Технічні науки

УДК 004.9

Маркін Іван Дмитрович

студент

Інституту прикладного системного аналізу

Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського»

Маркин Иван Дмитриевич

студент

Института прикладного системного анализа

Национального технического университета Украины

«Киевский политехнический институт имени Игоря Сикорского»

Markin Ivan

Student of the

Institute of applied systems analysis of the

National technical university of Ukraine

"Ihor Sikorskiy Kyiv Politechnical Institute"

Науковий керівник:

Кухарєв Сергій Олександрович

асистент

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

ПОРІВНЯННЯ МЕТОДІВ ЗАПОВНЕННЯ ПРОПУСКІВ ДАНИХ
СРАВНЕНИЕ МЕТОДОВ ЗАПОЛНЕНИЕ ПРОПУСКОВ ДАННЫХ
COMPARISON OF FILLING MISSING VALUES METHODS

Анотація. *Висвітлено застосування та порівняння існуючих методів заповнення пропусків даних.*

Ключові слова: *пропуски даних, ЕМ-алгоритми, регресія.*

Аннотация. *Освещены применения и сравнение существующих методов заполнение пропусков данных.*

Ключевые слова: *пропуски данных, EM-алгоритм, регрессия.*

Summary. *The application and comparison of existing methods of filling missing data.*

Key words: *missing data, EM-algorithm, regression.*

Вступ. У практичних завданнях аналізу даних вибірки часто містять в собі пропущені значення. Причини можуть бути різними, наприклад, відсутність відповіді респондента на конкретне запитання анкети, відмова датчика для вимірювань показника, помилки в програмному забезпеченні під час запису даних. Часто викиди даних також можна розглядати як пропуски. До викидів можна віднести данні, які явно суперечать даним з усієї вибірки.

За рідкісним винятком алгоритми машинного навчання не працюють з вибірками, що мають пропущені значення. Тому виникає необхідність у процедурі заповнення даних – процедурі попередньої обробки. Існують різні підходи до вирішення даного завдання, які різняться за своєю природою, областю застосування і обчислювальною складністю.

Невдалий вибір методу заповнення пропусків може не тільки не поліпшити, а й сильно погіршити результати. У даній статті розглянуті існуючі методи обробки пропусків, які отримали широке застосування на практиці, їх переваги та недоліки.

Перш ніж перейти до розв'язування задачі заповнення пропусків даних необхідно виявити механізм формування пропусків. Розрізняють три основних механізми формування пропусків в даних: MCAR, MAR, MNAR. Далі розглянуто кожен з цих трьох механізмів.

MCAR (Missing Completely At Random) – механізм рівномірного формування пропусків, тобто ймовірність пропуску для кожного запису однакова. Прикладом MCAR є випадкова вибірка групи населення, де кожен член має однаковий шанс потрапити у вибірку. Члени популяції які не брали участі в опитування і є MCAR.

MAR (Missing At Random) – ймовірність пропуску може бути обрахована в залежності від іншої наявної в даних інформації. На практиці дані зазвичай пропущені не випадково, для них існує певна закономірність. Люди, які займають керівні посади і/або які отримали вищу освіту частіше, ніж інші респонденти, не відповідають на питання про свої доходи. Оскільки посада і освіта сильно корелюють з доходами, то в такому випадку пропуски в графі доходи вже не можна вважати абсолютно випадковими, тобто говорити про випадок MCAR не представляється можливим. Важливо зазначити що механізм MAR частіше зустрічається на практиці ніж MCAR.

MNAR (Missing Not At Random) - механізм формування пропусків, при якому дані відсутні в залежності від невідомих чинників. MNAR передбачає, що ймовірність пропуску могла б бути описана на основі інших атрибутів, але інформація по цим атрибутам в наборі даних відсутня. Як наслідок, ймовірність пропуску неможливо виразити на основі інформації, що міститься в наборі даних. Відомий приклад MNAR з області медичних досліджень полягає в тому що респондент з більшою ймовірністю не братиме участь в опитувані, якщо лікування призводить до дискомфорту. Такі пропуски даних не є випадковими, а тому мають бути змодельовані, інакше дослідник повинен прийняти деякі упередженості в своїх висновках.

Різноманітні ситуації та причини виникнення пропусків в даних призвело до появи багатьох досліджень в даній області. Велика кількість методів розв'язування задачі заповнення пропусків даних вимагає

систематизації підходів та класифікації методів [1]. У вказаній вище праці приведені основні принципи методів відновлення даних. Таким чином більшість розроблених методів підпадають під наведену схему класифікації (рис.1).

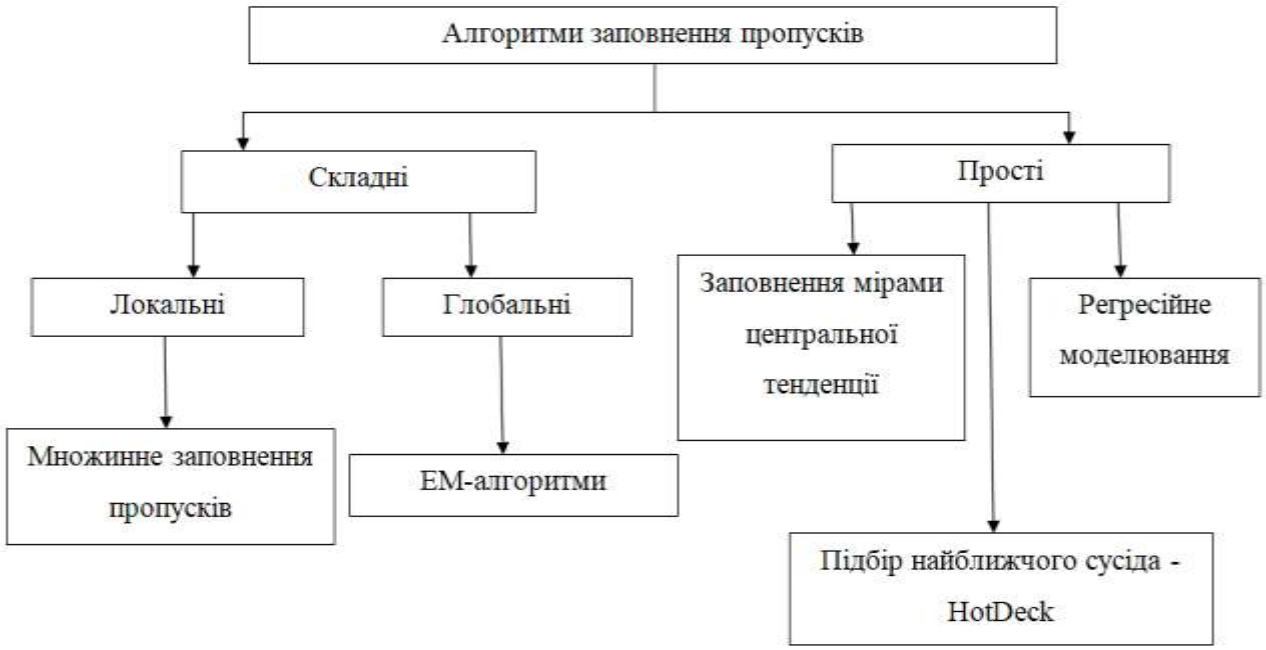

Рис. 1. Класифікація алгоритмів заповнення даних

Прості алгоритми – засновані на простих арифметичних операціях, відстані між об'єктами, регресійним моделюванням.

Складні алгоритми – ітеративні алгоритми. Даний тип алгоритмів передбачає оптимізацію деякого функціонала, який в свою чергу обчислює точність підставлених значень. Складні алгоритми можна розділити на глобальні та локальні.

Глобальні алгоритми при передбачені кожного пропущеного значення використовують усі об'єкти вибірки.

Локальні алгоритми при передбачені кожного пропущеного значення використовують об'єкти які знаходяться в певному околі передбачуваного значення.

Можна стверджувати, що теорія відновлення пропусків даних постійно розвивається, з'являються нові алгоритми та удосконалюються

існуючі. Це пов'язано з тим що не існує алгоритму який був би прийнятний та давав кращі результати в абсолютній більшості випадків.

Заповнення мірами центральної тенденції (середнім по всій вибірці або середніми по групах) - застосування має сенс тільки в разі проходження даних умові MAR, дану групу методів легко можна реалізувати; недоліки - спотворення розподілу даних, зменшення дисперсії.

Заповнення по регресії. В основу даної групи методів покладені добре відомі алгоритми регресійного аналізу [2]. З умов застосування даного методу можна виділити вимогу про приналежність даних умові MAR (хоча для окремих випадків можливе застосування більш слабких вимог) і вимоги, які стосуються виконання передумов регресійного аналізу. Недоліки подібних методів очевидні: якість передбачення (відновлення пропусків) безпосередньо залежить від успішного вибору взятої за основу регресійній моделі.

Метод заміни пропущеного значення середнім з найближчих присутніх елементів змінної. Даний метод є ефективним розвитком методу заміни пропусків загальним середнім, і експерименти показують гарну точність методу в разі одиночних пропусків на досить гладких рядах даних. Завдяки простоті реалізації можна навіть рекомендувати використання даного методу в наведених вище умовах, але тільки в них. Наявність в даних групових пропусків або сильні флуктуації значень ряду зводять ефективність методу до нуля.

EM-алгоритм - відноситься до категорії методів моделювання [3]. Особливість цих методів - побудова моделі породження пропусків з подальшим отриманням висновків на підставі функції правдоподібності, побудованої за умови справедливості даної моделі, з оцінюванням параметрів методами типу максимальної правдоподібності. Відзначимо, що якщо інші методи відновлення пропусків вимагають, щоб дані

відповідали умові MAR (або MCAR як жорсткішого), то для даних методів можлива побудова моделей, що враховують конкретну специфіку області, як наслідок, можлива постановка слабших умов до даних. Недолік - необхідність побудови моделі породження пропусків.

Вибір методу заповнення пропусків може залежати від типів ознак, в яких існують пропуски, від кількості об'єктів, що мають пропущені значення, і від причини їх виникнення. У кожному завданні необхідний індивідуальний підбір методу обробки пропущених значень.

Прості методи заповнення пропусків (заповнення модою, середньою та спеціальною значенням) показують якість порівнянну з просунутими методами, тому застосування витратних за часом роботи методів може бути необґрунтованим в рішенні задач аналізу даних.

Література

1. Kalton G., Kasprzyk D. The Treatment of Missing Survey Data: Survey Methodology. 1968. V. 12. P. 1-16.
2. Двоенко С.Д. Неиерархический дивизимный алгоритм кластеризации. Автоматика и телемеханика. 1999. № 4. С. 117-124.
3. Загоруйко Н.Г., Йолкіна В.Н., Алгоритм заполнения пропусков в эмпирических таблицах (алгоритм Zet). Эмпирическое предсказание и распознавание образов. Новосибірськ, 1975. Вид. 61. Вычислительные системы. С. 3-27.