УДК 004.622
DOI: 10.31891/2219-9365-2020-66-2-15

КОМАР М. П.
Західноукраїнський національний університет

МЕТОДИ ВІДНОВЛЕННЯ ВІДСУТНІХ ДАНИХ У ІНТЕРФЕЙСІ ВЕЛИКИХ ДАНИХ

У статті досліджено відомі методи та запропоновано новий метод відновлення відсутніх даних.

Найбільшою проблемою відомих методів відновлення відсутніх даних є лише обробка структурованих даних, тоді, коли напівструктуровані та неструктуровані дані домінують у реальних прикладних задачах сучасного світу. Крім того, дуже важливо аналізувати приховані залежності в наборі даних, а також враховувати характер набору даних та передбачати відсутність даних для кожного джерела даних окремо. Тому, перспективним є використання парадигми великих даних для попередньої обробки та обробки інформації з різних джерел, що складається з безперервних числових даних та категоріальних даних. Тим не менше, природа відсутніх даних у різних джерелах даних теж різна. Отже, основною ідеєю є аналіз даних з різних джерел на основі специфіки та природи цих джерел. Ідея запропонованого методу обчислення відсутніх даних полягає в обробці структурованих та напівструктурованих даних на основі ієрархії об'єктів, а також набору функціональних залежностей та розробки правил асоціації. Це питання дуже важливе для інтерфейсів великих даних, оскільки більша частина інформації доступна в напівструктурованому вигляді. Запропонований метод створює додаткові значення даних за допомогою доменних та функціональних залежностей на основі декількох методів обчислення та додає ці значення до наявних навчальних даних.

Ключові слова: великі дані, аналіз даних, обробка даних, відсутні дані, відновлення даних, функціональні залежності, правила асоціації, кластерний аналіз, машинне навчання, нейронна мережа.

KOMAR M.
West Ukrainian National University, Ternopil, Ukraine

METODS OF MISSING DATA IMPUTATION IN BIG DATA INTERFACE

Known methods are investigated and a new method of missing data imputation is proposed.

The biggest problem of the known methods of missing data imputation is only the processing of structured data, when semi-structured and unstructured data dominate the real applications of the modern world. In addition, it is very important to analyze the hidden dependencies in the data set, as well as to take into account the nature of the data set and to assume the absence of data for each data source separately. Therefore, it is promising to use the big data paradigm for pre-processing and processing information from different sources, consisting of continuous numerical data and categorical data. However, the nature of the missing data in different data sources is also different. Thus, the main idea is to analyze data from different sources based on the specifics and nature of these sources.

The idea of the proposed method of calculating missing data is to process structured and semi-structured data based on a hierarchy of objects, as well as a set of functional dependencies and the development of association rules. This issue is very important for big data interfaces, as most of the information is available in a semi-structured form. The proposed method creates additional data values using domain and functional dependencies based on several calculation methods and adds these values to the existing training data.

Keywords: big data, data analysis, data processing, missing data, data recovery, functional dependencies, association rules, cluster analysis, machine learning, neural network.

Вступ. Постановка проблеми. Процеси прийняття рішень в різних галузях (науці, бізнесі, економіці) сильно залежать від наявності даних, з яких можна отримати корисну інформацію. В процесів прийняття рішень можна застосовувати прогнозні моделі, які використовують отримані дані в якості вхідних даних. Такі моделі не працюють, коли один або кілька входів відсутні. У багатьох рішеннях просто ігнорувати пропущені дані не можна. Це пов'язано головним чином з тим, що неповні дані можуть призвести до необ'єктивних результатів статистичного моделювання або навіть збитків при керуванні технологічними процесами. З цієї причини дуже важливо приймати рішення на основі повних даних. Оскільки, багато методів машинного навчання, таких як нейронні мережі, метод опорних векторів та багато інших методів обчислювального інтелекту не можуть бути використані для прийняття рішень, якщо дані не повні. У таких випадках оптимальний результат прийняття рішення все одно повинен зберігатися, незважаючи на відсутні дані. У випадках неповних векторів даних першим кроком до прийняття рішення є оцінка відсутніх значень. Після оцінки відсутніх значень можна використовувати інструменти відновлення, заповнення даних. Імпутація даних стає все більш важливою в таких сферах, де кількість зразків, як правило, невелика, а вартість отримання нових висока, що робить ігнорування неповних даних не правильним. Таким чином, створення теоретичних основ та методів, які можуть призвести до повноти даних, є важливою задачею. Метою роботи є дослідження та розробка методів відновлення відсутніх даних в інтерфейсах великих даних.

1. Механізм відсутніх даних. За механізмом відсутності дані визначають наступним чином [1–4]:

1. Дані відсутні абсолютно випадково (missing completely at random – MCAR) – відсутність значень в даних не залежить від будь-яких значень – наявних або відсутніх.

2. Дані відсутні випадково (missing at random – MAR) – відсутні значення умовно залежні від наявних значень, а не від відсутніх. Ймовірність того, що значення X_i опущено, не пов'язана з самим X_i, але вона залежить від інших змінних в аналізованій таблиці.

3. Дані відсутні не випадково (missing not at random – MNAR) – відсутність значень залежить від значення відсутньої змінної. Ймовірність того, що значення X_i опущено, пов'язана з самим X_i.

Загальна проблема MCAR полягає в тому, що дані за цією схемою на початку аналізу насправді можуть бути різними через невідомі або неприйнятні фактори. На цій підставі неповні дані часто мають схему, що не відповідає MCAR, але вони можуть бути віднесені до схеми з випадковими пропусками MAR у випадку, якщо неповні дані можуть мати місце залежно від відомих показників. Якщо неповні дані неможливо віднести ні до MCAR, ні до MAR, вони класифікуються як MNAR. Це означає, що результати моделювання на основі таких даних матимуть упереджені оцінки, якщо модель відсутнього значення невідома або не враховується.

Основним недоліком MCAR, MAR та MNAR є відсутність зв’язку з джерелом даних або структурою даних. Це означає неможливість передбачити місце пропущених даних, тоді як такий тип передбачення дуже важливий для потокових даних.

2. Методи та алгоритми відновлення відсутніх даних. В якості фундаментальної технології аналізу великих даних використовується кластеризація, яка поділяє об'єкти на різні кластери на основі подібності [5–7]. Традиційні алгоритми кластеризації даних зосереджені на повній обробці даних, таких як кластеризація зображень [8], кластеризація звуку [9] та кластеризація тексту [10]. Гетерогенні методи кластеризації даних, останнім часом, цікавлять науковців [11–13].

Крім того, було запропоновано багато алгоритмів – наприклад, в [14] оптимізовано уніфіковану цільову функцію за допомогою ітераційного процесу, і розроблено алгоритм спектральної кластеризації для кластеризації різнорідних даних на основі теорії графів. У [15] запропоновано алгоритм нечітких c-середніх значень високого порядку для розширення звичайного алгоритму нечітких c-середніх з векторного простору на тензорний простір. Для кластеризації даних в системах Інтернету речей (IoT) запропоновано алгоритм c-середніх значень високого порядку, заснований на тензорах [16]. Ці алгоритми ефективні для покращення продуктивності кластеризації неоднорідних даних. Однак вони можуть отримати лише результати кластеризації та позбавлені подальшого аналізу неповних даних з низькими розмірами. Тому їх продуктивність обмежена неоднорідними даними у середовищі великих даних. Що ще важливіше, інші існуючі алгоритми кластеризації об’єктів не враховують відновлення даних та відсутність даних.

Метод середнього заміщення (MS) дозволяє вирішити проблему неповноти даних, замінюючи кожну відсутність середньою змінною. Типи MS – це підміна медіани, моди, середнього значення підгрупи даних [17], значення з найвищою частотою (найбільш загальне значення), в деяких випадках замінює мінімальне/максимальне значення. Цей метод може призвести до багатьох небажаних результатів, таких як заниження реальної дисперсії, негативне упередження кореляцій та неправильне представлення загальної сукупності [18].

Проста hot-deck імпутація замінює кожне відсутнє значення випадковим значенням, взятим із існуючого набору даних [19]. Його суттєвим недоліком є спотворення кореляцій.

Імпутація Cold-deck (CD) реалізує заміну кожного проходу на деяке постійне значення із зовнішнього джерела [20]. Особливим випадком CD є нульова підміна. Як і послідовна та випадкова гаряча заміна, вона має ті ж недоліки, що і проста гаряча заміна.

Перевага підходів на основі заміщення полягає в тому, що вони дають внутрішньо узгоджений набір значень, але не можуть визначити взаємозв'язки між змінними. Методи заповнення прогалин на основі моделі оцінюють значення параметрів моделі, які послідовно використовуються для обчислення прогалин. Ці методи враховують взаємозв'язки між змінними, але вони досить складні, оскільки спричиняють помилки, коли всі параметри оцінюються одночасно.

Оцінка регресії (RE) передбачає заміщення прогалин даних прогнозованими значеннями, отриманими з рівняння регресії, побудованого з повного набору даних. Недоліки заповнення регресією включають наступне: необхідність точно ідентифікувати моделі регресії, перебільшення кореляції та коваріації, ймовірність виходу прогнозованих значень за межі логічного ряду та необхідність великих обсягів даних для отримання послідовних оцінок.

Метод максимальної ймовірності (ML) виводить оцінені параметри таким чином, що ймовірність відтворення даних на основі відомих значень максимізується. Метод ML розглядається як один із підходів до обробки неповних даних [21], і його широко рекомендують використовувати, оскільки він не призводить до упередженості при наявності відсутніх схем значень MCAR та MAR [22]. Однак метод ML не вирішує цього питання за наявності MNAR, хоча величина такого упередження, як правило, набагато менша, ніж традиційні підходи.

Метод максимізації очікувань (EM) заснований на загальному ітераційному алгоритмі, який реалізує послідовне заповнення відсутніх значень за їхніми оцінками та максимізує умовне сподівання

ймовірності отримання повних даних до процесу конвергенції [22]. Основним недоліком методів ML та EM є їх ітераційна властивість і висока часова складність в подальшому, особливо для обробки великих даних.

Метод k-найближчого сусіда (kNN) заснований на визначенні найближчого сусіда для кожного пропущеного значення за допомогою певної метрики [23]. Після пошуку k-найближчих сусідів за кожний прохід воно замінюється середнім значенням характеристик для сусідів, що містять повні дані. Окремим випадком методу kNN є зважений метод kNN [24]. Найбільша складність у його застосуванні полягає у визначенні адекватного ступеня близькості.

Алгоритм опорних векторних машин (SVM) [25] дозволяє уникнути оцінки ймовірності даних, які є стабільними. Вибір ядра впливає на якість обчислення відсутніх даних. Ось чому реалізація цього алгоритму залежить від набору даних та його параметрів.

Метод випадкових лісів (RF) [26] заснований на побудові класифікатора, сформованого групою дерев рішень, навчальним набором повних даних та прогнозуванням відсутніх значень у наборі тестів. На першому кроці всі прогалини в наборі тестів заповнюються середнім значенням на основі наявних даних. Далі розраховується матриця близькості для першого набору даних. Середні ваги, розраховані на матриці близькості для першої ітерації, використовуються для заміщення пропусків на наступній ітерації і т. д. до досягнення критерію зупинки [27]. Хоча цей метод є ітеративним, він дозволяє розпаралелювати обробку за рахунок зменшення часової складності.

Метод асоціативних правил (AR) передбачає побудову асоціативних правил щодо надмірності наявних даних [28, 29]. По-перше, усі отримані правила сортуються за якимись параметрами (наприклад, автентичністю). Потім здійснюється пошук набору даних за кожним пропуском та відповідним правилом, яке не суперечить іншим значенням, що знаходяться у відновлюваному рядку. Алгоритм AR має кілька модифікацій. Наприклад, один з них, FP-алгоритм може ефективно паралельно працювати, саме тому його можна використовувати для попередньої обробки великих даних. Однак часову складність цього методу доводиться вдосконалювати.

Метод багатошарового персептрону (MLP) використовує навчену мережу для оцінки відсутніх значень. Оскільки вхідними даними є непусті набори даних, виходами є неповні значення змінних, які потрібно визначити [22, 30, 31]. Заповнення відсутніх значень за допомогою MLP складається з трьох етапів:

– формування повного (навчального) та неповного (тестового) набору;

– побудова MLP на навчальному наборі, де значення змінної є значеннями прогалин, а вхідними значеннями, відповідно, є заповнені значення змінної;

– прогнозування невідомих значень для кожної неповної структури даних за допомогою навченої мережі.

Недоліком цього підходу є прив'язка до набору змінних, що містять відсутні значення, тому потрібно будувати окрему модель MLP для кожної комбінації.

Метод самоорганізованої карти (SOM) [32] дозволяє навчати мережу на основі неповного набору даних. Він запускається завдяки здатності алгоритму ігнорувати відсутні значення та обчислювати відстань між поточним спостереженням і вузлом та використовувати навчену карту для оцінки значень. Недолік SOM такий же, як і MLP.

Глибоке навчання (DL) активно застосовується у багатьох додатках завдяки його потужній здатності до імпутації даних [33]. Глибоко вбудована кластеризація (DEC) вчиться перетворювати з простору даних у простір об’єктів низьких розмірів [34]. У [35] показано здатність представлення ознак автокодера (VAE). VAE вивчає багатогранну структуру даних і досягає високих показників кластеризації [36]. Крім того, VAE має потужну здатність до вилучення та реконструкції функцій, і це може бути хорошим інструментом для обробки неповних даних.

У роботі [37] запропоновано алгоритм VAE, який може покращити ефективність кластеризації мультимедійних даних. На відміну від багатьох існуючих технологій, цей алгоритм вивчає функції даних, проектуючи вдосконалену мережу VAE, і використовує fuzzy c-means алгоритм на основі тензора для кластеризації даних у просторі функцій.

У роботі [38] пропонується розширення варіаційного автокодера (VAE), який називається варіаційним автокодером із зваженими втратами (VAE-WL). Він використовує спеціальну функцію втрат, яка більше підходить для обчислення відсутніх значень.

Найбільшою проблемою методів, проаналізованих вище, є обробка структурованих даних лише тоді, коли напівструктуровані та неструктуровані дані домінують у реальних ситуаціях [39]. Крім того, дуже важливо проаналізувати приховані залежності в наборі даних, а також врахувати характер набору даних та передбачити відсутність даних для кожного джерела даних окремо. Ось чому парадигма великих даних використовується для попередньої обробки та обробки інформації з різних джерел, що складається з безперервних числових даних та категоріальних даних. Тим не менше, природа відсутніх даних у різних джерелах даних теж різна. Отже, основною ідеєю є аналіз даних з різних джерел на основі специфіки та природи цих джерел.

3. Запропонований метод відновлення відсутніх даних. Ми пропонуємо проаналізувати структуру наборів даних, щоб знайти типи невизначеності та створити функцію ідентифікації (посередника) для загальної схеми великих даних на основі всіх джерел даних [39]. Після цього загальну схему можна використовувати для відновлення відсутніх даних безпосередньо у джерелі даних.

Запропонований метод відновлення відсутніх даних базується на функціональних залежностях та правилах асоціації та складається з двох частин:

– видобування ймовірнісних виробничих залежностей на основі інтерфейсу великих даних;

– використання ймовірнісних виробничих залежностей для відновлення відсутніх даних.

Видобування ймовірнісних виробничих залежностей.

Аналіз великих обсягів даних вимагає виявлення груп атрибутів, які утворюють функціональні залежності. Однак, е реальному світі набори даних набагато більш поширені, а важливі залежності визначені тільки на підмножині значень групи ключових атрибутів. Причому залежності можуть з'являтися не тільки для кортежів в реляційних джерелах даних, але і між різними частинами різних кортежів. Будемо називати їх імовірнісною виробничою залежністю (PPD).

Імовірнісна виробнича залежність – це виробниче правило у виборі базового співвідношення, яке діє для значного числа об'єктів в цьому виборі. Поріг значущості повинен визначатися фахівцями або на основі розрахунків ймовірності помилкового вибору цієї залежності. Основна відмінність між асоціативними правилами і PPD полягає в тому, що PPD буде згенеровано з існуючої функціональної залежності в наборі даних.

$$F_1 : K = \{a_i\}, a_i \in A, D = \{a_i\},$$
$$a_i \in A : P(k \in K \to d \in D) = p, \quad (1)$$

де *k* та *d* – набори груп атрибутів *K* і *D*, відповідно

Основним показником надійності такої залежності є відношення кількості об'єктів, які є в такій PPD, до кількості об'єктів у вибірці:

$$P(F_1) = \frac{|\sigma_{k \in K \wedge d \in D}(R)|}{|\sigma_{k \in K}(R)|}. \quad (2)$$

Правило класифікації називається ймовірнісним продуктивним зв'язком між підмножинами атрибутів *X* і *Y* в консолідованих великих даних *Bd*, яка зустрічається в навчальному наборі *bd* з рівнем довіри *s*, де:

$$(\mathrm{X} = \mathrm{x}) \to (Y - y) \quad (3)$$

Правило класифікації побудовано на навчальному наборі даних зі схемою *Bd* з відомими значеннями міток класів. Це правило створено для схеми *Bd*, тому на нього не вплине нова сутність, яка надходить у вибірку великих даних (перевірка незалежності набору даних).

Мітка класу називається лінгвістичною змінною або звичайною характеристикою об'єктів, які є значеннями підмножини атрибутів *Y* і позначають об'єкти з загальними (наприклад, ступенем довіри *s*) значеннями підмножини атрибутів *X*. Домени атрибутів, що належать підмножині атрибутів *Y*, $y \in dom(Y) = \pi_y(Bd)$ повинні містити кінцевий і відомий набір значень.

Мітки класів вибираються з відомого набору значень (в межах області дослідження є фіксованими), а клас об'єктів, який тільки що був введений в консолідоване сховище даних, заснований на правилах класифікації [29]. Теги будуть додаватися автоматично, так як нові джерела даних також додаються в каталог великих даних. Використання ймовірнісних виробничих залежностей для відновлення відсутніх даних. Щоб класифікувати об’єкти (або заповнити відсутні дані), потрібно побудувати правила класифікації. Як правило, великі дані можуть зберігати інформацію про декілька типів класів, і для кожного класу існує підмножина функцій. Цю саму функцію можна використовувати для визначення декількох типів класів.

Правилами класифікації називають PPD, які виконуються для певної підмножини кортежів для консолідованих великих даних *bd*. Розроблено наступний алгоритм імпутації даних.

Для аналізу важлива також послідовність подій, які часто відбуваються. Якщо в таких послідовностях виявляються закономірності, можна з деякою мірою ймовірності передбачити виникнення подій в майбутньому.

Висновки. Запропоновано метод відновлення відсутніх даних у наборах великих даних, який на відміну від існуючих методів, створює додаткові значення даних на основі функціональних залежностей та правил асоціації та додає ці значення до наявних навчальних даних, що, в свою чергу, дало змогу підвищити ефективність подальшого аналізу даних.

Даний метод полягає в обробці структурованих та напівструктурованих даних на основі ієрархії об’єктів, а також набору функціональних залежностей та розробки правил асоціації. Це питання дуже важливе для інтерфейсів великих даних, оскільки більша частина інформації доступна в напівструктурованому вигляді. Запропонований метод створює додаткові значення даних за допомогою доменних та

функціональних залежностей на основі декількох методів обчислення та додає ці значення до наявних навчальних даних. Правильність обчислених значень перевіряється на класифікаторі, побудованому на вихідному наборі даних.

Література

1. He Y. Missing data analysis using multiple imputation. Circulation: Cardiovascular Quality and Outcomes. 2010. 3(1). 98-105.
2. Little R. J. A., Rubin D. B. Statistical analysis with missing data. John Wiley & Sons, 2019. 464 p.
3. Ahmat Zainuri, N., Jemain, A. A., Muda, N. A Comparison of Various Imputation Methods for Missing Values in Air Quality Data. JSM, 2015, 44, 449-456.
4. Leke, C. A., Marwala, T. Introduction to Missing Data Estimation. Deep Learning and Missing Data in Engineering Systems. 2019, 1-20. https://doi. org/10.1117/12.2053057 [Access 18.07.2020].
5. Zhang, S.; Yang, Z.; Xing, X.; Gao, Y.; Xie, D.; Wong, H.S. Generalized Pair-Counting Similarity Measures for Clustering and Cluster Ensembles. IEEE Access. 2017, 5, 16904–16918.
6. Hoecker, M.; Polsterer, K.L.; Kugler, S.D.; Heuveline, V. Clustering of Complex Data-Sets Using Fractal Similarity Measures and Uncertainties. In Proceedings of the 2015 IEEE 18th International Conference on Computational Science and Engineering, Porto, Portugal, 21–23 October 2015; pp. 82–91.
7. Zhou, L.;Wu, D.; Zheng, B.; Guizani, M. Joint physical-application layer security for wireless multimedia delivery. IEEE Commun. Mag. 2014, 52, 66–72.
8. Li, F.; Qiao, H.; Zhang, B. Discriminatively boosted image clustering with fully convolutional auto-encoders. Pattern Recognit. 2018, 83, 161–173.
9. Gebru, I.D.; Alameda-Pineda, X.; Forbes, F.; Horaud, R. EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 38, 2402–2415.
10. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A. Unsupervised feature selection technique based on genetic algorithm for improving the Text Clustering. In Proceedings of the 2016 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 13–14 July 2016; pp. 1–6.
11. Saadaoui, F.; Bertrand, P.R.; Boudet, G.; Rouffiac, K.; Dutheil, F.; Chamoux, A. A Dimensionally Reduced Clustering Methodology for Heterogeneous Occupational Medicine Data Mining. IEEE Trans. Nanobiosci. 2015, 14, 707–715.
12. Zhou, Q. Research on heterogeneous data integration model of group enterprise based on cluster computing. Clust. Comput. 2016, 19, 1275–1282.
13. Ramachandran, N.; Perumal, V. Delay-aware heterogeneous cluster-based data acquisition in Internet of Things. Comput. Electr. Eng. 2018, 65, 44–58.
14. Meng, L.; Tan, A.H.; Xu, D. Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering. IEEE Trans. Knowl. Data Eng. 2014, 26, 2293–2306.
15. Li, P.; Chen, Z.; Yang, L.T.; Zhao, L.; Zhang, Q. A privacy-preserving high-order neuro-fuzzy C-means algorithm with cloud computing. Neurocomputing 2017, 256, 82–89.
16. Zhang, Q.; Yang, L.T.; Chen, Z.; Li, P. High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT. Inf. Fusion 2018, 39, 72–80.
17. Kowarik, A., Templ, M. Imputation with the R Package. VIM, 2016. https://doi.org/10.18637/jss.v074.i07 [Access 18.07.2020].
18. Lin, T. H. A Comparison of Multiple Imputation with EM Algorithm and MCMC Method for Quality of Life Missing Data. Quality and Quantity, 2010, 44, 277-287.
19. Myers, T. A. Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data. Communication Methods and Measures, 2011, 5, 297-310.
20. Shakhovska, N., Strubytskyi, R. Model of Data Warehouse with Uncertain Consolidated Data. Applied Mathematics & Information Sciences, 2015, 9(4). http://dx.doi.org/10.12785/amis/090412 [Access 18.07.2020].
21. Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., Petersen, I. Missing Data and Multiple Imputation in Clinical Epidemiological Research. Clinical Epidemiology, 2017, 9, 157-166.
22. Nelwamondo, F. V., Mohamed, S., Marwala, T. Missing Data: A Comparison of Neural Network and Expectation Maximisation Techniques, 2007, arXiv:0704.3474 [Access 18.07.2020].
23. Lim, S. Y., Mohamad, M. S., Chai, L. E., Deris, S., Chan, W. H., Omatu, S., Ibrahim, Z. Investigation of the Effects of Imputation Methods for Gene Regulatory Networks Modelling Using Dynamic Bayesian Networks. Distributed Computing and Artificial Intelligence, 13th International Conference, 2016, 413-421.
24. Aydilek, I. B., Arslan, A. A Hybrid Method for Imputation of Missing Values Using Optimized Fuzzy C-Means with Support Vector Regression and a Genetic Algorithm. Information Sciences, 2013, 233, 25-35.
25. Alhroob, A., Alzyadat, W., Almukahel, I., Altarawneh, H. Missing Data Prediction Using Correlation Genetic Algorithm and SVM Approach. Population, 2020, 11(2). https://doi.org/10.14569/IJACSA.2020.0110288 [Access 18.07.2020].
26. Siroky, D. S. Navigating Random Forests and Related Advances in Algorithmic Modeling. Statistics Surveys, 2009, 3, 147-163.
27. Rahman, Md. G., Islam, M. Z. Missing Value Imputation Using Decision Trees and Decision Forests by Splitting and Merging Records: Two Novel Techniques. Knowledge-Based Systems, 2013, 53, 51-65.
28. Raković, L., Sakal, M., Matković, P., Marić, M. Shadow IT-Systematic Literature Review. Information Technology and Control, 2020, 49(1), 144-160.
29. Shakhovska, N., Kaminskyy, R., Zasoba, E., Tsiutsiura, M. Association Rules Mining in Big Data. International Journal of Computing, 2018, 17(1), 25-32.
30. Jung, S., Moon, J., Park, S., Rho, S., Baik, S. W., Hwang, E. Bagging Ensemble of Multilayer Perceptrons for Missing Electricity Consumption Data Imputation. Sensors, 2020, 20(6). https://doi.org/10.3390/ s20061772 [Access 18.07.2020].
31. Sachenko A., Kochan V., Turchenko V. Instrumentation for Gathering Data. IEEE Instrumentation and Measurement Magazine, 2003, 6 (3), 34-40.
32. Jurgelevičius, A., Sakalauskas, L. Big Data Mining Using Public Distributed Computing. Information Technology and Control, 2018, 47(2), 236-248.
33. Mohammadi, M.; Al-Fuqaha, A.; Sorour, S.; Guizani, M. Deep Learning for IoT Big Data and Streaming Analytics: A Survey. IEEE Commun. Surv. Tutor. 2018, 20, 2923–2960.
34. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. arXiv 2015, arXiv:1511.06335 [Access 18.07.2020].

35. Kingma, D.P.;Welling, M. Auto-Encoding Variational Bayes. arXiv 2013, arXiv:1312.6114 [Access 18.07.2020].
36. Li, X.; Chen, Z.; Poon, L.K.M.; Zhang, N.L. Learning Latent Superstructures in Variational Autoencoders for Deep Multidimensional Clustering. arXiv 2018, arXiv:1803.05206 [Access 18.07.2020].
37. Yu X, Li H, Zhang Z, Gan C. The Optimally Designed Variational Autoencoder Networks for Clustering and Recovery of Incomplete Multimedia Data. *Sensors*. 2019; 19(4):809.
38. Ricardo Cardoso Pereira , Joana Cristo Santos , Jos´e Pereira Amorim , Pedro Pereira Rodrigues and Pedro Henriques Abreu. Missing Image Data Imputation using Variational Autoencoders with Weighted Loss // In Proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2020), 2-4 October 2020. http://www.i6doc.com/en/ [Access 18.07.2020].
39. Shakhovska, N., Vovk, O., Kryvenchuk, Y. Uncertainty Reduction in Big Data Catalogue for Information Product Quality Evaluation. Eastern-European Journal of Enterprise Technologies, 2018, 1, 12-20.

References

1. He Y. Missing data analysis using multiple imputation. Circulation: Cardiovascular Quality and Outcomes. 2010. 3(1). 98-105.
2. Little R. J. A., Rubin D. B. Statistical analysis with missing data. John Wiley & Sons, 2019. 464 p.
3. Ahmat Zainuri, N., Jemain, A. A., Muda, N. A Comparison of Various Imputation Methods for Missing Values in Air Quality Data. JSM, 2015, 44, 449-456.
4. Leke, C. A., Marwala, T. Introduction to Missing Data Estimation. Deep Learning and Missing Data in Engineering Systems. 2019, 1-20. https://doi. org/10.1117/12.2053057 [Access 18.07.2020].
5. Zhang, S.; Yang, Z.; Xing, X.; Gao, Y.; Xie, D.; Wong, H.S. Generalized Pair-Counting Similarity Measures for Clustering and Cluster Ensembles. IEEE Access. 2017, 5, 16904–16918.
6. Hoecker, M.; Polsterer, K.L.; Kugler, S.D.; Heuveline, V. Clustering of Complex Data-Sets Using Fractal Similarity Measures and Uncertainties. In Proceedings of the 2015 IEEE 18th International Conference on Computational Science and Engineering, Porto, Portugal, 21–23 October 2015; pp. 82–91.
7. Zhou, L.;Wu, D.; Zheng, B.; Guizani, M. Joint physical-application layer security for wireless multimedia delivery. IEEE Commun. Mag. 2014, 52, 66–72.
8. Li, F.; Qiao, H.; Zhang, B. Discriminatively boosted image clustering with fully convolutional auto-encoders. Pattern Recognit. 2018, 83, 161–173.
9. Gebru, I.D.; Alameda-Pineda, X.; Forbes, F.; Horaud, R. EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 38, 2402–2415.
10. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A. Unsupervised feature selection technique based on genetic algorithm for improving the Text Clustering. In Proceedings of the 2016 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 13–14 July 2016; pp. 1–6.
11. Saadaoui, F.; Bertrand, P.R.; Boudet, G.; Rouffiac, K.; Dutheil, F.; Chamoux, A. A Dimensionally Reduced Clustering Methodology for Heterogeneous Occupational Medicine Data Mining. IEEE Trans. Nanobiosci. 2015, 14, 707–715.
12. Zhou, Q. Research on heterogeneous data integration model of group enterprise based on cluster computing. Clust. Comput. 2016, 19, 1275–1282.
13. Ramachandran, N.; Perumal, V. Delay-aware heterogeneous cluster-based data acquisition in Internet of Things. Comput. Electr. Eng. 2018, 65, 44–58.
14. Meng, L.; Tan, A.H.; Xu, D. Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering. IEEE Trans. Knowl. Data Eng. 2014, 26, 2293–2306.
15. Li, P.; Chen, Z.; Yang, L.T.; Zhao, L.; Zhang, Q. A privacy-preserving high-order neuro-fuzzy C-means algorithm with cloud computing. Neurocomputing 2017, 256, 82–89.
16. Zhang, Q.; Yang, L.T.; Chen, Z.; Li, P. High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT. Inf. Fusion 2018, 39, 72–80.
17. Kowarik, A., Templ, M. Imputation with the R Package. VIM, 2016. https://doi.org/10.18637/jss.v074.i07 [Access 18.07.2020].
18. Lin, T. H. A Comparison of Multiple Imputation with EM Algorithm and MCMC Method for Quality of Life Missing Data. Quality and Quantity, 2010, 44, 277-287.
19. Myers, T. A. Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data. Communication Methods and Measures, 2011, 5, 297-310.
20. Shakhovska, N., Strubytskyi, R. Model of Data Warehouse with Uncertain Consolidated Data. Applied Mathematics & Information Sciences, 2015, 9(4). http://dx.doi.org/10.12785/amis/090412 [Access 18.07.2020].
21. Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., Petersen, I. Missing Data and Multiple Imputation in Clinical Epidemiological Research. Clinical Epidemiology, 2017, 9, 157-166.
22. Nelwamondo, F. V., Mohamed, S., Marwala, T. Missing Data: A Comparison of Neural Network and Expectation Maximisation Techniques, 2007, arXiv:0704.3474 [Access 18.07.2020].
23. Lim, S. Y., Mohamad, M. S., Chai, L. E., Deris, S., Chan, W. H., Omatu, S., Ibrahim, Z. Investigation of the Effects of Imputation Methods for Gene Regulatory Networks Modelling Using Dynamic Bayesian Networks. Distributed Computing and Artificial Intelligence, 13th International Conference, 2016, 413-421.
24. Aydilek, I. B., Arslan, A. A Hybrid Method for Imputation of Missing Values Using Optimized Fuzzy C-Means with Support Vector Regression and a Genetic Algorithm. Information Sciences, 2013, 233, 25-35.
25. Alhroob, A., Alzyadat, W., Almukahel, I., Altarawneh, H. Missing Data Prediction Using Correlation Genetic Algorithm and SVM Approach. Population, 2020, 11(2). https://doi.org/10.14569/IJACSA.2020.0110288 [Access 18.07.2020].
26. Siroky, D. S. Navigating Random Forests and Related Advances in Algorithmic Modeling. Statistics Surveys, 2009, 3, 147-163.
27. Rahman, Md. G., Islam, M. Z. Missing Value Imputation Using Decision Trees and Decision Forests by Splitting and Merging Records: Two Novel Techniques. Knowledge-Based Systems, 2013, 53, 51-65.
28. Raković, L., Sakal, M., Matković, P., Marić, M. Shadow IT-Systematic Literature Review. Information Technology and Control, 2020, 49(1), 144-160.
29. Shakhovska, N., Kaminskyy, R., Zasoba, E., Tsiutsiura, M. Association Rules Mining in Big Data. International Journal of Computing, 2018, 17(1), 25-32.
30. Jung, S., Moon, J., Park, S., Rho, S., Baik, S. W., Hwang, E. Bagging Ensemble of Multilayer Perceptrons for Missing Electricity Consumption Data Imputation. Sensors, 2020, 20(6). https://doi.org/10.3390/ s20061772 [Access 18.07.2020].
31. Sachenko A., Kochan V., Turchenko V. Instrumentation for Gathering Data. IEEE Instrumentation and Measurement Magazine, 2003, 6 (3), 34-40.

32. Jurgelevičius, A., Sakalauskas, L. Big Data Mining Using Public Distributed Computing. Information Technology and Control, 2018, 47(2), 236-248.

33. Mohammadi, M.; Al-Fuqaha, A.; Sorour, S.; Guizani, M. Deep Learning for IoT Big Data and Streaming Analytics: A Survey. IEEE Commun. Surv. Tutor. 2018, 20, 2923–2960.

34. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. arXiv 2015, arXiv:1511.06335 [Access 18.07.2020].

35. Kingma, D.P.;Welling, M. Auto-Encoding Variational Bayes. arXiv 2013, arXiv:1312.6114 [Access 18.07.2020].

36. Li, X.; Chen, Z.; Poon, L.K.M.; Zhang, N.L. Learning Latent Superstructures in Variational Autoencoders for Deep Multidimensional Clustering. arXiv 2018, arXiv:1803.05206 [Access 18.07.2020].

37. Yu X, Li H, Zhang Z, Gan C. The Optimally Designed Variational Autoencoder Networks for Clustering and Recovery of Incomplete Multimedia Data. *Sensors*. 2019; 19(4):809.

38. Ricardo Cardoso Pereira , Joana Cristo Santos , Jos´e Pereira Amorim , Pedro Pereira Rodrigues and Pedro Henriques Abreu. Missing Image Data Imputation using Variational Autoencoders with Weighted Loss // In Proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2020), 2-4 October 2020. http://www.i6doc.com/en/ [Access 18.07.2020].

39. Shakhovska, N., Vovk, O., Kryvenchuk, Y. Uncertainty Reduction in Big Data Catalogue for Information Product Quality Evaluation. Eastern-European Journal of Enterprise Technologies, 2018, 1, 12-20.

Надійшла / Paper received: 28.09.2020 Надрукована / Paper Printed : 01.12.2020