УДК 004.8:004.43:378

ВИКОРИСТАННЯ БАГАТОВИМІРНОГО НОРМАЛІЗУЮЧОГО ПЕРЕТВОРЕННЯ БОКСА-КОКСА ДЛЯ ПІДВИЩЕННЯ ЯКОСТІ КЛАСТЕРИЗАЦІЇ МЕТОДОМ K-MEANS

Приходько С.Б., д.т.н., проф.[1]; Кудін О.О., к.т.н. [2]
[1,2]Національний університет кораблебудування імені адмірала Макарова
Україна, Миколаїв
[1]sergiy.prykhodko@nuos.edu.ua; [2]oleh.o.kudin@gmail.com

Анотація. *В роботі на прикладі відомого чотиривимірного набору даних Ireses показана можливість підвищення якості кластеризації методом K-Means завдяки використанню нормалізації даних за допомогою багатовимірного перетворення Бокса – Кокса. Використання попередньої стандартизації означеного набору даних навпаки призвело до зменшення значень показників якості у порівнянні з початковим і нормалізованим варіантами набору даних Ireses.*

Ключові слова: *метод кластеризації K-Means; багатовимірне перетворення Бокса – Кокса; показники якості кластеризації; скоригований індекс Ренда; скоригована взаємна інформація.*

Вступна частина. При практичному використанні методу кластерного аналізу K-Means зустрічаються такі багатовимірні помічені набори даних у яких окремі елементи однієї групи мають числові значення ознак, характерні для іншої групи тієї ж вибірки. В цих випадках метод K-Means іноді створює кластери, які перетинаються між собою. Отже, виникає питання: Чи можливо коректно змінити вхідні дані перед кластеризацією для підвищення її якості, а саме, коректного розподілу елементів набору даних по кластерам відповідно до їх міток?

Мета роботи. Порівняти результатів кластеризації методом K-Means багатовимірного поміченого набору даних для варіантів первинного, стандартизованого та нормалізованого (за допомогою багатовимірного нормалізуючого перетворення Бокса – Кокса) набору для визначення кращого варіанту за якістю кластеризації.

Основна частина. Для проведення обчислювальних експериментів був обраний відомий набір даних Irises (https://archive.ics.uci.edu/dataset/53/iris). Цей набір даних містить чотири стовпчики з даними про довжину чашолистків (sepal length), ширину чашолистків (sepal width), довжину пелюсток (petal length), ширину пелюсток (petal width) та один стовпчик міток, який визначає сорт ірісу (setosa, versicolor, virginica).

При стандартизації набору було використано наступне рівняння:

$$x_i^s = \frac{x_i - \overline{x}}{\sigma_x} + 4, \qquad (1)$$

де x_i^s – стандартизоване значення i-ої ознаки (i = 1, 2, …, n); n – кількість елементів набору даних; x_i – емпіричне значення ознаки; $\overline{x}$ – середнє значення ознаки; середньоквадратичне відхилення. У (1) додано 4 для того, щоб усі стандартизовані значення були додатними.

Для нормалізації набору даних Ireses, як і в роботі [1], ми застосовували чотиривимірне перетворення Бокса-Кокса. Як і в [1], використовувалися такі змінні: Y – sepal length; X_1 – sepal width; X_2 – petal length; X_3 –petal width. Оцінки параметрів чотиривимірного перетворення

Бокса – Кокса з [1] для зазначених змінних є наступними: $\hat{\lambda}_y = -0,320736$; $\hat{\lambda}_1 = 1,010215$; $\hat{\lambda}_2 = 0,626423$; $\hat{\lambda}_3 = 0,593540$.

При оцінюванні якості кластеризації для всіх варіантів набору даних Irises було використано скоригований індекс Ренда (adjusted rand score) та скоригована взаємна інформація (adjusted mutual info score) [2, 3].

В таблицях 1-3 представлено розподіл ірісів по кластерам. Можна побачити, що найкраще іріси розподілились по кластерам відповідно до їх міток у табл. 3, яка відповідає варіанту нормалізованих даних на основі багатовимірного перетворення Бокса-Кокса. Відповідно до табл. 4 цей варіант даних також має найбільші значення скоригованого індексу Ренда та скоригованої взаємної інформації.

Таблиця 1 – Первинні дані

	setosa	versicolor	virginica
Кластер 1	50	0	0
Кластер 2	0	48	14
Кластер 3	0	2	36

Таблиця 2 – Стандартизовані дані

	setosa	versicolor	virginica
Кластер 1	50	0	0
Кластер 2	0	39	14
Кластер 3	0	11	36

Таблиця 3 – Нормалізовані дані

	setosa	versicolor	virginica
Кластер 1	50	0	0
Кластер 2	0	6	48
Кластер 3	0	44	2

Таблиця 4 – Показники якості кластеризації

Тип даних	Індекс Ренда	Взаємна інформація
Первинні	0,73	0,76
Стандартизовані	0,62	0,66
Нормалізовані	0,85	0,84

При детальному аналізі розподілу ірісів по кластерам видно, що найгірше вони розподілені по кластерам при використанні для кластеризації методом K-Means стандартизованих даних. Для стандартизованих даних значення скоригового індексу Ренда теж мають найменші значення. Тобто у даному випадку стандартизація емпіричних даних погіршила результати кластеризації методом K-Means.

Відповідно до табл. 4 найбільше значення показників якості мають результати кластеризації даних, нормалізованих за допомогою багатовимірного перетворення Бокса-Кокса. Результати розподілу типів даних по кластерам також отримані для нормалізованих даних (табл. 3) мають кращі значення у порівнянні з емпіричним та стандартизованим варіантом.

Отже, використання багатовимірного нормалізуючого перетворення Бокса – Кокса для перетворення набору даних Irises дозволило покращити показники якості кластеризації. Це можна пояснити тим, що при нормалізації даних за допомогою багатовимірного нормалізуючого перетворення враховується кореляція між змінними.

Висновки. На основі проведених комп’ютерних експериментів можна зробити наступні висновки:

1. Використання багатовимірного нормалізуючого перетворення Бокса – Кокса дозволило підвищити показники якості кластеризації поміченого набору даних Irises у порівнянні з первинним та стандартизованим варіантами цього набору.
2. Застосування попередньої стандартизації до первинного набору даних Irises призвело до погіршення показників якості кластеризації. Це вказує на те, що використання попередньої стандартизації вхідного набору даних не завжди дозволяє підвищити показники якості кластеризації за методом K-Means.

Використання інструментів ШІ. При проведені обчислень використано модулі cluster та metrics бібліотеки scikit-learn для мови програмування Python версії 1.9.1. Наукові положення, висновки та результати є власним інтелектуальним внеском авторів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Prykhodko, S., & Prykhodko, N. (2023). Building nonlinear regression models for estimating the number of clusters and their initial centroids. *Computer Science and Information Technologies :2023 IEEE 18th International Conference (CSIT)*, (pp. 1-4). Lviv, Ukraine. https://doi.org/10.1109/CSIT61576.2023.10324095

[2] Gates, A.J, & Ahn, Y.-Y. (2017). The impact of random models on clustering similarity. *Journal of Machine Learning Research*, *1* (18), 1-28. https://www.jmlr.org/papers/volume18/17-039/17-039.pdf

[3] Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *1* (11), 2837-2854. https://jmlr.csail.mit.edu/papers/volume11/vinh10a/vinh10a.pdf

Prykhodko S.B., Kudin O.O.

USING THE MULTIVARIATE BOX-COX NORMALIZING TRANSFORMATION TO IMPROVE THE QUALITY OF K-MEANS CLUSTERING

Abstract*. The work shows the possibility of improving the quality of K-Means clustering by data normalization using the multivariate Box-Cox transformation based on the well-known four-dimensional Ireses data example. The use of pre-standardization of the specified data set, on the contrary, led to a decrease in the values of quality indicators compared to the initial and normalized versions of the Ireses data set.*

Keywords*: K-Means clustering method; multivariate Box-Cox normalizing transformation; clustering quality metrics; Adjusted Rand Index; Adjusted Mutual Information.*