УДК 519.816
DOI: 10.20535/SRIT.2308-8893.2016.2.10

ВИЯВЛЕННЯ ТА ОБРОБЛЕННЯ НЕВИЗНАЧЕНОСТЕЙ У ФОРМІ НЕПОВНИХ ДАНИХ МЕТОДАМИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ

Н.В. КУЗНЄЦОВА

Розглянуто методи оброблення пропущених даних і запропоновано їх класифікацію з урахуванням видів вхідних даних, типів та форматів даних, причин пропусків, зумовлених проявом впливу невизначеностей навколишнього світу і об'єкта моделювання. Досліджено спільні ознаки та відмінності існуючих методів оброблення, визначено особливості їх застосування для дозаповнення пропущених даних залежно від характеру невизначеностей. Показано, що традиційний підхід до заповнення пропусків середнім значенням не дозволяє отримати достовірні прогнози у багатьох випадках через зміну характеру вибірки. Запропоновано використання методів інтелектуального аналізу даних для оброблення пропущених значень та наведено приклад заповнення пропусків даних методами регресійного аналізу, зокрема за допомогою оцінок прогнозів.

ВСТУП

Невизначеності навколишнього світу та об'єктів, що у ньому функціонують, безпосередньо або опосередковано впливають на діяльність людини, потребують урахування під час прогнозування розвитку множини супутніх процесів. Очікувані результати від застосування тих чи інших засобів можуть бути незадовільними або зовсім непередбачуваними внаслідок дії випадкових зовнішніх факторів. Коли ж зовнішні фактори строго визначені або хоча б відомі, то невизначеність природи та обмежень може бути врахована і, відповідно, можна запропонувати методи їх оброблення. Так, у задачах системного аналізу [1, 2] у загальному випадку розрізняють три основні види невизначеностей: невизначеність цілей; ситуаційну і природну невизначеність (невизначеність знань про можливі ситуації у процесі функціонування складних систем); інформаційну невизначеність (невизначеність поведінки навколишнього середовища та дій реального партнера чи супротивника).

Розв'язування задачі розкриття концептуальної невизначеності щодо системного аналізу потребує розкриття множини різнорідних невизначеностей на підставі єдиних принципів, прийомів і критеріїв [3]. На практиці розв'язуються задачі дослідження невизначеності цілей розроблення і перспектив конкурентоспроможності виробу, аналізуються невизначеність динаміки ринків попиту та пропозиції і невизначеність активної протидії конкурентів, невизначеність динаміки розроблення, виробництва, збуту та експлуатації певного виробу тощо.

Одним з проявів інформаційної невизначеності є невизначеність, зумовлена пропусками даних. Об'єктивні характеристики певних процесів можуть бути змінені або навіть спотворені внаслідок втрати частини даних під час

їх отримання, передавання чи зберігання. Постає потреба у відновленні таких пропущених даних і, що важливо, у підбиранні тих алгоритмів, за якими вони будуть відновлюватись, оскільки неправильне або недостатньо достовірне відновлення може завдати більше шкоди, ніж самі пропуски даних.

Роботу присвячено аналізу інформаційної невизначеності у формі пропусків статистичних даних та методів заповнення наявних пропусків з метою підвищення адекватності математичних моделей та оцінок прогнозів, які обчислюються за цими моделями.

ПОСТАНОВКА ЗАВДАННЯ

Мета дослідження — аналіз та класифікація методів оброблення пропусків даних для узгодження даних різних типів та форматів, зумовлених проявом впливу невизначеностей навколишнього світу й об'єкта моделювання; розроблення рекомендацій для розроблення коректного підходу до оброблення неповних даних, які дадуть змогу підвищити прогнозну якість моделей, побудованих на відновлених за цією методикою пропущених даних. Зокрема, це завдання є важливим для оброблення вибірок невеликих розмірів, коли некоректне оцінювання пропущених змінних є вкрай небажаним і може спричинити похибки подальшого прогнозування поведінки системи і побудови прогнозних моделей. У роботі будуть проаналізовані спільні ознаки та відмінності існуючих методів та особливості алгоритмічних засобів оброблення пропущених значень.

НЕПОВНОТА ДАНИХ ЯК РЕАЛІЗАЦІЯ НЕВИЗНАЧЕНОСТІ

Неповнота даних зумовлюється такими причинами: пропусками, неуважністю під час уведення інформації; браком інформації з об'єктивних причин; незнанням; некомпетентними відповідями на поставлені запитання, зокрема, через зумисне приховування інформації [4]. Залежно від причини пропуски можуть істотно впливати на результати та спричиняти значні збитки організації, яка вчасно не отримала необхідну інформацію.

Невизначеності насправді трапляються у повсякденному житті. Потреба у моделюванні та прогнозуванні за неповними даними виникає у різних сферах: фінансах, транспорті, виробництві, сільському господарстві, логістиці, фізиці, соціології тощо.

Поглиблене вивчення процесів за допомогою математичних моделей дозволяє дослідити кількісні зв'язки між вхідними та вихідними змінними, а також фактори, які впливають на вихідні змінні при варіації вхідних у широкому діапазоні, і розглянути поведінку процесів на будь-яких часових інтервалах у прийнятному масштабі часу. Математична модель, що будується для цієї мети, може бути надскладною і трудомісткою, оскільки вона має враховувати тонкощі взаємодії кількісних і якісних змінних із можливим урахуванням реального часу, тобто з використанням імітаційного моделювання. За допомогою математичних моделей можна виявити ефекти і явища, які недоступні безпосереднім спостереженням за допомогою приладів. Крім цього, під час проектування нових систем у різних галузях можна швидко змінювати варіанти реалізації системи завдяки можливості її швидкого

дослідження на моделі, виявити вплив початкових умов та обмежень на ключові змінні.

Прогнозування значень змінних виконується, як правило, на основі набагато простіших моделей ніж поглиблене вивчення процесів. Таке спрощення моделі також може внести додаткову інформаційну невизначеність.

Поняття структури моделі охоплює такі параметри: порядок, вимірність моделі, наявність нелінійностей і їх характер, час запізнення (для часових рядів), тип збурень тощо.

Вибір структури моделі, що адекватна процесу, є непростою задачею, що розв'язується в інтерактивному режимі. Спочатку структуру моделі оцінюють наближено на підставі дослідження закономірностей перебігу процесу, аналізу кореляційних функцій, візуального аналізу даних. При цьому вибирають декілька найбільш імовірних структур (кандидатів). Потім обчислюють оцінки параметрів моделей-кандидатів і вибирають оптимальну з них, використовуючи відповідні статистичні характеристики якості моделей.

Якщо жодна з моделей-кандидатів не може вважатися адекватною для конкретного застосування, то необхідно досліджувати на інформативність експериментальні дані, які можуть бути недостатньо інформативними для оцінювання моделі. У такому випадку потрібно буде повторно чи додатково збирати експериментальні дані (якщо це взагалі можливо) і коригувати структуру моделі.

Наприклад, розглянемо задачу визначення місця розташування транспортних засобів для контролю комунального транспорту системою EasyWay у разі неповних даних від GPS і маршруту складної форми. Інформація про місце розташування необхідна для прогнозування часу прибуття транспорту на зупинку. Щоб його розрахувати, можна використати найпростішу структуру моделі, що враховує відстань S і швидкість руху транспорту v: $t = \frac{S}{v}$. Така модель не враховує нерівномірність руху транспорту, наявність перешкод на шляху, особливості дорожнього покриття, погодні умови тощо. І навіть уточнена модель не може врахувати всі фактори, зокрема кількість пасажирів та час їх посадки на кожній зупинці. Прогнозований час можна показувати на сайтах, мобільних додатках та інформаційних табло на зупинках, що є важливим і зручним для пасажирів, зменшує час очікування і робить рух транспорту більш передбачуваним. У випадку, коли транспорт перебуває там, де сигнал GPS слабкий або його немає, постає питання прогнозування неповних даних (пропущених даних сигналу GPS) для уточнення місцеперебування та прогнозування орієнтовного часу. Тут може бути і невизначеність стану природи, зумовлена ситуаційною невизначеністю — можливим випаданням опадів, створенням складностей проїзду, аварійних ситуацій тощо. Для прогнозування неповних даних можуть застосовуватись різні методи і підходи залежно від причин появи таких невизначеностей, установлених існуючих і відомих закономірностей.

Поняття «розширена невизначеність» виникає під час оброблення результатів вимірювання у фізиці, метрології, географії, військовій справі. *Розширена невизначеність* (expanded uncertainty) — це величина, що визначає довірчий інтервал для результату вимірювання, у межах якого ймовірно міс-

титься більша частина розподілу значень, які обґрунтовано можуть бути приписані вимірюваній величині.

Таким чином, розширена невизначеність визначає межі *інтервалу невизначеності* для результату вимірювання y. Права межа цього інтервалу: $y+U$, а ліва: $y-U$. Величина розширеної невизначеності, а отже, і ширина цього інтервалу, залежать від обраного під час розрахунку рівня довіри p, який менший або дорівнює одиниці [5].

Значення рівня довіри повинно бути досить великим, щоб була висока впевненість у тому, що інтервал невизначеності містить істинне значення. Водночас із підвищенням p ширина інтервалу збільшується, що ускладнює його практичне використання для прийняття рішень за результатами вимірювань. Тому доводиться вибирати у певному розумінні «компромісне» значення рівня довіри. У більшості випадків значення p припускають рівним 0,95. Це означає, що інтервал невизначеності включатиме 95% усіх значень, які можуть бути результатом вимірювання, або з імовірністю 0,95 покриватиме істинне значення вимірюваної фізичної величини. Разом з тим під час особливо відповідальних вимірювань, які мають великий вплив на життя чи здоров'я людей, значення рівня довіри може досягати 0,99 і більше.

Інформаційна невизначеність часто виникає у задачах оброблення статистичних даних і пов’язана з недоотриманням, запізненням або втратою частини інформації з будь-яких причин. Це притаманно фінансовій, економічній і соціологічній галузях. Аналіз таких причин може дати додаткове розуміння суті пропусків і допомогти у виборі моделі їх заповнення.

ІСНУЮЧІ МЕТОДИ ЗАПОВНЕННЯ ПРОПУСКІВ ДАНИХ

Існує багато засобів заповнення пропусків уже після етапу збирання даних: заповнення середнім значенням, пропорційне розміщення спостережень з пропущеними даними за вже існуючими градаціями шкали, розрахунок можливого значення за допомогою регресійної моделі тощо.

Зрозуміло, що використання будь-яких засобів заповнення пропусків може змістити структуру вибірки, яка буде отримана на основі існуючих неповних даних, у бік структури неповних даних, що може спотворити реальний розподіл спостережень у вибірці і зменшити фактичну значущість отриманих результатів.

Обираючи конкретний алгоритм для заповнення пропусків, варто враховувати, що можливість його застосування істотно залежить від методу аналізу даних, який передбачається використати надалі.

Сьогодні існують алгоритми, які дають змогу обробляти пропуски необхідною інформацією, такі як метод Hot Deck, метод Барлета, алгоритми Resampling, Zet, Zetbraid, EM-оцінювання, регресійне моделювання та прогнозування значень [6–9]. Особливістю цих алгоритмів є заповнення пропусків значеннями, які підбираються самим алгоритмом.

Метод Hot Deck. Цей метод використовує підстановку замість пропущеного значення найближчого інформаційного об’єкта. Пропущені дані можна підбирати як з усієї сукупності повних спостережень, так і з деякої

підгрупи — кластера, до якого належить цільовий об'єкт. Для заповнення пропуску за обраною характеристикою цільового об'єкта використовується значення цієї характеристики в об'єкта, найближчого до цільового. Тип функції відстані для визначення спостереження, найближчого до цільового (з пропуском), вибирається виходячи з типу досліджуваних даних, уявлень про характер зв'язку між змінними і завдання конкретного дослідження.

Метод Барлета. Цей метод складається з двох етапів: підстановки замість пропуску початкових згенерованих значень на першому етапі; проведення на другому етапі коваріаційного аналізу цільової змінної і побудова дихотомічного індикатора повноти спостережень за цільовою змінною. Індикатор повноти спостережень завжди дорівнює 0, за винятком одного єдиного випадку: *i*-е значення — це цільова змінна і воно є пропущеним, тоді індикатор набуває значення 1 [8].

Алгоритм ZET. Суть цього алгоритму полягає у підборі кожного значення для заповнення пропуску не за всією сукупністю спостережень, а з деякої її частини, яка називається компонентною матрицею, що складається з компонентних рядків і стовпців. Компонентність деякого рядка являє собою величину, обернено пропорційну декартовій відстані за цільовим рядком (неповного спостереження з пропуском) у просторі, осями якого задані змінні — характеристики об'єктів [7, 9].

За даними компонентної матриці надалі будується функціональна залежність прогнозного значення від відповідного значення у компонентній матриці, на основі якої потім прогнозується значення пропуску.

Алгоритм ZetBraid. Основна відмінність цього методу від попереднього полягає у тому, що в цьому алгоритмі закладено механізм об'єктивного відбору розмірності компетентної матриці. При роботі алгоритму відбувається послідовний почерговий відбір компетентних рядків та стовпців і щоразу формується нова компетентна матриця. Потім за заданим критерієм визначається її ефективність при прогнозуванні пропусків [7].

Resampling. Це ітеративний метод, який передбачає зміну рядків з пропущеними даними випадково вибраними рядками з матриці повних спостережень, а далі будується регресійне рівняння для прогнозування пропущеного значення. Процедуре регресійного моделювання повторюють декілька разів, після чого значення отриманих регресійних коефіцієнтів усереднюють і отримують кінцеве значення, яке дає максимальну точність прогнозу пропущеного значення [8].

Множиннна вставка. Метод розроблений у 1970-х рр. XX ст. Дональдом Рубіним [10]. Технологія множинної вставки пропусків передбачає підстановку одразу кількох значень замість кожного пропущеного. Значна розбіжність цих значень означає невизначеність моделі і не дозволяє зробити висновки про їх типи і причини появи. Дані, що містять набір заповнених пропусків, зберігаються в окремих масивах, кожен з яких потім аналізується як такий, що містить повні спостереження без пропусків.

Наразі цей метод вважається доволі перспективним і реалізований у більшості комерційних програмних додатків.

ЕМ-оцінювання [11]**.** Метод максимізації математичного сподівання (ЕМ — expectation maximization) або ЕМ-оцінювання надає можливість не лише відтворювати пропущені значення з використанням двоетапного іте-

ративного алгоритму, але й оцінювати середнє значення, коваріаційні та кореляційні матриці для кількісних змінних. ЕМ-алгоритм у загальному випадку являє собою ітераційну процедуру, призначену для розв'язання задач оптимізації деякого функціонала через аналітичний пошук екстремуму функції.

На Е-кроці обчислюється очікуване значення (expectation) вектора прихованих змінних G за поточним наближенням вектора параметрів Θ. На *M*-кроці розв'язується задача максимізації правдоподібності і обчислюється наступне наближення вектора Θ за поточними значеннями векторів G і Θ.

Ідею реалізації ЕМ-алгоритму можна подати так:

– обчислити початкове наближення вектора параметрів Θ;

– повторювати:

$G = ESTEP(\Theta)$,

$\Theta = MSTEP(G, G)$;

– поки G і Θ не стабілізуються (настає збіжність до усталених значень).

У класичному варіанті алгоритму формально задачу максимізації очікування можна виразити таким чином: $Q^{m+1} = \arg\max_{\Theta} Q(\Theta; \Theta^m)$. Тут Θ означає розраховане очікуване умовне значення пропущеної характеристики для певного спостереження.

Регресійне моделювання [12]**.** Пропущені значення за допомогою регресійних моделей відновлюються за два етапи.

1. На першому етапі за сукупністю повних спостережень будується регресійна модель і оцінюються коефіцієнти рівняння, де залежною змінною є цільова змінна — пропущене значення, яке необхідно відновити.

2. За отриманим на попередньому етапі рівнянням, у яке підставляються відомі значення незалежних змінних (предикторів), для кожного цільового об'єкта розраховується пропущене значення за залежною цільовою змінною. У випадку інтервальних та абсолютних змінних розраховується конкретне значення, а для порядкових і номінальних значень з деякою ймовірністю передбачається категорія, до якої має бути віднесений об'єкт.

Вибір типу регресійної моделі для розрахунку пропущених значень змінної визначається кількістю вимірювань цільової залежної змінної (значення якої необхідно відновити) і незалежних змінних, за якими передбачатимуться пропущені значення.

У праці [13] розглядаються можливості оцінювання пропусків даних за допомогою *байєсівського компонентного аналізу та локального методу найменших квадратів* і порівнюються можливості їх сукупного використання. Також показано, що спільне використання обох методів дає змогу отримати вищу якість прогнозів пропущених значень, але при цьому істотно збільшуються обчислювальні витрати.

Для оцінювання і відновлення пропущених значень надзвичайно важливою є оцінка втрат інформації через неповноту спостережень і якість оцінок пропущених значень залежно від типу цільової змінної та частки пропусків початкових даних.

Зрозуміло, що коректність і ефективність роботи цих алгоритмів визначаються підбиранням найбільш подібного значення до пропуску, а для цього необхідно враховувати причину пропуску даних. Сучасні комп'ютерні аналітичні системи, такі як SPSS, GeNIe, SAS Enterprise Miner, грунтуються на використанні логічних дерев для умовного обчислення значень та їх заміни на середні величини або медіани. Зокрема, для розроблення скорингових карт рекомендуються методи підстановки [14], що враховують інші характеристики даних. Однак присвоєння найбільш часто вживаних значень або середніх значень спричинить так звані «сплески», що спотворить реальну ситуацію з розподілом груп у вибірках і призведе до втрати надзвичайно важливої інформації. Тому пропонується виносити пропущені дані в окрему групу, замінюючи пропущені значення певним спеціальним значенням поза нормальними значеннями і включати їх в аналіз як окрему категорію.

Багато аналітиків мають переконання, що пропущені значення не потрібно взагалі включати в аналіз і вилучити їх цілком з начальної вибірки даних. Такий метод корисний, якщо аналітики не схильні накладати додаткові ризики того, що пропущені значення будуть відновлені некоректно і таким чином можуть збільшити ризик віднесення таких випадків до нормальної категорії. Однак цей метод нераціональний у випадку, коли обсяг навчальної вибірки надзвичайно малий і видалення таких даних є критичним, або коли необхідно побудувати скорингові моделі, що відображають реальні, а не «ідеальні» дані і містять пропущені значення. Такі дані необхідно додатково обробляти до прийняття рішення.

Загальну класифікацію методів заповнення пропусків, що використовуються в різних інструментальних засобах інтелектуального аналізу даних, показано на рисунку. Для детального аналізу обрано середовище SAS Enterprise Miner, у якому реалізовано різні методи заміни пропущених значень, а також передбачено відсутність обов'язкової заміни. У SAS Enterprise Miner окремо передбачено можливість вставки для вхідних вузлів та цільових змінних, а також є можливість такого заповнення не на всій вибірці даних, а спочатку на навчальній вибірці, і у випадку отримання задовільних результатів — поширити таку заміну і на перевірну вибірку. Розглянемо детальніше різні методи заповнення пропусків залежно від типу змінних (категоріальні або неперервні).

Категоріальні змінні

Якщо як пропущені спостереження обрати дані за категоріальною змінною, то можливі такі методи заміщення:

Count **—** заміна пропущеного значення для категоріальної змінної найчастішим значенням спостереження.

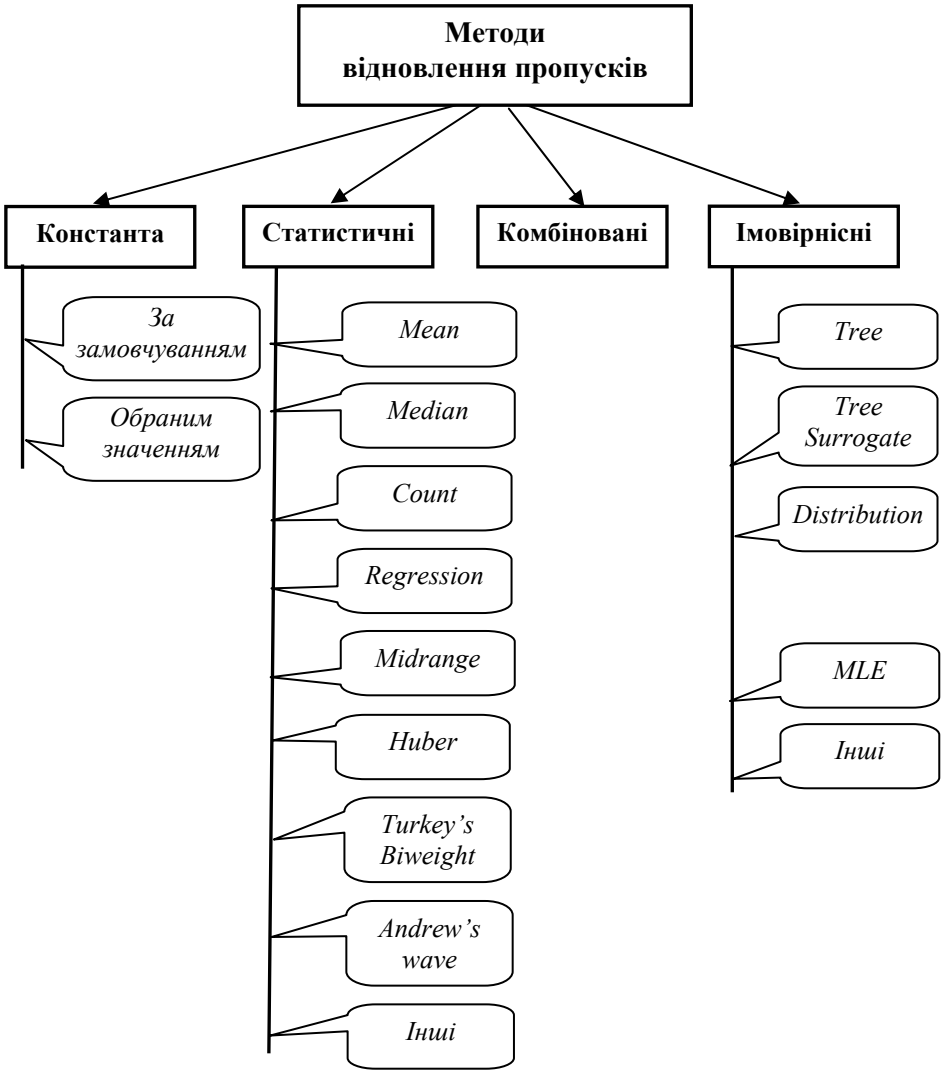
Default Constant Value **—** заміна введеним вручну значенням для категоріальної змінної.

Distribution — заміна значенням, розрахованим за ймовірнісним розподілом для наявних спостережень. Такий спосіб не спотворює розподіл вибірки.

Tree **—** заміна пропуску оціненим у результаті цільового аналізу значенням. Вхідні і відбраковані змінні використовуються як предиктори. Змінні, які важливі для моделі як цільові, не можуть бути використані для

заповнення. Оскільки відновлені значення для кожної змінної з пропусками ґрунтуються на інших вхідних змінних, то такий метод має бути точнішим.

Tree surrogate — використовується попередній метод дерева щеплення з наступною модифікацією правил щеплення. Правило заміни є зворотним до головного правила щеплення дерева. Коли правило щеплення діє на змінну, яка має пропуски, застосовується правило заміни. Якщо пропущені значення запобігають головному правилу виконати щеплення і всім правилам заміни спостережень, то головне правило призначає спостереження у гілці, що приведе до отримання відсутніх значень.



Класифікація методів заповнення пропущених даних

Неперервні (статистичні) змінні

Mean — заміна пропущених інтервальних значень середнім арифметичним. Це незміщена оцінка середнього популяції. *Mean* є найпоширенішою стати-

стикою для заміни пропущених значень, якщо значення змінної мають приблизно симетричний розподіл (наприклад, дзвоноподібний нормальний розподіл). Цей метод використовують за замовчуванням для неперервних змінних з пропущеними значеннями.

Median – використовується певне середнє, установлене для заміни пропущеного інтервального значенння 50-го процентилю, яке є середнім значенням або середнім арифметичним двох середніх значень для множини чисел, розміщених у порядку зростання. Середнє і медіана однакові для симетричного розподілу. Медіана менш чутлива до екстремальних значень, ніж середнє або півсума крайніх значень. Таким чином, медіана підходить краще для заміни відсутніх значень для змінних, які мають спотворені розподіли. Медіана також використовується для порядкових даних.

Midrange — використовується параметр півсуми крайніх значень (середній діапазон) для заміни відсутніх неперервних значень змінної значенням суми максимального значення для змінної плюс мінімального значення для змінної, поділеної на два. Midrange є швидше відображенням тенденції; його легко розрахувати.

Методи *Distribution, Tree, Tree Surrogate* реалізуються аналогічно тому, як це виконується для категоріальних змінних.

Mid-minimum Spacing — використовується середній мінімальний інтервал, застосовується числова константа для визначення пропорції даних, що включаються в інтервал.

Huber — метод, у якому для заміни пропущеного значення використовується оцінка, описана нижче [15]. У разі, коли лінійна регресійна функція втрат, визначена як $l(r) = \sum_i r_i^2$, швидко зростає зі збільшенням значень залишків, тоді альтернативним є використанням абсолютного значення функції втрат замість квадрата залишків, тобто $l(r) = \sum_i |r_i|$.

Елегантним компромісом між цими двома функціями втрат стала запропонована Пітером Хубером у 1964 р. така функція [15]:

$$l(r) = \sum_i \rho(r_i), \text{ де } \rho(r_i) = \begin{cases} r_i^2, & \text{if } |r_i| \le c, \\ c(2|r_i| - c), & \text{if } |r_i| > c. \end{cases}$$

Хубер вважав, що правильним вибором є значення $c = 1{,}345$, і показав, що асимптотично це 95%-й інтервал. Цей метод так само ефективний як і метод найменших квадратів, якщо реальний розподіл близький до нормального (і набагато ефективніший у багатьох інших випадках).

Tukey's Biweight — метод, у якому оцінка для функції втрат визначається за критерієм Tukey's Biweight (відомим також як Tukey's bisquare) [15, 16]:

$$\rho'(r_i) = \begin{cases} r_i \left(1 - \left(\frac{r_i}{c} \right)^2 \right), & \text{if } |r_i| \le c, \\ 0, & \text{if } |r_i| > c. \end{cases}$$

Для цієї функції втрат зазвичай використовується значення $c = 4{,}685$; воно забезпечує асимптотичну ефективність на рівні 95%, так само, як і лінійна регресія для нормального розподілу.

Andrew's wave – метод, згідно з яким оцінка визначається так:

$$w(r_i) = \begin{cases} \dfrac{c}{\pi\, r_i} \sin\left(\dfrac{\pi r_i}{c}\right), & \text{if} \quad |\, r_i \,| \le c, \\ 0, & \text{if} \quad |\, r_i \,| > c. \end{cases}$$

За замовчуванням $c = 1{,}34\pi$.

Default Constant — пропуск замінюється визначеним уведеним символом.

ПРИКЛАД ЗАПОВНЕННЯ ПРОПУСКІВ ОЦІНКАМИ ПРОГНОЗІВ

Для заповнення невеликої кількості пропусків можна скористатись моделлю авторегресії першого порядку AP:

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k), \quad E[\varepsilon(k)] = 0. \tag{1}$$

Збільшимо незалежну змінну, час на одиницю і запишемо рівняння знову:

$$y(k+1) = a_0 + a_1 y(k) + \varepsilon(k+1).$$

Якщо коефіцієнти a_0, a_1 відомі, то можна знайти умовне математичне сподівання на основі відомої інформації до моменту k включно:

$$E_k[y(k+1)] = E_k[y(k+1) \,|\, y(k), y(k-1), \ldots, \varepsilon(k), \varepsilon(k-1), \ldots] =$$

$$= a_0 + a_1 E_k[y(k)] = a_0 + a_1\, y(k),$$

оскільки $y(k)$ у момент k є відомою константою. За аналогією запишемо рівняння (1) для моменту $k+2$

$$y(k+2) = a_0 + a_1 y(k+1) + \varepsilon(k+2)$$

і знайдемо умовне математичне сподівання:

$$E_k[y(k+2)] = a_0 + a_1 E_k[y(k+1)] = a_0 + a_1 E_k[a_0 + a_1 y(k)] =$$

$$= a_0 + a_0 a_1 + a_1{}^2 y(k).$$

Для наступного моменту часу маємо:

$$E_k[y(k+3)] = a_0 + a_0 a_1 + a_0 a_1{}^2 + a_1{}^3 y(k).$$

Таким чином, для загального випадку прогнозування на s кроків можна записати:

$$\hat{y}(k+s) = E_S[y(k+s)] = a_0 \left(\sum_{i=0}^{S-1} a_1{}^i \right) + a_1{}^S y(k) = a_0 \sum_{i=0}^{S-1} a_1{}^i + a_1{}^S y(k). \tag{2}$$

Рівняння (2) називають функцією прогнозування на довільну кількість кроків s. Прогноз являє собою збіжний процес, якщо $|a_1| < 1$, тобто

$$\lim_{s\to\infty} E_k[y(k+s)] = \frac{a_0}{1-a_1}, \quad |a_1| < 1, \tag{3}$$

де a_1 — знаменник геометричної прогресії. Вираз (3) свідчить про те, що для будь-якого стаціонарного процесу АР чи АРКС оцінка умовного прогнозу асимптотично при $s \to \infty$ збігається до безумовного середнього.

Виконаний порівняльний аналіз різних методів заповнення пропущених значень показав, що поняття «найкращого» методу для заміни пропущених є некоректним. Вибір методу може істотно залежати не лише від конкретної предметної галузі, у якій ці пропущені значення трапляються, а й від припущень аналітика щодо типу розподілу реальних (пропущених) даних. Найчастіше аналітики застосовують метод середнього для заміни пропущених значень, а це означає, що робиться припущення про належність даних до нормального розподілу (а це швидше виняток з правил). Заміна пропущених значень середнім, медіаною або іншою оцінкою є звичайно більш простим способом, однак це може істотно спотворити істинний розподіл вибірки. Тобто такі заміни можливі лише у випадку мінімального впливу на характер вибірки.

ВИСНОВКИ

У реальних задачах оброблення статистичних даних найбільшою складністю залишається необхідність класифікації невизначеностей різних типів і зумовлених ними пропусків, утрат і неточних значень. Для кожної предметної галузі, виходячи з особливостей даних, з якими доводиться працювати, час від часу фіксуються одноманітні помилки, похибки, пропуски, а тому через певний час можна вибрати ефективні алгоритми опрацювання таких невизначеностей та пов'язаних з ними пропущених значень, характерних саме для цієї галузі. Обрані методи можуть бути використані для оброблення даних в інших галузях і навіть бути високоефективними у задачах іншої специфіки. Основною метою роботи аналітиків є саме виявлення і напрацювання таких рекомендацій для конкретних сфер застосування, які можуть бути типовими для розв'язання різноманітних фінансово-економічних завдань, задач логістики, прогнозування продажів, маркетингових досліджень тощо.

Поетапне розв'язання задачі заповнення пропущених даних передбачає аналіз суті процесу, що описується певною послідовністю даних, підбір структури моделі заповнення пропусків, вибір адекватних методів інтелектуального аналізу даних для заповнення пропущених даних, реалізація цих методів сучасними інструментальними засобами.

Перспективою для подальших досліджень автори вважають виконання на початковому етапі ґрунтовного аналізу причин появи пропусків та екстремальних значень. Доцільно також застосовувати комбінації методів різних типів — імовірнісних, статистичних та інтелектуального аналізу даних з метою збереження особливостей вхідної вибірки.

ЛІТЕРАТУРА

1. *Згуровский М.З.* Системный анализ: Проблемы. Методология. Приложения. / М.З. Згуровский, Н.Д. Панкратова; НАН Украины. Ин-т приклад. систем. анализа. — К.: Наук. думка, 2005. — 743 с.
2. *Згуровський М.З.* Основи системного аналізу: підруч. для студ. вищ. навч. закл. / М.З. Згуровський, Н.Д. Панкратова. — К.: Вид. група ВНУ, 2007. — 543 с.
3. *Панкратова Н.Д.* Рациональный компромисс в системной задаче концептуальной неопределенности / Н.Д. Панкратова // Кибернетика и системный анализ. — 2002. — № 4. — С. 162–180.
4. *Кузнєцова Н.В.* Практичні підходи до визначення та урахування невизначеностей, що формують фінансові ризики / Н.В. Кузнєцова // Тр. Одес. политехн. ун-та. — Одесса, 2014. — Вып. 2(44). — С. 160–170.
5. *Вікіпедія* [Електронний ресурс]. — Режим доступу: https://uk.wikipedia.org.
6. *Зангиева И.К.* Решение проблемы неповноты данных массовых опросов / Российская социология завтрашнего дня: сб. студ. работ / И.К. Зангиева. — М.: Изд. дом ГУ-ВШЭ, 2008. — Вып. 3. — С. 84–95.
7. *Снитюк В.Е.* Эволюционный метод восстановления пропусков в данных / В.Е. Снитюк // Интеллектуальный анализ информации. — К., 2006. — С. 262–271.
8. *Злоба Е.* Статистические методы восстановления пропущенных данных / Е. Злоба, И. Яцкив // Computer Modelling & New Technologies. — 2002. — **6**, № 1. — P. 51–61.
9. *Загоруйко Н.Г.* Методы распознавания и их применение / Н.Г. Загоруйко. — М.: Сов. радио, 1972. — 216 с.
10. *Rubin D.B*. An Overview of Multiple Imputation / D.B. Rubin // Proc. Survey Research Methods Section of the American Statistical Association. — 1988. — P. 79–84.
11. *Dempster A.P*. Likelihood from Incomplete Data via the EM Algorithm / A.P. Dempster, N.M. Laird, D.B. Rubin // Journal of the Royal Statistical Society. Series B (Methodological). — 1977. — **39**, N 1. — P. 1–38.
12. *Бідюк П.І.* Моделі і методи прикладної статистики / П.І. Бідюк, Л.О. Коршевнюк, Н.В. Кузнєцова . — К.: НУТУ «КПІ», 2014. — 722 с.
13. *Shi F*. Missing Value Estimation for Microarray Data by Bayesian Principal Component Analysis and Iterative Local Least Squares / F. Shi, D. Zhang, J. Chen, H.R. Karimi // Mathematical Problems in Engineering. Article ID 162938. — 2013. — P. 17.
14. *Siddiqi N.* Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring / N. Siddiqi. — 2005. — 196 p.
15. *Owen M.* Tukey's Biweight Correlation and the Breakdown [Електронний ресурс] / M. Owen. — 2005. — Режим доступу: http://pages.pomona.edu/~jsh04747/Student%20Theses/MaryOwen10.pdf
16. *Breheny P*. Robust regression [Електронний ресурс] / P. Breheny. — Режим доступу: http://web.as.uky.edu/statistics/users/pbreheny/764-F11/notes/12-1.pdf.

Надійшла 18.06.2015