DOI: 10.31319/2519-8106.2(43)2020.219259
УДК 519.614
Л.М. Божуха[2], к.фіз.-мат. наук, доцент, bozhukha.li@gmail.com
О.С. Косухіна[1], к.т.н., доцент, e_kos@ukr.net
О.В. Косухін[1], здобувач вищої освіти третього (доктор філософії) рівня**,** kosukhin24@gmail.com
Д.І. Божуха[1], здобувач вищої освіти другого (магістр) рівня,
[1]Дніпровський державний технічний університет, м. Кам'янське
[2]Дніпровський національний університет ім. Олеся Гончара, м. Дніпро

ПРО МЕТОДИ ЗНАХОДЖЕННЯ ВЛАСНИХ ЗНАЧЕНЬ SVD-РОЗКЛАДАННЯ МАТРИЦІ

Обрання алгоритмів аналізу даних на етапі зменшення розмірності даних з втратою мінімальної кількості інформації залежать від набору даних та подальшого використання зменшеного простору ознак в подальших алгоритмах (наприклад, при роботі з зображеннями та обробці текстових даних).

В роботі отримано результати використання точних, чисельних та ітераційних методів сингулярного розкладання прямокутних матриць та виконаний порівняльний аналіз роботи їх алгоритмів. Опрацьовування методів проводилося на зображеннях для отримання наочності, що не зменшує область їх використання для роботи з текстовими даними.

***Ключові слова**: svd-метод; сингулярні числа; власні вектори; алгоритм сингулярного розкладання.*

The choice of data analysis algorithms at the stage of the data reducing dimensionality with the loss of the information minimum amount depends on the data set and the subsequent use of reduced feature space in the subsequent algorithms (for example, when working with images and text processing).

The results of using exact, numerical and iterative methods of singular decomposition of rectangular matrices are obtained and a comparative analysis of their algorithms is performed. The methods were processed on images to obtain clarity, which does not reduce the scope of their use for working with textual data.

***Keywords**: svd-method; singular numbers; eigenvectors; singular decomposition algorithm.*

Постановка проблеми

Зменшення розмірності даних з втратою мінімальної кількості інформації базується на скороченні кількості випадкових змінних шляхом отримання множини головних змінних. Цю задачу можна розділити на декілька кроків. Першим кроком вважають обрання ознак, а другим — виділення вагомих ознак. Математичною основою щодо виділення ознак є перетворення багатовимірного простору в простір невеликої кількості вимірів. Для вирішення цієї задачі можна використати лінійне та нелінійне перетворення або деякі підходи тензорного числення.

Класичним прикладом лінійної техніки перетворення можна вважати метод головних компонент (principal component analysis, PCA), для реалізації якого можуть бути використані обчислення власних векторів і чисел коваріаційної матриці початкових даних або метод сингулярного розкладу матриць (singular-value decomposition, SVD). Існує великий спектр чисельних методів для обчислення власних значень та власних векторів.

Аналіз останніх досліджень та публікацій

При використанні методів власних [1] і сингулярних векторів [2] щодо аналізу даних зображень/текстів основним недоліком є використання значних обчислювальних ресурсів. Для роботи з зображеннями можна використати інструментарій модального аналізу.

Обрання алгоритмів аналізу даних на етапі зменшення розмірності даних з втратою мінімальної кількості інформації залежать від набору даних та подальшого використання зменшеного простору ознак в подальших алгоритмах.

Формулювання мети дослідження

В роботі ставиться задача отримання результатів використання чисельних та ітераційних методів сингулярного розкладання матриць та виконання порівняльного аналізу роботи цих алгоритмів.

Виклад основного матеріалу

Для роботи з даними потрібно розглянути дві основні задачі — задачу стиснення-відновлення зі збереженням достатньої якості і задачу ідентифікації, коли потрібно побудувати стислий образ з великим коефіцієнтом стиснення, і це відображення має бути взаємно-однозначним. Технологія стиснення-відновлення може бути заснована на сингулярному розкладанні матриці [4]. Для багатьох класів зображень сингулярні числа дуже швидко зменшуються. Можна залишити малу кількість сингулярних чисел, при цьому норма зміниться незначно і зміни зображення будуть майже непомітні. Цей феномен і дає можливість стиснення-відновлення (з втратами).

Чорно-біле зображення розмірності $n \times n$ пікселів можна розглядати як матрицю тієї ж розмірності, значеннями елементів якої служать числа, відповідні інтенсивності білого кольору для кожного пікселя. Тому, як і для будь-якої матриці, до зображення можна застосувати сингулярне розкладання (SVD-розкладання).

Розглянемо алгоритм SVD-стиснення на прикладі.

1. Нехай задано в форматі *.bmp зображення *A*. Зображенню *A* відповідає цілочисельна матриця *A* довільного розміру (512×512), (256×256) або ін. Здійснюємо сингулярне розкладання матриці, використовуючи чисельний апарат знаходження власних значень та векторів матриць $A^T A$ та $A A^T$ (для зручності була використана бібліотека Alglib).

2. Сингулярні числа швидко зменшуються, і починаючи з деякого номера стають досить малими, а, отже, значення, що залишаються, можна не враховувати при відновленні вихідного зображення. Таким чином, замість матриці *A* розмірності $n \times n$ потрібно зберігати матриці розмірності $n \times k$, $k \times n$ і рядок з k чисел. Тобто отримуємо стиснення з коефіцієнтом

$$k = \frac{n^2}{k(2n+1)}.$$

Алгоритм SVD-стиснення та знаходження власних значень та власних векторів у розробленому програмному забезпеченні реалізується за допомогою підпрограм. Зокрема, виділимо такі:

1) Підпрограма для знаходження власних значень (і власних векторів) симетричної матриці:

Вхідними параметрами є: *A* — симетрична матриця, яка задається його верхньою або нижньою трикутною частиною; масив [*0..N-1, 0..N-1*], де *N* — розмір матриці *A*; ZNeeded — прапор, який контролює, чи потрібні власні вектори чи ні (якщо ZNeeded дорівнює 0, то власні вектори не повертаються, а якщо 1, то власні вектори повертаються); IsUpperA — формат зберігання матриці *A*; *B1*, *B2* — границі напіввідкритого інтервалу (*B1*, *B2*) для пошуку власних значень.

Вихідними параметрами є: *M* — число власних значень ($M \geq 0$); *W* — масив знайдених власних значень; масив, індекс якого змінюються в інтервалі [*0..M-1*]; прапор *Z* — якщо ZNeeded дорівнює 0, то значення *Z* не змінюється, а якщо, то *Z* буде містити власні вектори; масив, індекси якого змінюються в інтервалах [*0..N-1*, *0..M-1*]. При цьому власні вектори зберігаються в стовпцях матриці.

Результатом роботи підпрограми є:

- значення True, якщо робота успішна; *M* містить число власних значень у даному напівінтервалі (може бути дорівнює 0); *W* містить власні значення, *Z* містить власні вектори (якщо необхідно);

- False, якщо підпрограма методу бісекції не змогла знайти власні значення в заданому інтервалі або якщо підпрограма зворотної ітерації не змогла знайти всі відповідні власні вектори. У цьому випадку власні значення і власні вектори не повертаються, *M* дорівнює 0.

2) Підпрограма сингулярного розкладання матриці:

Вхідними параметрами є: *A* — симетрична матриця, яка задається її верхньою або нижньою трикутною частиною, [*0..N-1*, *0..N-1*], де *N* — розмір матриці *A*; ZNeeded — прапор, який контролює, чи потрібні власні вектори чи ні, якщо ZNeeded дорівнює 0, то власні вектори не повертаються, а якщо 1, то власні вектори повертаються; IsUpperA — формат зберігання матриці *A*; *B1*, *B2* — межі напіввідкритого інтервалу (*B1*, *B2*) для пошуку власних значень у.

Вихідними параметрами є: *M* — число власних значень ($M \geq 0$); *W* — масив знайдених власних значень; масив, індекс якого змінюються в інтервалі [*0..M-1*]; *Z* — якщо ZNeeded дорівнює 0, то *Z* не змінюється, а якщо 1, то *Z* містить власні вектори; масив, індекси якого змінюються в інтервалах [*0..N-1*, *0..M-1*]. Власні вектори зберігаються в стовпцях матриці.

Результатом роботи підпрограми є:

- True, якщо операція успішна, *M* містить кількість власних значень (може бути 0), *W* містить власні значення, *Z* містить власні вектори (якщо необхідно).

- False, якщо підпрограма методу бісекції не змогла знайти власні значення в заданому інтервалі або якщо зворотній ітерації підпрограми не вдалося знайти всі відповідні власні вектори. У цьому випадку власні значення і власні вектори не повертаються, *M* дорівнює 0.

В роботі розроблено програмний продукт у програмному середовищі Microsoft Visual Studio на мові програмування C#. При обранні зображення запускається метод сингулярного розкладання матриці яскравості та виконується перевірка правильності розкладу при збереженні даних трьох матриць розміру 400*400. Головне вікно програми представлено на рис. 1.

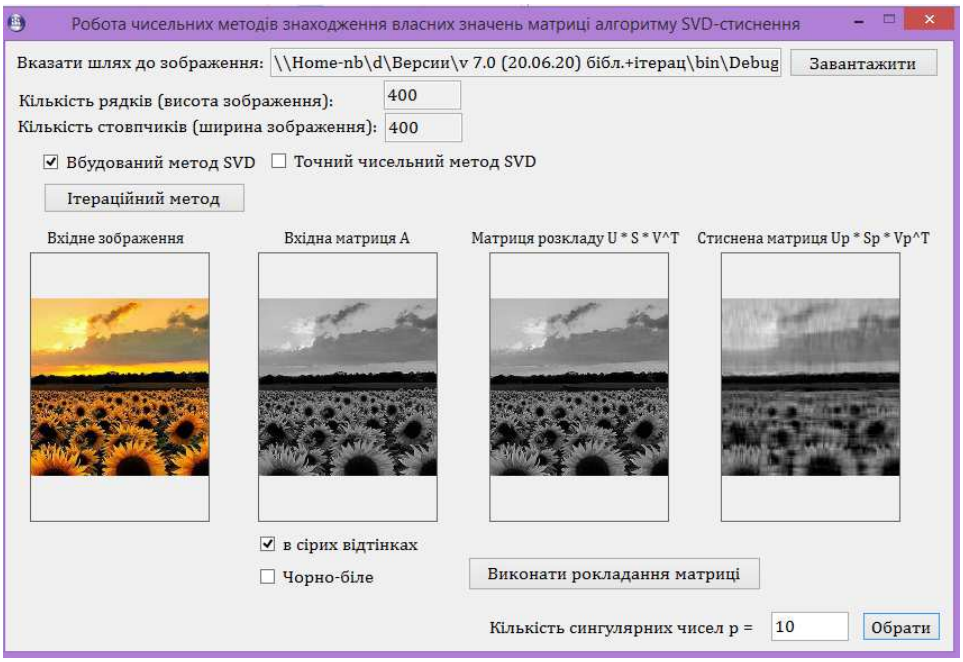

Рис. 1. Головне вікно програми

Для показу роботи методу SVD можна обрати різну кількість сингулярних значень та побудувати відповідні апроксимаційні матриці зображень.

Результати роботи точного чисельного методу можна побачити на рис. 1 у формі «Матриця розкладу U*S*V^T». На цьому етапі роботи для обчислення власних векторів сингулярного розкладання використаний підхід, який представлений у роботі [1]. Знаходження власних значень виконано на основі методу Левер‘є.

Дуже часто при розв‘язанні прикладної задачі необхідно знайти не обов‘язково всі власні значення. Іноді достатньо знайти тільки найбільше власне значення або набір власних значень, які розташовані в порядку спадання. Для розв’язання часткової задачі знаходження власних

значень ефективними є ітераційні методи. Довільний ітераційний процес знаходження найбільшого власного значення λ_1 базується на розв'язанні системи рівнянь $AX = \lambda_1 X$. При повторенні декілька разів операції множення AX права частина рівняння буде збільшуватися на максимальне по модулю число λ_i. Цей підхід ітераційного методу реалізований у представленому проекті аналізу методів. В якості додаткових ітераційних методів знаходження власних значень і векторів обиралися покомпонентний метод, метод скалярних добутків, метод Левер'є, метод Фадєєва та метод обертання Якобі для симетричних матриць.

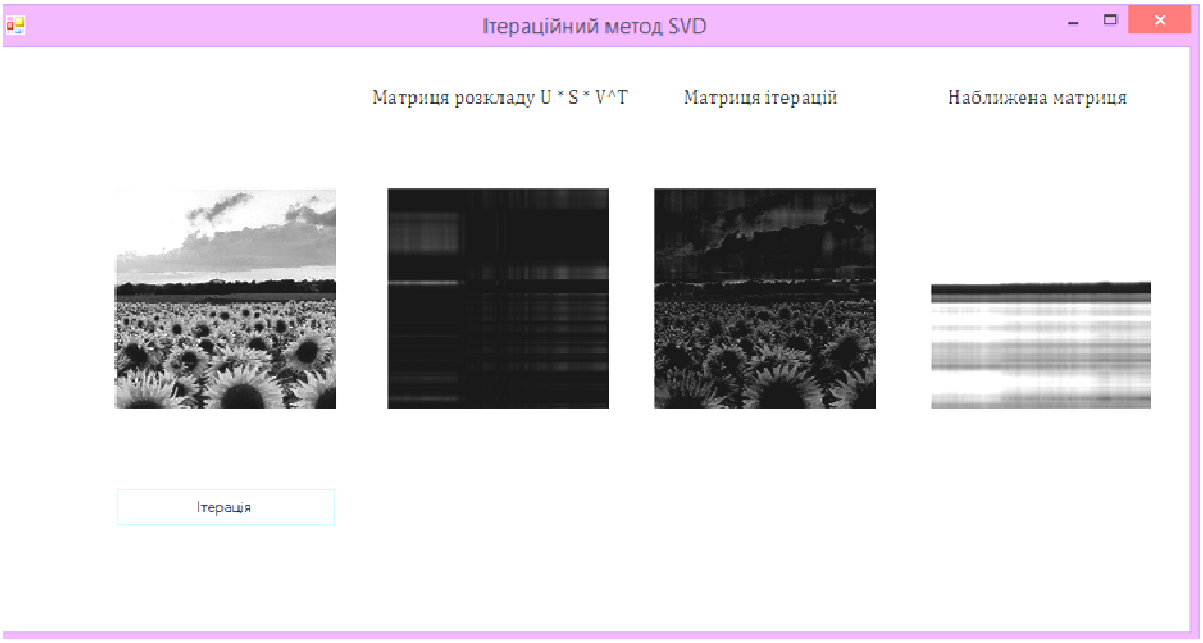

Рис. 2. Результати роботи ітераційного алгоритму (2 ітерації)

Виділяють ітераційний алгоритм знаходження власних чисел та матриць сингулярного розкладання SVD. Ставиться задача послідовного знаходження векторів U_k, V_k та сингулярних чисел λ_k $(k = \overline{1,r}, r \geq min(m,n))$, r — ранг матриці A розміру $m \times n$. Ітераційний процес проходить за відповідною схемою. Робота представленого ітераційного алгоритму представлена на рис.2. Користувачу надається можливість слідкувати за ітераційним процесом побудови матриці зображення з переглядом результатів матриці $A = V \Sigma W$, матриці ітерацій та наближеної матриці.

Висновки

Використання технології стиснення-відновлення, яка заснована на сингулярному розкладанні матриці, надає інструмент роботи з багатьма класами даних при швидкому зменшенні відсортованої множини сингулярних чисел.

У даній роботі отримано результати використання ітераційних методів сингулярного розкладання прямокутних матриць та виконаний порівняльний аналіз роботи їх алгоритмів. Отримані результати використання ітераційних методів сингулярного розкладання матриць при роботі з даними можуть бути використані для роботи з текстовими даними (корпусами текстів). Для обробки великого обсягу текстових даних (наприклад, тематичне моделювання) залишається проблема розрідженості матриць та коректного використання метрик простору.

Використання точних чисельних методів для знаходження власних векторів та власних значень у сукупності суттєво ускладнює процес обчислення за рахунок високої кількості обчислень при врахуванні великих розмірів вхідних матриць даних. Складність цього підходу полягає у використанні чисельних методів обчислення визначників, розв'язання систем лінійних рівнянь, знаходження коренів трансцендентних рівнянь.

Список використаної літератури

1. Характеристический полином, собственные числа, собственные векторы матрицы [Електронний ресурс]. URL: http://pmpu.ru/vf4/algebra2/charpoly
2. Сингулярное разложение матрицы [Електронний ресурс]. URL: https://www-cloudfront-alias.coursera.org/ learn/vvedeniye-v-nauku-o-dannykh
3. Г.Т. Олійник, Т.В. Савельєва, О.М. Пригодюк Розв'язання фахових задач із застосуванням ПЕОМ: посіб. з інформатики і системології для студентів напрямів підготовки 6.040106 – екологія, охорона навколишнього середовища та збалансоване природокористування (екологія та охорона навколишнього середовища), 6.051301 – хімічна технологія (хімічна технологія неорганічних речовин), 6.051701 – харчові технології та інженерія (технології продуктів бродіння і виноробства), 6.060101 – будівництво (промислове та цивільне будівництво). Черкаси: ЧДТУ, 2011. 180 с.
4. В.Г. Лежнев, А.Н. Марковский Математические алгоритмы сжатия изображений: учебное пособ. Краснодар: КГУ, 2015. 55 с.

ABOUT METHODS OF FINDING OWN VALUES FOR SVD-DECOMPOSITION MATRIX

Bozhukha L., Kosukhina E., Kosukhin A., Bozhukha D.

The choice of data analysis algorithms at the stage of the data reducing dimensionality with the loss of the information minimum amount depends on the data set and the subsequent use of reduced feature space in the subsequent algorithms (for example, when working with images and text processing).

To work with the data, two main tasks are considered — the task of compression-recovery with sufficient quality and the problem of identification, when you want to build a compressed image with a high compression ratio, and this mapping should be mutually unique. Compression-recovery technology can be based on the singular decomposition of the matrix [4]. For many image classes, singular numbers decrease very rapidly. You can leave a small number of singular numbers, while the rate will change slightly and the changes in the image will be almost invisible. This phenomenon allows compression-recovery (with losses).

The results of using exact, numerical and iterative methods of singular decomposition of rectangular matrices are obtained and a comparative analysis of their algorithms is performed. The methods were processed on images to obtain clarity, which does not reduce the scope of their use for working with textual data.

The use of compression-recovery technology, which is based on the singular decomposition of the matrix, provides a tool for working with many classes of data while rapidly reducing the sorted set of singular numbers.

In this paper, the results of using iterative methods of singular decomposition of rectangular matrices are obtained and a comparative analysis of their algorithms is performed. The obtained results of using iterative methods of singular decomposition of matrices when working with data can be used to work with text data (text corpora). For the processing of large amounts of text data (for example, thematic modeling) there is a problem of sparse matrices and the correct use of space metrics.

The use of accurate numerical methods to find eigenvectors and eigenvalues in the aggregate significantly complicates the calculation process due to the high number of calculations, taking into account the large size of the input data matrices. The complexity of this approach lies in the use of numerical methods for calculating determinants, solving systems of linear equations, finding the roots of transcendental equations.

References

[1] *Harakteristicheskiy polinom, sobstvennyie chisla, sobstvennyie vektoryi matritsyi* [Characteristic polynomial, eigenvalues, eigenvectors of a matrix] Retrieved from URL: http://pmpu.ru/vf4/algebra2/charpoly

[2] *Singulyarnoe razlozhenie matritsyi* [*Singular value decomposition of a matrix*] Retrieved from URL: https://www-cloudfront- alias.coursera.org/ learn/vvedeniye-v-nauku-o-dannykh

[3] G.T. Oliynyk, TV Savelyeva, OM Prigodyuk (2011) Rozv'yazannya fahovih zadach Iz zastosuvannyam PEOM [Solving professional problems using a PC] (manual on computer science and systemology). Cherkasy: ChTTU

[4] Lezhnev V.G., Markovsky A.N. (2015), Matematicheskie algoritmyi szhatiya izobrazheniy [*Mathematical algorithms for image compression*]: (uchebnoe posob.). Krasnodar: KGU (in Russia)