УДК 004.048

МОДЕЛЬ ОЦІНКИ ЯКОСТІ НОРМАЛІЗАЦІЇ ДАНИХ НА ОСНОВІ ЗАСТОСУВАННЯ КРИТЕРІЇВ ЯКОСТІ КЛАСИФІКАЦІЇ ОБ'ЄКТІВ

Л. М. Ясінська-Дамрі, Б. В. Дурняк

Українська академія друкарства,
вул. Під Голоском, 19, Львів, 79020, Україна

Наведено модель оцінки якості нормалізації даних під час застосування різних типів методів. Як основний критерій оцінки якості відповідного методу було використано точність класифікації, яка виконувалася на нормалізованих даних. Чотири різні типи наборів даних, завантажені з репозиторію машинного навчання (UCI Machine Learning), були застосовані як експериментальні дані. Різні методи нормалізації, доступні з пакета clusterSim мови програмування R, були використані до даних у процесі моделювання. Якість процедури нормалізації у кожному випадку оцінювали на основі результатів класифікації даних за допомогою розрахунку точності розподілу об'єктів по класах. Нейронна мережа «багатошаровий персептрон» була використана як класифікатор на цьому етапі. Результати моделювання засвідчили, що етап нормалізації даних суттєво впливає на точність класифікації, водночас вибір методу нормалізації залежить від типу даних, і його, отже, потрібно проводити у кожному випадку індивідуально.

Ключові слова: *нормалізація даних, критерії якості класифікації, багатошаровий персептрон, передобробка даних, точність класифікації.*

Постановка проблеми. Сучасний стан наукового напряму, орієнтованого на інтелектуальний аналіз даних та машинне навчання, визначає різке збільшення обсягу даних. Водночас виникає потреба щодо зменшення кількості атрибутів або/та об'єктів за допомогою виокремлення найбільш інформативних параметрів для створення ефективних систем прогнозування, класифікації, кластеризації, прийняття рішень тощо [1–4]. Здебільшого експериментальні дані є «необробленими», неповними, мають різні діапазони змін значень атрибутів, і через це їх потрібно трансформувати або корегувати на етапі передобробки. Процедура попередньої обробки даних передбачає обробку відсутніх значень та нормалізацію векторів відповідних атрибутів на основі заздалегідь проведеного статистичного аналізу. У наш час існують різні техніки реалізації цих етапів. Обробка відсутніх значень може бути виконана з використанням різних типів регресійних моделей. Точність та ефективність реалізації цієї процедури залежать від типу регресійної використовуваної залежності.

Наступним дуже важливим етапом передобробки даних є процедура нормалізації або упорядкування значень атрибутів відповідно до однакового діапазону з однаковою нормою (в ідеалі — одиничною). Нормалізовані дані дають змогу

порівнювати досліджувані об'єкти, враховуючи набір усіх атрибутів, де кожен із них має однакову вагу. Сьогодні існує багато методів нормалізації, однак немає універсальної найкращої методики реалізації цього етапу, ефективної для всіх типів даних. Навіть більше: вибір методу нормалізації, з огляду на тип досліджуваних даних, суттєво впливає на ефективність наступного етапу розв'язання порушеної проблеми. Об'єктивний вибір відповідних методів нормалізації може бути здійснений лише на основі кількісних критеріїв якості з огляду на мету поставленої задачі та тип досліджуваних даних. Цей факт свідчить про актуальність теми дослідження в окресленій предметній галузі.

У цій статті наведено результати дослідження щодо об'єктивного вибору методу нормалізації з огляду на тип досліджуваних даних на основі застосування точності класифікації даних як основного критерію для оцінки якості нормалізованих даних.

На рис. 1 зображено блок-схему реалізації процедури вибору оптимального методу нормалізації з урахуванням типу досліджуваних даних. Вона передбачає такі етапи:

- формування матриці експериментальних даних;
- формування вектора доступних методів нормалізації, реалізація яких передбачається у процесі моделювання;
- вибір і налаштування методу класифікації об'єктів, що містять нормалізовані дані;
- формування критеріїв якості класифікації даних;
- нормалізація даних із їхньою подальшою класифікацією;
- розрахунок критеріїв якості класифікації даних;
- аналіз отриманих результатів, вибір оптимального методу нормалізації за критеріями якості класифікації відповідних об'єктів.

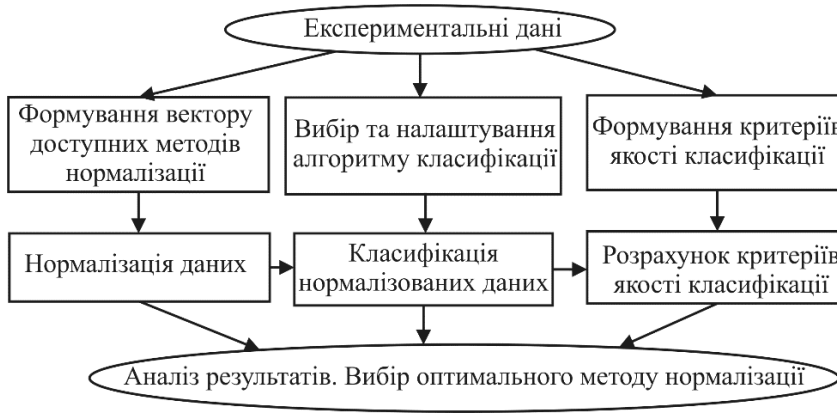

Рис. 1. Блок-схема процедури вибору оптимального методу нормалізації даних

Вищенаведені задачі розв'язуються у межах цього дослідження.

Аналіз останніх досліджень та публікацій. Чимало праць присвячено розв'язанню проблеми передобробки даних із різною природою. Так, у статті [5]

автори подали результати досліджень щодо аналізу даних з метою визначення шляхів їхньої подальшої передобробки та вибору функцій для підвищення точності методів машинного навчання. Етап попередньої обробки даних у цьому разі охоплював усунення шуму та нормалізацію даних. У статті [6] запропоновано метод квантильної нормалізації, який дає змогу трансформувати дані відповідно до нормального розподілу. Автори порівняли наведений метод з іншими відомими методами нормалізації і засвідчили, що запропонований варіант діє послідовно та ефективно, незалежно від вихідного розподілу. Результати досліджень щодо порівняльного аналізу різних методів нормалізації послідовності РНК наведені у науковій статті [7]. Автори подали результати моделювання із застосуванням двох типів даних: нормалізованих і ненормалізованих за семикратного перехресного контролю, використовуючи контрольні зразки клітин HepaRG людини, і нормалізовані дані із застосуванням семи методів нормалізації. Результати досліджень засвідчили перевагу методу квантильної нормалізації, порівняно з іншими, що використовувалися у процесі моделювання.

Однак, незважаючи на певні дослідження у цій предметній галузі, потрібно відзначити, що задача об'єктивного вибору оптимального методу нормалізації з урахуванням типу досліджуваних даних сьогодні не має однозначного розв'язку. У цій статті наведено один із можливих способів розв'язання задачі, заснований на послідовному застосуванні різних доступних методів нормалізації з вибором оптимального методу з погляду максимального значення точності класифікації даних, що досліджуються.

Мета статті — порівняльний аналіз різних методів нормалізації даних із використанням різноманітних типів наборів даних із метою вибору оптимального методу з погляду кількісного критерію якості класифікації досліджуваних об'єктів.

Виклад основного матеріалу дослідження. Методи нормалізації даних, що застосовувалися у процесі моделювання, наведені у таблиці. Вони доступні у пакеті *clusterSim* [8] мови програмування R [9].

Як класифікатор була використана нейронна мережа «багатошаровий персептрон». Оптимальні параметри мережі з погляду глобального мінімуму поверхні похибки (кількість нейронів, шарів, вагові коефіцієнти) визначалися за допомогою крос-валідації із застосуванням функції *train()* пакета *caret* [10] мови програмування R. Точність класифікації нормалізованих даних була використана як основний критерій для встановлення оптимальних параметрів нейронної мережі:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

де *TP, TN, FP, FN* — правильні позитивні, правильні негативні, неправильні позитивні і неправильні негативні випадки, передбачені класифікатором на тестових даних. У процесі моделювання дані розбивали на дві підмножини у відношенні: 70 % — для навчання мережі і 30 % — для тестування ефективності її роботи.

Таблиця

Методи нормалізації пакета *clusterSim*, що застосовувались у процесі моделювання

Індекс методу	Тип нормалізації	Індекс методу	Тип нормалізації
n1	(*x-mean*)/*sd*	*n7*	*x/range*
n2	(*x-median*)/*mad*	*n8*	*x/max*
n3	(*x-mean*)/*range*	*n9*	*x/mean*
n4	(*x-median*)/*range*	*n10*	*x/median*
n5	$\frac{x-mean}{max\left(abs\left(x-mean\right)\right)}$	*n11*	*x/sqrt*(*SSQ*)
n5a	$\frac{x-median}{max\left(abs\left(x-median\right)\right)}$	*n12*	$\frac{x-mean}{sqrt\left(sum\left(\left(x-mean\right)^2\right)\right)}$
n6	*x/sd*	*n12a*	$\frac{x-median}{sqrt\left(sum\left(\left(x-median\right)^2\right)\right)}$
n6a	*x/mad*	*n13*	(*x-midrange*)/(*0.5*range*)

Чотири різні типи наборів даних, що завантажені з репозиторію машинного навчання (UCI Machine Learning), були використані як експериментальні дані [11]:

– База даних ірисів Фішера (Iris) [12]. Містить 150 зразків трьох класів (по 50 у кожному: Setosa, Versicolour, Virginica) та чотири числові атрибути (довжина чашолистка (см), ширина чашолистка (см), довжина пелюстки (см) і ширина пелюстки (см).
– Набір даних про насіння (Seeds) [13]. Дані містять інформацію про ядра, що належать до трьох різних сортів пшениці: «кама», «роза» та «канадка», по 70 елементів у кожному класі, випадково вибрані внаслідок проведення експерименту. Кожен із досліджуваних об'єктів містить сім атрибутів: площа, периметр, компактність, довжина ядра, ширина ядра, коефіцієнт асиметрії, довжина борозенки ядра.
– Дані про склад вина [14]. Містить три види вина (три класи). Кожен із досліджуваних об'єктів характеризується 13 атрибутами: алкоголь, яблучна кислота, зола, лужність золи, магній, загальні феноли, флавоноїди, нефлавоноїдні феноли, проантоціанідини, інтенсивність забарвлення, відтінок, OD280/OD315 розведених вин, пролін.
– База даних для ідентифікації скла [15]. Містить сім видів скла (7 класів) та 7 атрибутів: показник заломлення, натрій, магній, алюміній, кремній, калій, кальцій.

На рис. 2 відтворено характер розподілу об'єктів у класах для наборів даних Iris та Seeds. У цьому разі використано дві головних компоненти. Як видно, об'єкти можна адекватно розділити на класи, хоча деякі перетинаються між собою. На рис. 3

наведено аналогічні результати для наборів даних Wine та Glass із застосуванням графіків у паралельних координатах (у цих випадках дві головні компоненти мали низькі значення дисперсії). Як можна бачити, набори даних Wine та Glass є більш складними, як і розподіл об'єктів на класи в цих випадках. На рис. 4 зображено характер розподілу атрибутів даних, що досліджуються із використанням діаграми розмаху. Відповідно до зображень, діапазони зміни відповідних атрибутів для всіх наборів даних дуже різні. Цей факт ускладнює якісну класифікацію даних або іншу процедуру машинного навчання чи інтелектуального аналізу даних.

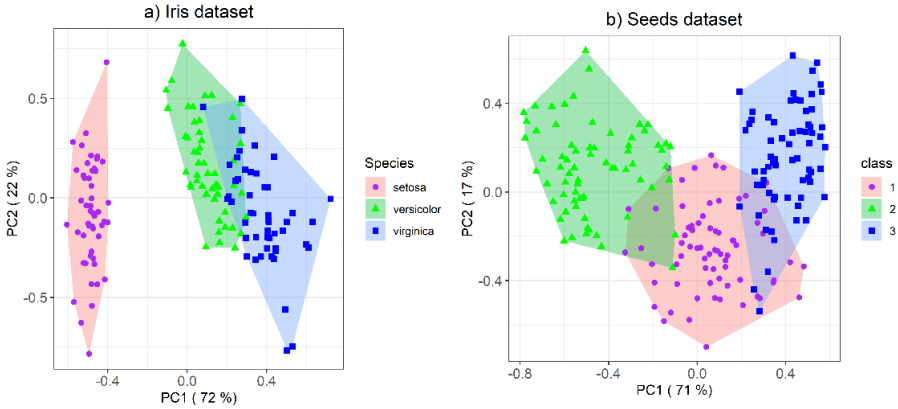

Рис. 2. Розподіл об'єктів за класами даних іриси Фішера (Iris) і насіння (Seeds)

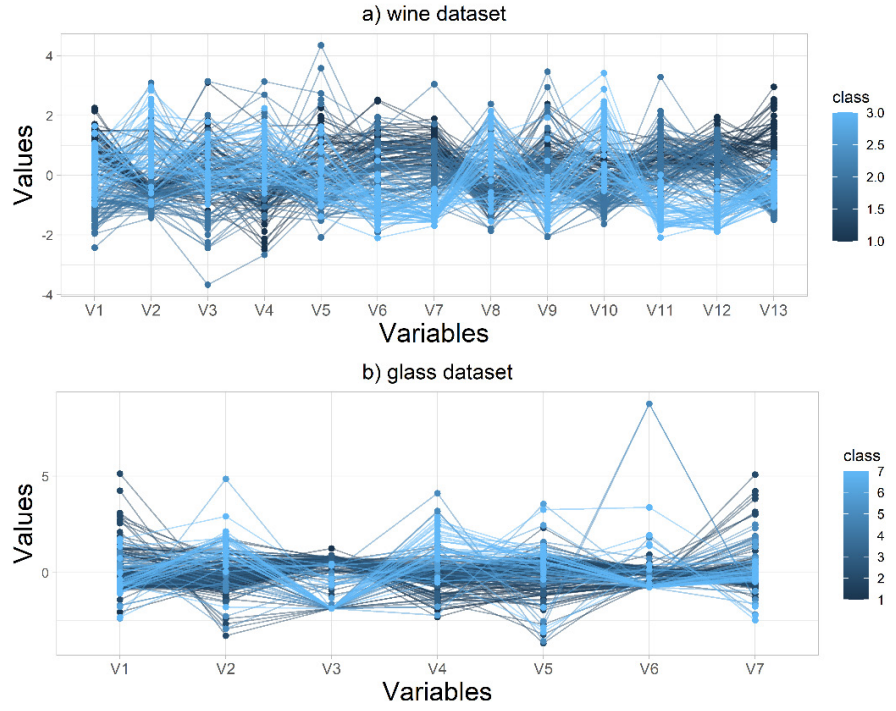

Рис. 3. Графіки у паралельних координатах для даних Wine і Glass

На рис. 5 відтворено точкові діаграми залежності точності класифікації об'єктів нормалізованих даних під час застосування нейронної мережі «багатошаровий персептрон» від методу нормалізації. У цьому разі метод *n0* відповідає ненормалізованим даним.

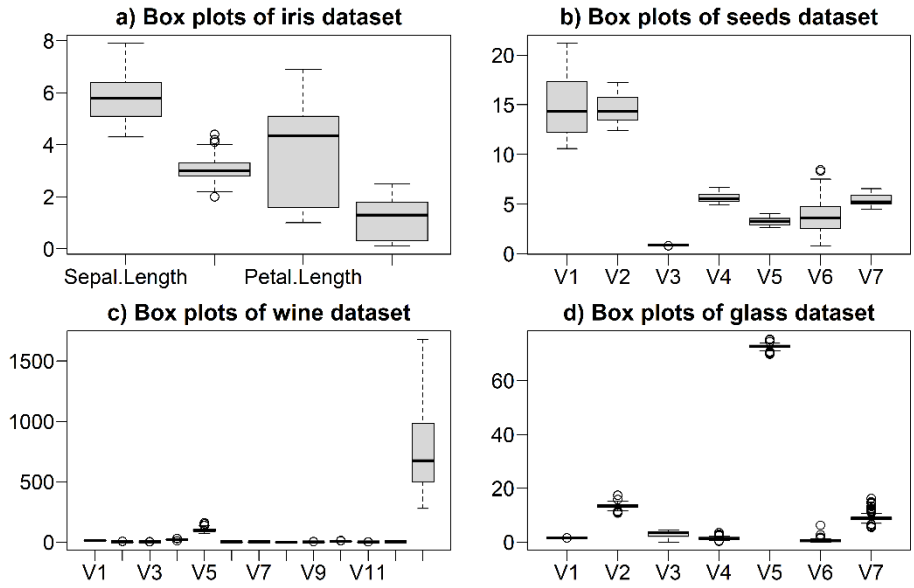

Рис. 4. Діаграми розмаху ненормалізованих даних

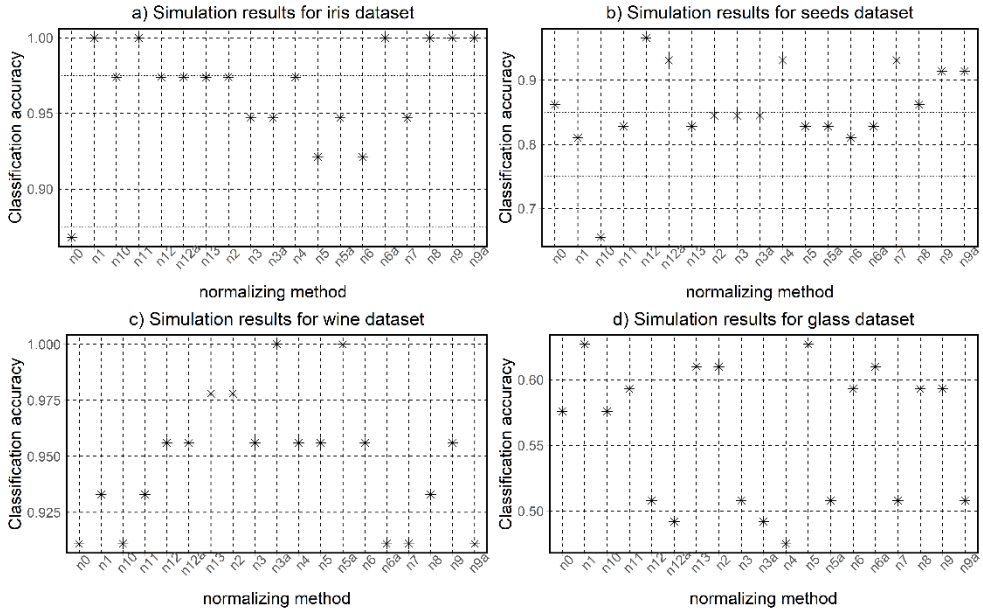

Рис. 5. Результати моделювання щодо визначення оптимального методу нормалізації атрибутів даних

Аналіз отриманих результатів дає змогу зробити висновок, що методи нормалізації, які відповідають максимальному значенню точності класифікації, різні для різноманітих наборів даних. Отже, методи нормалізації *n1*, *n6a, n8, n9* та *n11* є оптимальними для набору даних іриси Фішера. У цьому випадку отримано 100 %-ву точність класифікації після застосування тестового набору даних. Метод нормалізації *n11* є оптимальним для даних про насіння. У цьому випадку була отримана найвища (майже максимальна) точність класифікації. Методи *n3a* та *n5a* є оптимальними для даних Wine. Нормалізація даних за допомогою цих методів відповідає максимальному значенню точності класифікації на тестових даних. У разі використання набору даних про скло оптимальними є методи нормалізації *n1* та *n5*. Однак значення точності класифікації в цих випадках є меншим, порівняно з іншими використовуваними наборами даних. У цьому випадку необхідно розглянути атрибути, застосовуючи сучасні методи статистичного аналізу.

На рис. 6 зображено діаграми розмаху нормалізованих даних із використанням оптимальних методів нормалізації.

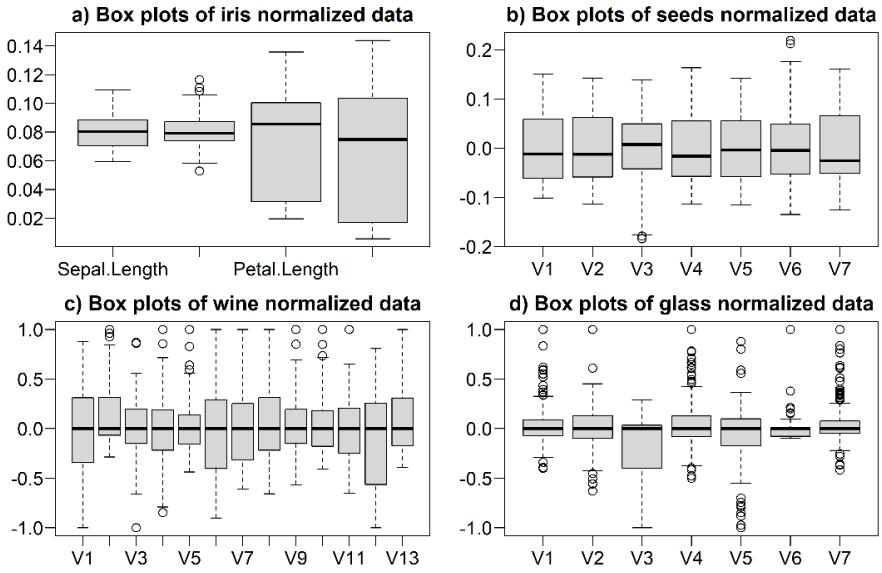

Рис. 6. Діаграми розмаху нормалізованих даних із використанням оптимальних методів нормалізації

Аналіз отриманих діаграм розмаху підтверджує наведені вище висновки про високу якість обробки (нормалізації) наборів даних про іриси Фішера, насіння та вино, оскільки значення усіх атрибутів розподіляються в одному діапазоні адекватно. У випадку використовуваного набору даних про скло дані містять багато викидів, і цей факт може впливати на точність класифікації.

Однак, на наш погляд, запропонований метод може дати нам змогу вибрати оптимальний метод нормалізації з огляду на досліджуваний набір даних.

Висновки. У статті наведено результати дослідження щодо вибору оптимального методу нормалізації даних із погляду максимального значення точності класифікації досліджуваних даних. Як експериментальні дані в процесі моделювання використовувались чотири різні типи наборів даних: про іриси Фішера, насіння, вино та скло. Створені діаграми розмаху розподілу значень атрибутів даних відтворили дуже різні діапазони варіацій атрибутів даних. У межах нашого дослідження були використані різні методи нормалізації, доступні в пакеті *clusterSim* мови програмування R. Нейронна мережа «багатошаровий персептрон» використовувалася як класифікатор у процесі моделювання. Оптимальні параметри мережі з погляду глобального мінімуму поверхні помилок оцінювали завдяки кросвалідації за допомогою функції *train()* пакета *caret* (програмне забезпечення R).

Результати моделювання засвідчили, що методи нормалізації, які відповідають максимальному значенню точності класифікації, різні для різноманітних наборів даних. Навіть більше: для трьох наборів даних отримано максимальне (або майже максимальне) значення точності класифікації тестових наборів даних. У разі використання набору даних про скло отримано дещо меншу точність класифікації. Цей факт можна пояснити наявністю багатьох викидів у даних, і це може вплинути на точність класифікації. Однак у будь-якому разі запропонований варіант може дати нам змогу об'єктивно вибрати оптимальний метод нормалізації даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. De Silva A., De Livera A., Lee K. et al. Multiple imputation methods for handling missing values in longitudinal studies with sampling weights: Comparison of methods implemented in stata. Biometrical Journal. 2021. Vol. 63 (2). Pp. 354–371.
2. Johnson T., Isaac N., Paviolo A. et al. Handling missing values in trait data. Global Ecology and Biogeography. 2021. Vol. 30 (1). Pp. 51–62.
3. Kim K. H., Kim K. J. Missing-data handling methods for lifelogs-based wellness index estimation: Comparative analysis with panel data. JMIR Medical Informatics. 2021. Vol. 8 (12), art. no. e20597.
4. Modulo 9 model-based learning for missing data imputation / Ngueilbaye A., Wang H., Mahamat D., Junaidu S. Applied Soft Computing. 2021. Vol. 103, art. no. 107167.
5. Sharma S., Sood M. Exploring feature selection technique in detecting sybil accounts in a social network. Advances in Intelligent Systems and Computing. 2020. Vol. 1166. Pp. 695–708.
6. Peterson R., Cavanaugh J. Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. Journal of Applied Statistics. 2020. Vol. 47 (13–15). Pp. 2312–2327.
7. Bushel P., Ferguson S. et al. Comparison of normalization methods for analysis of tempo-seq targeted RNA sequencing data. Frontiers in Genetics. 2020. Vol. 11, art. no. 594.
8. Clustersim package. URL: http://keii.ue.wroc.pl/clusterSim/.
9. Ihaka R., Gentleman R. R: a linguage for data analysis and graphics. Journal of Computational and Graphical Statistics. 1996. Vol. 5 (3). Pp. 299–314.
10. Caret package. URL: https://topepo.github.io/caret/.
11. UCI - machine learning repository. URL: https://archive.ics.uci.edu/ml/datasets.php.

12. Fisher R. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936. Vol. 7 (2). Pp. 179–188.
13. Seeds dataset. URL: https://archive.ics.uci.edu/ml/datasets/seeds.
14. Wine recognition data. URL: https://archive.ics.uci.edu/ml/datasets/wine.

REFERENCES

1. De Silva, A., De Livera, A., & Lee, K. et al. (2021). Multiple imputation methods for handling missing values in longitudinal studies with sampling weights: Comparison of methods implemented in stata: Biometrical Journal, 63 (2), 354–371 (in English).
2. Johnson, T., Isaac, N., & Paviolo, A. et al. (2021). Handling missing values in trait data: Global Ecology and Biogeography, 30 (1), 51–62 (in English).
3. Kim, K. H., & Kim, K. J. (2021). Missing-data handling methods for lifelogs-based wellness index estimation: Comparative analysis with panel data: JMIR Medical Informatics, 8 (12), art. no. e20597 (in English).
4. Ngueilbaye, A., Wang, H., Mahamat, D., & Junaidu, S. (2021). Modulo 9 model-based learning for missing data imputation: Applied Soft Computing, 103, art. no. 107167 (in English).
5. Sharma, S., & Sood, M. (2020). Exploring feature selection technique in detecting sybil accounts in a social network: Advances in Intelligent Systems and Computing, 1166, 695–708 (in English).
6. Peterson, R., & Cavanaugh, J. (2020). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era: Journal of Applied Statistics, 47 (13–15), 2312–2327 (in English).
7. Bushel, P., & Ferguson, S. et al. (2020). Comparison of normalization methods for analysis of tempo-seq targeted RNA sequencing data: Frontiers in Genetics, 11, art. no. 594 (in English).
8. Clustersim package. Retrieved from http://keii.ue.wroc.pl/clusterSim/ (in English).
9. Ihaka, R., & Gentleman, R. (1996). R: a linguage for data analysis and graphics: Journal of Computational and Graphical Statistics, 5 (3), 299–314 (in English).
10. Caret package. Retrieved from https://topepo.github.io/caret/ (in English).
11. UCI - machine learning repository. Retrieved from https://archive.ics.uci.edu/ml/datasets.php (in English).
12. Fisher, R. (1936). The use of multiple measurements in taxonomic problems: Annals of Eugenics, 7 (2), 179–188 (in English).
13. Seeds dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/seeds (in English).
14. Wine recognition data. Retrieved from https://archive.ics.uci.edu/ml/datasets/wine (in English).

doi: 10.32403/0554-4866-2021-1-81-35-44

MODEL OF THE DATA NORMALIZATION QUALITY EVALUATION ON THE BASIS OF APPLICATION OF THE OBJECTS CLASSIFICATION QUALITY CRITERIA

L. M. Yasinska-Damri, B. V. Durnyak

Ukrainian Academy of Printing,
19, Pid Holoskom St., Lviv, 79020, Ukraine
Lm.yasinska@gmail.com

The paper presents a comparative analysis of various types of normalization techniques. The accuracy of data classification which was carried out after data normalizing was used as the main criterion for evaluating the quality of the appropriate normalizing method. Four various types of datasets downloaded from the UCI Machine Learning Repository were used as the experimental data during the simulation process. Various normalization techniques available from package clusterSim of R software were applied to the experimental data. The quality of the data normalizing procedure was evaluated based on the use of data classification by the calculation of the accuracy of the distribution of the objects into classes. The neural network multilayer perceptron was used as the classifier at this step.

Four different types of datasets were used as the experimental data during the simulation procedure: Iris Plants, Seeds, Wine and Glass. The simulation results have shown that the data normalizing stage significantly influences the classification accuracy and selection of the normalization method depends on the type of data and, consequently, the selection of the normalizing technique should be carried out in each of the cases separately.

The analysis of the obtained results allows also concluding that the normalization methods that correspond to maximum value of the classification accuracy are different for various datasets. So, the normalization methods n1, n6a, n8, n9 and n11 are the optimal ones for the iris dataset. In this case, 100% classification accuracy is obtained for test dataset. The normalizing technique n11 is optimal one for the seeds data. The highest (almost maximal) classification accuracy was received in this case. The methods n3a and n5a are optimal ones for complex wine data. In the case of the glass dataset use, the n1 and n5 normalization methods are optimal ones.

Keywords: *data normalization, classification quality criteria, multilayer perceptron, data processing, classification accuracy.*

Стаття надійшла до редакції 19.05.2021.
Received 19.05.2021.