УДК 519.254:519.237.8

О. М. Мацуга, В. С. Шеремет
Дніпровський національний університет імені Олеся Гончара

КЛАСТЕРИЗАЦІЯ ДАНИХ З ПРОПУСКАМИ МЕТОДОМ K-СЕРЕДНІХ

Наведено огляд існуючих підходів до кластеризації даних з пропусками. Проведено обчислювальні експерименти для їх порівняння у разі застосування методу *k*-середніх. Для порівняння обрано методи заповнення пропусків середнім, медіаною і на основі методу головних компонент, а також методи кластеризації даних з пропусками KSC і *k*-POD. Одержані результати показали перевагу методу кластеризації KSC та заповнення пропусків на основі методу головних компонент.

Ключові слова: *кластеризація, метод k-середніх, пропуски, заповнення пропусків, зовнішні метрики якості, обчислювальний експеримент.*

Приведен обзор существующих подходов к кластеризации данных с пропусками. Проведены вычислительные эксперименты для их сравнения в случае применения метода *k*-средних. Для сравнения выбраны методы заполнения пропусков средним, медианой и на основе метода главных компонент, а также методы кластеризации данных с пропусками KSC и *k*-POD. Полученные результаты показали преимущества метода KSC и заполнения пропусков на основе метода главных компонент.

Ключевые слова: *кластеризация, метод k-средних, пропуски, заполнение пропусков, внешние метрики качества, вычислительный эксперимент.*

The clustering task is a very important data mining task arising in many applications. However, well known and widely used clustering methods cannot work with datasets that have missing values. The paper provides an overview of the existing approaches to data with missing values clustering and a comparative analysis of chosen approaches when applying the *k*-means clustering method. Commonly used approaches are: 1) deletion entities or features with missing values with further complete data clustering, 2) imputation (filling in the missing values) with further complete data clustering, 3) using clustering methods that work with missing data. The following methods within these approaches were chosen for comparison: mean substitution, median substitution, imputation based on principal component analysis and two *k*-means extensions – KSC and *k*-POD. In order to conduct the comparative analysis the software «ClustDMV» was developed on the C# using the .NET Core framework. The software is a desktop application with a

graphical interface.The following computational experiments were carried out using the created software. One hundred two-dimensional datasets were simulated in each experiment. All the datasets in one experiment were modeled with the same clusters parameters. Each dataset consists of 800 entities from four clusters. Then a certain percentage of omissions were randomly entered into the data. After that, either the missing values were filled in and *k*-means method was used with the complete data, or the KSC and *k*-POD methods were applied to the data with missing values. In order to evaluate the clustering quality Rand index was used. Indexes obtained on all 100 datasets were averaged.The paper presents the results of three typical experiments. According to the results the KSC method and imputation based on principal component analysis showed the best performance.

Keywords: *clustering, k-means method, missing values, impute missing values, external evaluation measures, computational experiment.*

Постановка проблеми. Задача кластерного аналізу полягає у розбитті заданого набору об'єктів на підмножини, що називаються кластерами, так, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися.

На практиці часто виникає необхідність розв'язувати цю задачу на даних з пропусками, тобто коли у деяких об'єктів відсутні значення деяких показників. Це становить проблему, оскільки усі поширені й загальновідомі методи кластеризації (*k*-середніх, ієрархічні, EM тощо) працюють лише на даних без пропусків. Для подолання цієї проблеми запропоновано декілька різних підходів, але актуальним досі залишається питання вибору найкращого з них. Тому в рамках даної роботи було поставлено за мету проаналізувати існуючі підходи до кластеризації даних з пропусками і провести їх порівняльний аналіз у разі застосування методу *k*-середніх. Останній обраний через те, що є одним з найвідоміших і найпоширеніших методів кластеризації [1].

Аналіз останніх досліджень і публікацій. Можна виділити три загальні підходи до кластеризації даних з пропусками:

1. Видалення з набору даних об'єктів або ознак, у яких є пропущені значення, і подальша кластеризація даних без пропусків.

2. Заповнення пропусків і подальша кластеризація повних даних.

3. Використання методів кластеризації, що враховують наявність пропусків у даних.

Недолік першого підходу полягає в тому, що він призводить до скорочення набору даних та неможливості кластеризувати об'єкти з пропущеними значеннями. Проте цей підхід простий в реалізації і може бути рекомендований, коли кількість пропусків дуже мала.

Другий підхід є найбільш розповсюдженим на практиці. В його

рамках запропоновано значну кількість методів заповнення пропусків, серед яких можна виділити такі [2–5]:

- Заповнення пропусків середнім значенням відповідного показника (mean substitution). Замість середнього інколи використовують медіану, а якщо показник з пропусками якісний, тоді відсутні значення заповнюють модою.
- Індикаторний метод (indicator method) – полягає у заповненні пропусків спеціальним значенням (як правило, нулем) та введенням додаткового бінарного показника, який набуває нульових значень на тих об'єктах, що не мають пропусків, і ненульових – на об'єктах, у яких є пропуски.
- Метод найближчого сусіда – для об'єкта, у якого відсутні значення деяких показників, шукають найближчий, у якого ці показники відомі, і заповнюють ними пропуски. Але необхідно, щоб у наборі даних була достатня кількість об'єктів без пропусків.
- Регресійний метод – пропущене значення конкретного показника заповнюють, прогнозуючи його на основі інших показників за допомогою моделі регресії. Тут може бути використана класична лінійна модель регресії, випадковий ліс регресійних дерев або інша модель. Очевидно, що метод можна застосовувати лише якщо показник з пропусками корелює з іншими показниками набору і є кількісним. Для заповнення пропусків у якісних показниках замість моделей регресії використовують класифікатори.
- Метод на основі сингулярного розкладання (singular value decomposition – SVD) – знаходять сингулярне розкладання матриці вхідних даних, тоді пропуски заповнюють на основі матриці, одержаної із сингулярного розкладання.
- Алгоритм ZET – запропонований Загоруйко Н.Г. і базується на припущеннях щодо надмірності даних у наборі, локальної компактності та лінійної залежності [6].
- Повторення результату попереднього спостереження (last observation carried forward – LOCF) – цей метод доцільно застосовувати у разі кластеризації часових рядів, у яких наступні спостереження, як правило, залежать від попередніх.

Також варто вказати метод множинного заповнення пропусків (multiple imputation for missing data), запропонований у 1977 році Рубіном [3; 5]. Цей метод включає в себе ідею багатокроковості й є у своїй реалізації важким, бо описується абстрактно й потребує конкретизації залежно від задачі, тому, незважаючи на численні рекомендації, так і не набув великої популярності.

В рамках третього підходу до кластеризації даних з пропусками, зазвичай, розробляють модифікації відомих методів. Наприклад, на базі методу *k*-середніх, що є предметом розгляду в даній роботі, розроблено методи KSC (*k*-means with Soft Constraints) [7] та *k*-POD [8]. Особливістю методу KSC є те, що він вимагає наявності хоча б одного показника, який не містить жодного пропущеного значення.

Слід зазначити, що ефективність розглянутих підходів та методів роботи з пропусками значною мірою залежить від механізму породження пропусків. Залежно від механізму, що призвів до появи пропусків, виділяють такі їх типи [2–5]:

- Повністю випадкові (missing completely at random, MCAR) – ймовірність появи таких пропусків не залежить ні від значень показника, в якому з'явилися пропущені значення, ні від значень інших показників. Наприклад, через втрату зразків крові окремого пацієнта для нього не будуть визначені деякі показники крові.
- Частково випадкові (missing at random, MAR) – ймовірність появи таких пропусків не залежить від значень спостережуваного показника, але обумовлена значеннями інших показників. Наприклад, пацієнти з певним діагнозом можуть мати протипоказання до деяких обстежень, тому у таких пацієнтів будуть відсутні значення показників цих обстежень.
- Невипадкові (missing not at random, MNAR) – ймовірність їх появи залежить від значень самого показника. Наприклад, значення показника може не потрапляти в діапазон чутливості вимірювального приладу, внаслідок чого буде відсутнє.

Вищерозглянуті підходи та методи орієнтовані, головним чином, на пропуски типу MCAR. Хоча, наприклад, автори методу *k*-POD зазначають, що результати роботи їх методу майже не залежать від типу пропусків [8].

Постановка задачі. Необхідно провести порівняльний аналіз різних підходів до кластеризації даних з пропусками типу MCAR методом *k*-середніх. Припускається, що набір даних містить лише кількісні показники.

Основний матеріал. Для порівняння у роботі було обрано методи, які можна вважати найбільш універсальними і застосовувати до довільних наборів даних з кількісними показниками:

- Заповнення пропусків середнім та медіаною.
- Заповнення пропусків на основі методу головних компонент (МГК). Він являє собою варіацію методу на основі SVD. Будують коваріаційну матрицю вхідних даних з ігноруванням пропусків.

Обчислюють власні числа і вектори цієї матриці. В МГК на їх основі будують головні компоненти, але в роботі знання власних чисел та векторів дозволяє одержати сингулярне розкладання початкової матриці даних, відновити її і тим самим заповнити пропуски.

- Використання модифікацій методу *k*-середніх, що дозволяють кластеризувати дані з пропусками. Це методи KSC та *k*-POD.

З метою проведення їх порівняльного аналізу було розроблено програмний продукт «ClustDMV». У процесі його розробки використано такі технології:

- Середовище розробки – Visual Studio.
- Мова програмування – C# (.NET Core).
- Додаткові бібліотеки – Accord.Math, Accord.Statistics, Algomera, CenterSpace. Перші дві були використані у процесі реалізації EM алгоритму, а дві останні – для побудови дендрограми в ієрархічних методах.

Програмний продукт «ClustDMV» має такий функціонал:

- Завантаження або генерація даних. Передбачено генерацію повних даних із заздалегідь відомим розбиттям по кластерах, а також генерацію даних з пропусками на їх основі.
- Заповнення пропусків середнім або медіаною відповідної ознаки, а також на основі МГК.
- Кластеризація даних з пропущеними значеннями методами KSC та *k*-POD, що є модифікаціями методу *k*-середніх.
- Кластеризація даних без пропусків методом *k*-середніх (алгоритм Ллойда). Також реалізовано кластеризацію ієрархічними методами та EM алгоритмом.
- Оцінювання якості роботи різних підходів до кластеризації даних з пропусками шляхом порівняння результатів кластеризації із заздалегідь відомими на основі таких індексів: Rand, Jaccard, Fowlkes-Mallows, Sorensen-Dice (усі індекси набувають значень від 0 до 1). Результати порівняння відображаються у вигляді наочних діаграм.

Головне вікно застосунку наведено на рис. 1.

За допомогою створеного програмного забезпечення було проведено ряд обчислювальних експериментів на змодельованих наборах даних. У ході кожного експерименту моделювалось 100 двовимірних наборів даних згідно суміші нормальних розподілів. Усі 100 наборів моделювалися з однаковими параметрами кластерів. Кожен набір містив 800 об’єктів, розділених на 4 сферичні кластери (по 200 об’єктів у кожному кластері).

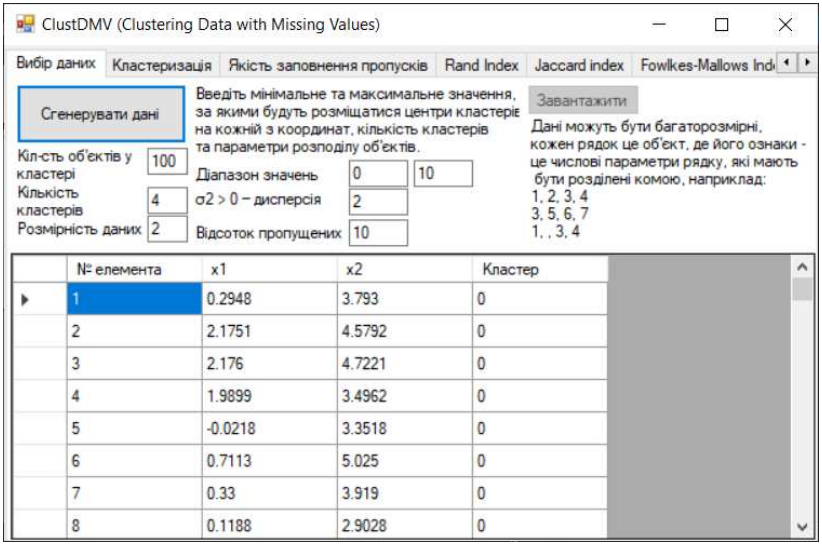

Рисунок 1 – Головне вікно розробленої програми «ClustDMV»

У кожному експерименті на кожному зі 100 наборів даних виконувалося таке. Випадковим чином вносився заданий відсоток пропусків. Тоді пропуски або заповнювалися (середнім, медіаною, на основі МГК) і проводилася кластеризація повного набору методом *k*-середніх, або дані з пропусками кластеризувалися методом KSC (з $w = 0.5$) чи *k*-POD. Щоб оцінити якість кластеризації, одержане розбиття порівнювалося із відомим на основі Rand Index. Вищі значення індексу свідчили про кращу якість кластеризації і були більш бажані. Значення індексу, одержані для усіх 100 наборів, усереднювалися.

Нижче наведено і проаналізовано результати трьох експериментів. Під час першого експерименту моделювались набори даних, в яких кластери були сильно відокремлені один від одного. У ході другого експерименту розглядалися набори з кластерами, що також не перетиналися, але були розташовані більш близько. У третьому експерименті було розглянуто випадок, коли кластери перетиналися. В усіх експериментах для кожного кластеру дисперсійно-коваріаційну матрицю було задано як одиничну матрицю. Центри кластерів для було задано такими:

- Експеримент 1: (5, 22), (18, 23), (18.5, 11), (35, 29).

- Експеримент 2: (0, 10), (4, 10), (4, 4), (7, 6).
- Експеримент 3: (0, 4), (3, 5), (4, 7), (6, 3).

Усередненні значення Rand Index для кожного з трьох експериментів наведено нижче (табл. 1–3). Чим вищі значення індексу, тим якіснішою є кластеризація даних. Одержані результати дозволяють зробити такі висновки:

- Якщо кластери дуже добре відокремлені один від одного, то найкращу якість кластеризації забезпечує метод KSC майже при будь-кому відсотку пропущених значень. Усі методи заповнення пропусків у такому випадку значно спотворюють дані і, як наслідок, забезпечують гіршу якість кластеризації.
- Якщо кластери досить близько розташовані або взагалі перетинаються, то найкращу якість кластеризації поряд з методом KSC показує заповнення пропусків на основі МГК.

Таблиця 1

Усереднене значення Rand Index ± його середньоквадратичне відхилення під час експерименту 1

Відсоток пропусків, %	**Заповнення пропусків середнім**	**Заповнення пропусків медіаною**	**Заповнення пропусків на основі МГК**	**Метод KSC**	**Метод *k*-POD**
10	0,708 ±0,00884	0,849 ±0,00236	**0,989 ±0,0013**	0,944 ±0,00118	0,837 ±0,00255
20	0,657 ±0,00985	0,721 ±0,00289	0,852 ±0,00149	**0,869 ±0,00046**	0,728 ±0,00059
30	0,638 ±0,00132	0,665 ±0,0012	0,722 ±0,00111	**0,820 ±0,0002**	0,683 ±0,0002
40	0,622 ±0,00027	0,640 ±0,00147	0,654 ±0,00292	**0,769 ±0,00227**	0,622 ±3,3E-05
50	0,615 ±0,00230	0,622 ±0,00042	0,646 ±0,00059	**0,706 ±0,00054**	0,560 ±0,00052
60	0,612 ±0,00128	0,602 ±0,0023	**0,647 ±0,00221**	0,639 ±0,00188	0,500 ±0,01261

Таблиця 2

Усереднене значення Rand Index ± його середньоквадратичне відхилення під час експерименту 2

Відсоток пропусків, %	Заповнення пропусків середнім	Заповнення пропусків медіаною	Заповнення пропусків на основі МГК	Метод KSC	Метод *k*-POD
10	0,783 ±0,00565	0,784 ±0,01586	0,878 ±0,01005	**0,894 ±0,00089**	0,845 ±0,00042
20	0,701 ±0,00196	0,689 ±0,00643	**0,841 ±0,00552**	0,809 ±0,00016	0,743 ±0,00078
30	0,679 ±0,00048	0,661 ±0,00333	**0,793 ±0,00098**	0,761 ±0,00135	0,693 ±0,00075
40	0,635 ±0,00215	0,635 ±0,00098	**0,724 ±0,00063**	0,710 ±0,0015	0,649 ±0,00157
50	0,606 ±0,00156	0,603 ±0,00154	**0,674 ±0,0043**	0,652 ±0,0015	0,596 ±0,00154
60	0,568 ±0,00568	0,573 ±0,00518	**0,650 ±0,00038**	0,586 ±0,00058	0,543 ±0,00386

Таблиця 3

Усереднене значення Rand Index ± його середньоквадратичне відхилення під час експерименту 3

Відсоток пропусків, %	Заповнення пропусків середнім	Заповнення пропусків медіаною	Заповнення пропусків на основі МГК	Метод KSC	Метод *k*-POD
10	0,747 ±0,00072	0,771 ±0,00357	**0,842 ±0,02166**	0,825 ±0,00136	0,786 ±0,00258
20	0,685 ±0,00216	0,677 ±0,00519	**0,778 ±0,00314**	0,761 ±0,00169	0,695 ±0,00233
30	0,656 ±0,00035	0,652 ±0,00131	**0,730 ±0,00197**	0,724 ±0,00056	0,647 ±0,00046
40	0,619 ±0,00283	0,626 ±0,00112	0,681 ±0,0064	**0,688 ±0,00192**	0,595 ±0,00108
50	0,581 ±0,0012	0,604 ±0,00208	0,634 ±0,00418	**0,645 ±0,00093**	0,545 ±0,00266
60	0,546 ±0,00789	0,578 ±0,00183	**0,620 ±0,01022**	0,592 ±0,00081	0,494 ±0,00104

Висновки. Проведено огляд існуючих підходів до кластеризації даних з пропусками і здійснено порівняльний аналіз декількох підходів у разі застосування методу *k*-середніх. Для порівняльного аналізу обрано методи заповнення пропусків середнім, медіаною і на основі методу головних компонент, а також методи кластеризації даних з пропусками KSC і *k*-POD (останні базуються на методі *k*-середніх). Розроблено програмний продукт «ClustDMV», який забезпечує проведення кластеризації даних з пропусками та порівняльного аналізу обраних методів. За допомогою розробленого програмного продукту проведено обчислювальні експерименти, результати яких показали, що найкращу якість кластеризації даних з пропусками забезпечує метод кластеризації KSC та заповнення пропусків на основі методу головних компонент.

Бібліографічні посилання

1. Steinley D. K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*. 2006. Vol. 59. P. 1–34.

2. Глушко О. Обработка пропусков в данных – часть 1. URL: https://basegroup.ru/community/articles/missing (дата звернення: 14.11.2019).

3. Литтл Р. Дж. А., Рубин Д. Б. Статистический анализ данных с пропусками. Москва: Финансы и статистика, 1990. 336 с.

4. Guan N.C., Yusoff M.S.B. Missing values in data analysis: Ignore or Impute? *Education in Medicine Journal*. 2011. Vol. 3 (1). P. e6–e11. DOI:10.5959/eimj.3.1.2011.or1

5. Pigott T.D. A Review of Methods for Missing Data. *Educational Research and Evaluation*. 2001. Vol. 7. No. 4. P. 353–383. DOI:10.1076/edre.7.4.353.8937

6. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999. 270 с.

7. Wagstaff K. Clustering with Missing Values: No Imputation Required. *Proceedings of the Meeting of the International Federation of Classification Societies*. 2004. P. 649–658. DOI:10.1007/978-3-642-17103-1_61

8. Chi J.T., Chi E.C., Baraniuk R.G. *k*-POD. A Method for *k*-Means Clustering of Missing Data. *The American Statistician*. 2016. Vol. 70. Issue 1. P. 91–99. DOI: 10.1080/00031305.2015.1086685

Надійшла до редколегії 15.11.2019.