in dependency structures: Properties and identification. *Cybernetics and Systems Analysis*. 2008. Vol. 44, N 6. P. 803–815.
63. Balabanov O.S. Vidtvorennya kauzalnych merezh na osnovi analizu markovskich vlastyvostej [Reconstruction of causal networks via analysis of Markov properties]. *Mathematical Machines and Systems*. 2016. N 1. P. 16–26. [In Ukrainian]
64. Granger C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*. 1969. Vol. 37. P. 424–459.
65. Swanson N.R. and Granger C.W.J. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *J. of the American Statistical Association*. 1997. Vol. 92, N 437, P. 357–367.
66. Gong M., Zhang K., Schölkopf B., Tao D. and Geiger P. Discovering temporal causal relations from subsampled data. *Proc. of the 32nd Intern. Conf. on Machine Learning*, 2015. P. 1898–1906.
67. Malinsky D. and Spirtes P. Learning the structure of a nonstationary vector autoregression. The 22nd Intern. Conf. on Artificial Intelligence and Statistics. *Proc. of*

Machine Learning Research, *PMLR*, 2019, Vol. 89. P. 2986–2994.
68. Harford T. Big data: A big mistake? *Significance*. 2014. Vol. 11, N 5. P. 14–19.
69. Bühlmann P. and van de Geer S. Statistics for high-dimensional data: Methods, theory and applications. Springer, 2011. 556 p.
70. Donoho D.L. High-dimensional data analysis: the curses and blessings of dimensionality – In: American Mathematical Society Conf. "*Math Challenges of the 21st Century*", 2000, Los Angeles. P. 1–32.
71. Bareinboim E., Tian J., Pearl J. Recovering from selection bias in causal and statistical inference. *Proc. of the 28th AAAI Conf. on Artificial Intelligence*. 2014. P. 2419–2416. (July 27–31, 2014, Québec Convention Center, Québec City, Québec, Canada).

Одержано 17.07.2019

Про автора:

Балабанов Олександр Степанович,
доктор фізико-математичних наук,
провідний науковий співробітник.
Кількість наукових публікацій в
українських виданнях – 60.
Кількість наукових публікацій в
зарубіжних виданнях – 12.
Індекс Хірша – 6.
http://orcid.org/0000-0001-9141-9074.

***Місце роботи автора*:**

Інститут програмних систем
НАН України,
03187, м. Київ-187,
проспект Академіка Глушкова, 40.
Тел.: (044) 526 3420.
E-mail: bas@isofts.kiev.ua