DOI 10.34185/1562-9945-3-128-2020-15
УДК 004.4

К.Ю. Островська, Р.В. Бедай

ДОСЛІДЖЕННЯ ПРОДУКТИВНОСТІ МЕТОДІВ НОРМАЛІЗАЦІЇ ОБ'ЄМНИХ ДАНИХ

Анотація. Робота присвячена дослідженню продуктивності методів нормалізації об'ємних даних. Робота відноситься до області обробки поста експериментальних і статистичних даних, яка полягає в перетворенні вхідного набору даних в вихідний в конкретному інтервалі (нормалізація).
В рамках роботи було досліджено актуальні на сьогоднішній день методи нормалізації з метою їх застосування для нормалізації числових даних зі збереженням співвідношення. Була розроблена бібліотека, реалізує підійшли під цей критерій методи, що дозволяє нормалізувати і візуалізувати вихідні дані.
Ключові слова: нормалізація, методи, бенчмарк, лаунчер, дамп, обробка даних, продуктивність, об'ємні дані.

Вступ. Нормалізація даних в даний час широко використовується в різних областях науки і техніки, а не тільки в середовищі інформаційних технологій.

Проблеми:

- нормалізація об'ємних даних. Це потенційно проміжний стан між простими даними і Big Data. У цьому випадку вже необхідно враховувати обсяг, але все ще не потрібно використовувати Machine Learning-рішення.

- відсутність спеціалізованої бібліотеки.

Також проблема актуальна для областей науки і техніки, де застосовується статистика на основі часу. Наприклад, існує зібрана статистика роботи якогось додатка в мілісекундах за різні періоди.

Для розуміння залежностей в роботі такого додатка буде необхідно нормалізувати статистику в один відрізок, щоб можна було робити точні висновки.

Для вирішення такого роду завдань і будуть розглядатися зазначені питання в рамках даної роботи.

Постановка задачі. Для дослідження продуктивності методів нормалізації необхідно провести теоретичний аналіз існуючих алгоритмів. Для і тому необхідні теоретичну оцінку складності O(N), де N - кількість сутностей в наборі даних. Також необхідно провести оцінку.

Залежно алгоритмічної складності від m - довжини (об'ємності) полів даних. Таким чином завдання вимагає теоретичної оцінки складності по двом параметрам O(m, N).

Далі потрібно буде виконати програмну реалізацію методів нормалізації з візуалізацією для наочного представлення результатів нормалізації до і після роботи алгоритму.

Після реалізації необхідно буде провести експериментальне порівняння і оцінити, наскільки теоретичні дані підтверджуються експериментальними.

Результатом роботи буде бібліотека, за допомогою якого користувач зможе виконувати нормалізацію даних і візуально оцінювати результати роботи за графіками. Користувачеві для цього буде доступний публічний інтерфейс. Основними програмними компонентами є наступні модулі:

- модуль, який відповідає за безпосередньо нормалізацію;

- модуль, який відповідає за візуалізацію;

- допоміжний модуль з бенчмарками для оцінки продуктивності того чи іншого алгоритму;

Кінцевому користувачеві в бібліотеці будуть доступні всі три модулі.

Бібліотека, яка буде отримана на виході, є універсальною.

Тобто її можна використовувати не тільки в рамках тих прикладів, що наводяться в рамках даної роботи, а й в будь-якому іншому випадку, коли потрібно нормалізувати дані, починаючи від використання в курсових роботах студентами технічних факультетів і закінчуючи лабораторними дослідженнями і додатків, що працюють зі статистикою.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Методи нормалізації даних.

1. Мінімаксний метод є найпростішим і популярним методом нормалізації. Він входить до групи статистичних методів нормалізації.

Для розрахунку потрібно знати тільки мінімальне і максимальне значення нормалізуемих ряду значень.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

де x' - нормалізоване поточне значення,

x - поточне ненормалізоване значення

x_{min} - мінімальне значення в ненормализованном числовому ряду, MAX

x_{max} - максимальне значення в ненормализованном числовому ряду.

Так як даний метод є лінійно перетворює, він потенційно підходить для вирішення поставленого завдання.

2. Метод нормалізації за середнім, належить до групи статистичних методів нормалізації.

Крім мінімального і максимального значення ненормалізованих числового ряду, тут потрібно знати і середнє значення по даному ряду.

$$x_{avg} = \frac{\sum_{1}^{N} x_n}{N}, \quad (2)$$

де N - кількість значень в числовому ряду.

Метод нормалізації за середнім представлений формулою (3).

$$x' = \frac{x - x_{avg}}{x_{max} - x_{min}} \quad (3)$$

де x' - нормалізоване поточне значення,

x - поточне ненормалізоване значення,

x_{avg}- середнє значення числового ненормалізованого ряду,

x_{min} - мінімальне значення в ненормалізованому числовому ряді,

x_{max} - максимальне значення в ненормалізованому числовому ряді.

3. Нормалізація стандартним відхиленням. Для обчислення використовуються статистичні характеристики, такі як середнє і стандартне відхилення (формула 4).

$$x' = \frac{x - x_{avg}}{\sigma_x} \quad (4)$$

де x' - нормалізоване поточне значення,

x - поточне ненормалізоване значення,

x_{avg} - середнє значення числового ненормалізованих ряду,

σ_x - стандартне відхилення ряду.

Стандартне відхилення обчислюється за формулою 5.

$$\sigma_x = \sqrt{\frac{\sum_1^n (x_n - x_{avg})^2}{n}}, \quad (5)$$

де x_n - n-й елемент числового ряду,

x_{avg} - середнє значення числового ненормалізованих ряду,

n - число елементів в ряду.

Для поставленого завдання будемо вважати його придатним.

В якості алгоритму сортування списків зі значеннями полів будемо використовувати сортування злиттям.

Загальна структура проекту наведена на UML-діаграмі, зображеної на рисунку 1.

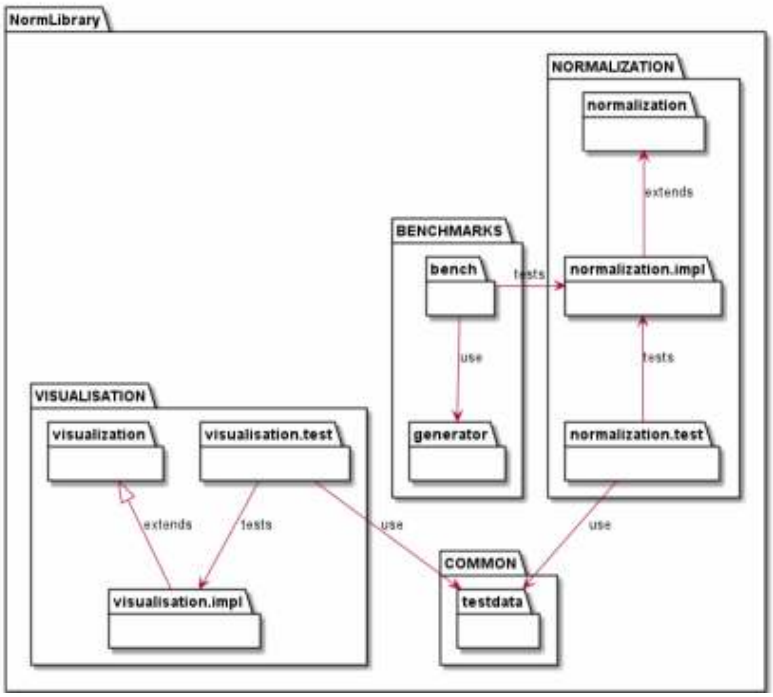

Рисунок 1 - Структура проекту бібліотеки

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Умовним "ядром" бібліотеки є пакет normalization.impl - саме всередині нього реалізовані в явному вигляді методи нормалізації. Цей пакет базується на пакеті normalization, що містить в собі відповідний інтерфейс.

Повна UML-діаграма модуля normalization приведена на рисунку 2.

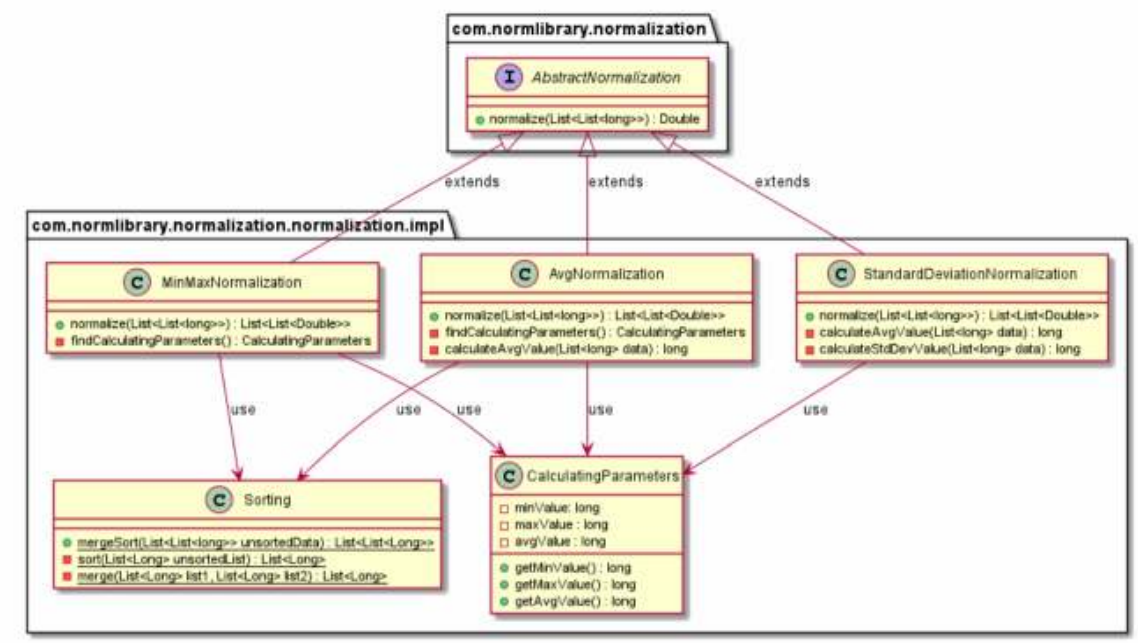

Рисунок 2 - Повна UML-діаграма модуля normalization

Допоміжним комплексом в бібліотеці є пакет benchmarks.

Даний пакет містить класи реалізує еталонні тести продуктивності (бенчмарки). Він необхідний для того, щоб перевірити на практиці твердження, зроблені в кінці розділу. Для перевірки використовується пакет generator в якому генеруються вихідні дані для тестування - масив масивів типу long, виходячи з поставлених вручну параметрів: кількість ознак (полів) і довжина набору, загальна для всіх ознак.

Оцінка продуктивності. Для оцінки продуктивності методів нормалізації будуть виконуватися бенчмарки для декількох дампов даних.

Дампи покриватимуть інтервал від 1 до 10 Гб, щоб виконати тестування навантаження на всьому інтервалі, на якому визначено поняття об'ємних даних. Виходячи з цього дампов буде 5, і вони будуть мати наступні розміри: 1,1 Гб, 2,3 Гб, 4,5 Гб, 6,4Гб, 9,1 Гб. Зняття бенчмарков реалізовано за допомогою фреймворка JMH.

Дампи генерувалися методом generateTestVolumeData класу DataGenerator. Метод приймає в якості параметрів кількість списків і кількість елементів в цих списках, а потім повертає згенеровані дані у вигляді списку списків типу Long.

В рамках підготовки бенчмарков в класі AbstractNormalizationBenchmark кількість елементів в списках було вибрано фіксованим в 750 тисяч елементів для зручності перемикання між дампами в процесі тестування, а змінним параметрів стало кількість списків, яке визначалося через клас-спадкоємець NormalizationState.

Розрахунок необхідної кількості списків виконувався виходячи з моделі даних в Java. Тип Long в даному випадку - це не примітивний тип, а тип класу- обгортки, тобто об'єкт. Тому він буде займати в пам'яті стандартні 64 біта для типу long (примітивного) плюс ще 32 біта, так як в даному випадку ми оперуємо посилальним типом і ці 4 байта займе безпосередньо посилання.

Далі на основі обсягу одного поля типу Long, кількості елементів і необхідного обсягу даних обчислювалося кількість списків для тестування.

Для кожного дампа даних отримаємо довірчий інтервал.

Довірчий рівень a обраний 99%. Для цього для кожного дампа по вимірювальним ітераціям розрахуємо похибку.

$$Z_{a/2} \cdot \frac{\sigma}{\sqrt{N}} \qquad (6)$$

де $Z_{a/2}$ - оцінка на основі коефіцієнта довіри,

a/2 - коефіцієнт довіри,

σ - стандартне відхилення.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

$$\sigma = \sqrt{\frac{\sum_{1}^{n}(x_n - x_{avg})^2}{n}}, \quad (7)$$

де x_n - n-й елемент числового ряду,

x_{avg} - середнє значення числового ряду,

n - число елементів в ряду.

Виходячи з довірчого рівня, отримуємо довірчий коефіцієнт 0,495. Йому відповідає Z-оцінка 2,58.

Результати обчислень за всіма дампи зведені в таблицю 1.

Таблиця 1

Довірчий інтервал для всіх методів по дампи

	Дамп, Гб									
	1,1		2,3		4,3		6,4		9,1	
	$\bar{t}$	±Δ	$\bar{t}$	±Δ	$\bar{t}$	±Δ	$\bar{t}$	±Δ	$\bar{t}$	±Δ
MM	11	0,3	23	10	60	10	69	36	110	38
A	11	0,3	25	10	58	12	81	32	106	40
SD	0,58	0,1	2	1,1	12	3	20	7	19	21

Розраховані значення з таблиці 10 відображені в графічному вигляді на рисунку 3.

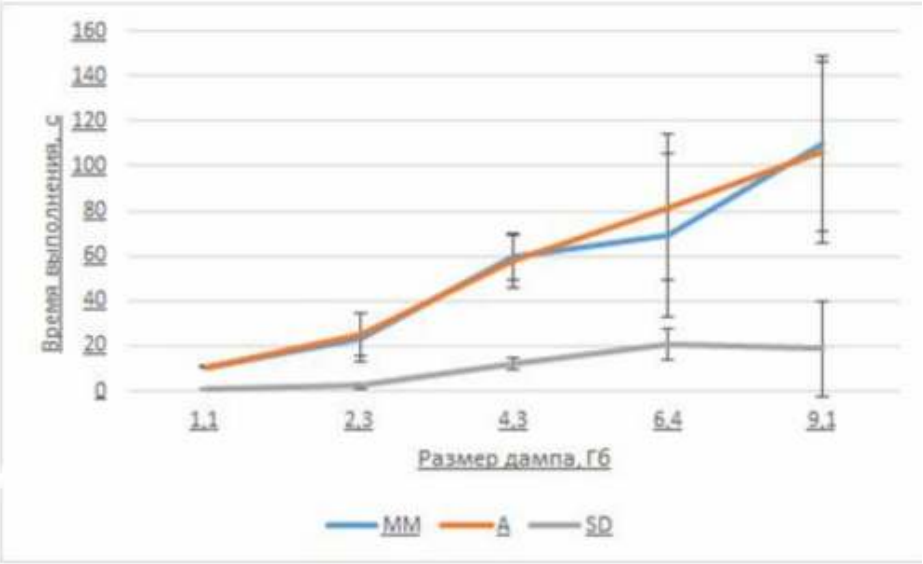

Рисунок 3 - Графічне представлення результатів виконання бенчмарків

Як видно з рисунка 3, теоретичний висновок, підтвердився практикою: метод нормалізації стандартним відхиленням менше складний. А значить і більш продуктивний, ніж мінімаксний метод або метод нормалізації за середнім. Це обумовлюється в першу чергу відсутністю необхідності сортувати дані, щоб отримати мінімальну і максимальне значення. В рамках планування розробки цей момент враховувався і отримання цих значень замість O(N) займає O(N·logN) часу. Але, незважаючи на це, метод нормалізації середньоквадратичним відхиленням показав себе в рази краще.

Отримані результати. Результати проведеного тестування навантаження підтверджують теоретичні викладки: метод нормалізації стандартним відхиленням має менший порядок складності, а, отже, більш високу продуктивність, ніж мінімаксний метод і метод нормалізації за середнім. Два останніх методу відповідно до теоретичними викладками мають однакову алгоритмічну складність, а на практиці мінімаксний метод показує трохи більше зростання складності з збільшенням обсягу даних.

Метод нормалізації стандартним відхиленням виграє по більшій частини за рахунок відсутності сортування, незважаючи на те, що сортування була введена для зменшення складності алгоритму.

Висновки. В рамках роботи був виконаний аналіз виду даних, сформовано його визначення. Для певного типу даних був зроблений відбір відповідних алгоритмів нормалізації, проведений їх аналіз і реалізація у вигляді бібліотеки. На готовій бібліотеці було проведено тестування навантаження з метою визначення найкращого методу нормалізації для певного типу даних.

Відібрані для реалізації методи нормалізації виправдали себе з точки зору постановки завдання: нормалізація виконується коректно.

Складність реалізованих алгоритмів також задовольняє поставленим вимогам. Відповідність поставленому завданню і вимогам підтверджено експериментальними даними.

Реалізована бібліотека показала себе ефективною, легковажною і зручною у використанні: цим були досягнуті поставлені в роботі мети.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Бібліотека відповідає заявленому функціоналу. Подальший план по реалізації бібліотеки: вивантаження в публічний репозиторій, підтримка, розширення, написання документації.

ЛИТЕРАТУРА / ЛІТЕРАТУРА

1. Майер-Шенбергер В. Большие данные. Революция, которая изменит то, как мы живём, работаем и мыслим / Майер-Шенбергер В., Кукьер К. — М.: Манн, Иванов, Фербер, 2014. — 240 с.
2. Стивенс Р. Алгоритмы. Теория и практическое применение. – Москва: Издательство «Э», 2016. –544 с.
3. Kreyszig E. Advanced Engineering Mathematics. – Wiley, 1979. – 880 с.
4. William H. Greene Econometric analysis. - New York: Pearson Education, Inc., 2003. - 1026 с.
5. Бенчмарки [Електронний документ]. (Https://wikipedia.org/wiki/Тест_производительности).
6. Дисперсія, стандартне відхилення і коефіцієнт варіації [Електронний документ]. - (https://statanaliz.info/metody/opisanie-dannyx/11-dispersiya-standartnoe-otklonenie-koeffitsient-variatsii).
7. Огляд методів попередньої обробки даних [Електронний документ]. - (http://www.math.spbu.ua/SD_AIS/documents/2019-12-341/2019-12-b-17.pdf).
8. Обробка об'ємних даних, Gwyddion [Електронний документ]. - (http://gwyddion.net/documentation/user-guide-ua/volume-data-processing.html).
9. Об'ємні дані [Електронний документ]. – (http://www.teamnet.ua/gruppa-teamnet/issledovanie-i-razrobotka/opros-mnenij- obemnye-dannye/).

REFERENCES

1. Mayer-Shenberger V. Bolshie dannyie. Revolyutsiya, kotoraya izmenit to, kak myi zhivYom, rabotaem i myislim / Mayer-Shenberger V., Kuker K. — M.: Mann, Ivanov, Ferber, 2014. — 240 s.
2. Stivens R. Algoritmyi. Teoriya i prakticheskoe primenenie. – Moskva: Izdatelstvo «E», 2016. –544 s.
3. Kreyszig E. Advanced Engineering Mathematics. – Wiley, 1979. – 880 s.
4. William H. Greene Econometric analysis. - New York: Pearson Education, Inc., 2003. - 1026 s.
5. Benchmarki [Elektronniy dokument].

(Https://wikipedia.org/wiki/Test_proizvoditelnosti).
6. DispersIya, standartne vIdhilennya I koefItsIEnt varIatsIYi [Elektronniy dokument]. - (https://statanaliz.info/metody/opisanie-dannyx/11-dispersiya-standartnoe-otklonenie-koeffitsient-variatsii).
7. Oglyad metodIv poperednoYi obrobki danih [Elektronniy dokument]. - (http://www.math.spbu.ua/SD_AIS/documents/2019-12-341/2019-12-b-17.pdf).
8. Obrobka ob'Emnih danih, Gwyddion [Elektronniy dokument]. - (http://gwyddion.net/documentation/user-guide-ua/volume-data-processing.html).
9. Ob'EmnI danI [Elektronniy dokument]. – (http://www.teamnet.ua/gruppa-teamnet/issledovanie-i-razrobotka/opros-mnenij- obemnye-dannye/).

Received 03.03.2020.
Accepted 04.03.2020.

Исследование производительности методов нормализации объемных данных

Работа посвящена исследованию производительности методов нормализации объемных данных

Работа относится к области постобработки экспериментальных и статистических данных, заключается в преобразовании входного набора данных в выходной в конкретном интервале (нормализация).

В рамках работы были исследованы актуальные на сегодняшний день методы нормализации с целью их применения для нормализации числовых данных с сохранением соотношения. Была разработана библиотека реализует подошли этому критерию методы, позволяет нормализовать и визуализировать выходные данные.

Performance study of volume normalization methods

Data normalization is currently widely used in various fields of science and technology, and not only in the information technology environment. Medicine, geodesy, radio engineering, soil science and many other fields of knowledge use data normalization for more convenient presentation of data and their subsequent analysis.

But, as in any area, there are problems. One of these problems is the normalization of voluminous data. This is a potentially intermediate state between simple data and Big Data. In this case, it is already necessary to take into account the volume, but there is still no need to use Machine Learning solutions. An additional question is the problem of normalizing data types implemented according to the rules / in the context of OOP: object fields can also be voluminous.

The second problem that goes hand in hand with any issue related to normalization is the lack of a specialized library.

The problem of normalizing this type of data may be encountered, for example, in the field of medicine, when the results of laboratory tests need to be normalized to what area is

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

convenient for research and / or practical application, and there is a lot of data and they are large numerical values.

Also, the problem is relevant for areas of science and technology, where time-based statistics are applied. For example, there are collected statistics on the operation of an application in milliseconds for various periods.

To understand the dependencies in the operation of such an application, it will be necessary to normalize the statistics in one segment so that accurate conclusions can be drawn.

To solve such problems, the above issues will be considered in the framework of this work.

The work is devoted to the study of the performance of volume normalization methods.

The work relates to the field of post-processing of experimental and statistical data, consists in converting the input data set to the output in a specific interval (normalization).

In the framework of the work, current normalization methods were studied with the aim of their application to normalize numerical data while maintaining the ratio. A library has been developed that implements methods that meet this criterion, allows you to normalize and visualize the output.

Островская Екатерина Юрьевна - к.т.н., доцент кафедри информационных технологыий и систе, Нациаональныя металлургическая академия Украины.

Бедай Роман Вадимович- магистр кафедры информационных технологий и систем, Национальная металлургическая академия Украины.

Островська Катерина Юріївна – к.т.н., доцент кафедри Інформаційних технологій та систем, Національна металургійна академія України.

Бедай Роман Вадимович - магістр кафедри Інформаційних технологій та систем, Національна металургійна академія України.

Ostrovska Kateryna - candidate of technical sciences, associate professor of the Department of Information Technology and System, National Metallurgical Academy of Ukraine.

Beday Roman - Master of the Department of Information Technologies and Systems, National Metallurgical Academy of Ukraine.