УДК 004.043

ОЦІНКА МЕТОДІВ КЛАСТЕРИЗАЦІЇ РІЗНОТИПОВИХ ДАНИХ
EVALUATION OF METHODS OF CLUSTERIZATION OF DIFFERENT TYPES OF DATA

[1]Бойко Н.І., [2]Ткачик О.А.
[1]Boyko N.I., [2]Tkachyk O.A.

[1,2]Національний університет "Львівська політехніка", Львів, Україна
ORCID: [1]orcid.org/0000-0002-6962-9363, [2]orcid.org/ 0000-0002-0728-4208
E-mail: [1]Nataliya.i.boyko@lpnu.ua, [2]oleksandr.a.tkachyk@lpnu.ua

DOI: 10.15673/atbp.v15i1.2508

Анотація. *Дослідницька робота вивчає взаємозалежність успішності роботи груп навчених суб'єктів від власних характеристик кожного члена групи. Описана можливість прогнозування співпраці окремих спеціалістів у команді в довгостроковій перспективі на основі методів машинного навчання, до яких належать різні моделі кластеризації й відповідні методи кластеризаційного аналізу. Окрім того, поданий алгоритм здобуття аналітичних даних для подальшого їх використання у можливій реалізації сервісу для підбору й групування персоналу та подальших досліджень. У роботі розглянуто декілька основних методів кластеризаційного аналізу. Визначена мета дослідження - оцінка методів та вибір найбільш кращого для дослідження і подальшої кластеризації ознак працівників офісних компаній. Було розглянуто переваги та недоліки основних алгоритмів з урахуванням потреб дослідження. У цьому дослідженні розглянуто потенційні джерела даних, процеси їх обробки та кластеризації обраними алгоритмами. В роботі практично перевірено відповідність обраного методу кластерного аналізу. Для аналізу був обраний FOREL алгоритм, який відповідає вимогам роботи на контрольній групі даних, зручній для наглядного представлення процесу кластеризації. Були оцінені особливості використання різних методів з різною кількістю ознак та обрано метод (k-середніх) для подальшого порівняння з основним підходом (FOREL). Досліджено можливість та доведено використання методів машинного навчання (FOREL, k-means зокрема) для полегшення процесу потреб найму та розподіленню працівників. У роботі був зроблений висновок, що обраний підхід годиться для обширного погляду на формування команд, а не на підгін всіх працівників під команду певного розміру (хоча це теж є можливим). Використання даного алгоритму може бути корисним і при доукомплектуванні команд. Для цього необхідно лиш розширити вибірку досліджуваних даних вже найманими працівниками зі своїми командами і оцінити, в якій команді "приживеться" потенційний командний гравець.*

Annotation. *Research work studies the interdependence of the success of groups of trained subjects on the individual characteristics of each group member. The possibility of predicting the cooperation of personal specialists in a team in the long term based on machine learning methods, including various clustering models and corresponding clustering analysis methods, is described. In addition, the algorithm for obtaining analytical data is presented for their further use in the possible implementation of the service for selecting and grouping personnel and other research. Several basic methods of clustering analysis are considered in work. The determined purpose of the study is to evaluate the processes and choose the best one for research and further clustering of the characteristics of employees of office companies. The advantages and disadvantages of the main algorithms were considered, taking into account the needs of the research. This study examines potential sources of data, processes of their processing and clustering by selected algorithms. In the work, the appropriateness of the chosen cluster analysis method was practically verified. The FOREL algorithm was selected for the analysis, which meets the requirements of working on a control group of data, convenient for a visual representation of the clustering process. The peculiarities of using different methods with different features were evaluated, and the (k-means) method was chosen for further comparison with the direct approach (FOREL). The possibility of using machine learning methods (FOREL, k-means in particular) to facilitate hiring needs and employee distribution has been investigated and proven. The work concluded that the chosen approach is suitable for a comprehensive view of team formation and not for fitting all employees to a team of a specific size (although this is also possible). The use of this algorithm can also be helpful when adding units. To do this, it is only necessary to expand the sample of researched data by already hired employees with their teams and evaluate which team a potential player will*

"take root".

Ключові слова: машинне навчання, кластерний аналіз, алгоритм, метод, FOREL алгоритм, k-means
Keywords: machine learning, cluster analysis, algorithm, method, FOREL algorithm, k-means

Вступ

Процес підбору спеціалістів завжди є складною задачею. Підібрати правильну людину, об'єднати співробітників у команди, які будуть приносити результат, відсіяти потенційно неефективних співробітників - все це потребує неабияких вмінь спеціально навчених людей. Питання підбору особливо загострилася у наші дні. Пандемія коронавірусу COVID-19, початок російсько-української війни 2022 р. суттєво вплинули на процес найму спеціалістів, в Україні зокрема. Брак живого спілкування, причини для постійного стресу, релокація, зміна життєвих обставин і способу життя вносять свої від'ємні корективи в можливість успішного підбору персоналу, навіть для досвідчених кадровиків.

Машинне навчання та його методи створені для полегшення життя людей, їх обов'язків та процесів. Вирішенням проблеми (хоча б частковим) тут вбачається використання методів кластерного аналізу для групування "кластеризації" офісних працівників на основі даних про ознаки кожного з них [8, 12]. Маючи датасет з характеристиками кожного з потенційних членів майбутньої команди, в теорії, методами машинного навчання можна добитись на виході комбінацію максимально ефективної команди. Команди, де кожен тіммейт буде другом, опорою так колегою для іншого. Цього вдалося б досягти на основі раніше з'ясованих даних про ознаки кожного співробітника, де методами кластеризації виявлено групи "кластери" осіб, які найкраще зможуть співпрацювати.

Аналіз літературних джерел

Співробітник є ключовим елементом організації. Успіх чи невдача компанії залежить від продуктивності співробітників. На основі дослідження "Employee's performance analysis and prediction using K-means clustering & decision tree algorithm." [1] від журналу Global Journal of Computer Science and Technology, можна зробити висновок, що змішана процедура, заснована на методі кластеризації даних і дерева рішень, може використовуватися органом влади для прогнозування роботи співробітників на наступний рік. У статті з результатами дослідження показано, як метод кластеризації даних можна застосувати для оцінки діяльності співробітника, а також у процесі прийняття рішень. Досліджено різні фактори оцінки ефективності, такі як особистість, пунктуальність, тактовність, вміння усно висловлюватися тощо [7, 18]. Результатом цієї роботи є прогнозування кількості працівників, які відбираються для підвищення або призначення та звільняються відповідно до їхньої роботи. Це дослідження допомагає з'ясувати неефективних співробітників, масштаби неефективності та допомагає усунути потенційну неефективність за допомогою відносно легкої системи.

У статті Jarman, Angur Mahmud - "Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method." [6] випущенів при Georgia Southern University у 2020-му році, згадується, що ієрархічні методи зв'язку здаються дуже простими - знайти евклідову відстань між кластерами, коли в кожному кластері є не більше одного випадку і частково це правда. Однак, якщо в кожному кластері є багато об'єктів, то обчислення евклідової відстані лише між двома об'єктами, що належить кожному умовному кластеру, недостатньо [1, 17].

Можемо зауважити, що існує позитивний досвід використання методів машинного навчання при вирішенні кадрових питань.

Актуальність дослідження. Нова реальність, що несподівано прийшла у наше повсякденне життя змусила всіх швидко адаптуватись. Разом з тим з'явилися й нові виклика та проблеми у кадровій сфері. Дослідження потенційної взаємодії офісних працівників між собою на основі їх попередніх характеристик методами машинного навчання, кластерним аналізом зокрема, може допомогти у вирішенні багатьох незручних, або непомітних, але від цього не менш значних кейсів під час підбору спеціалістів у команду, чи створенні нових команд. Також, результати даного дослідження можуть стати основою для нових, не обов'язково пов'язаних з формуванням команд досліджень (прогнозування успішності співробітника, прогнозування вигорання, підбір робочого місця, коригування рівня заробітної плати, покращення системи найму тощо).

Мета дослідження

З огляду на методи та дані, ця дослідницька робота вивчає взаємозалежність успішності роботи груп навчених суб'єктів від власних характеристик кожного члена групи. Можливість прогнозування співпраці окремих спеціалістів у команді в довгостроковій перспективі на основі методів машинного навчання, до яких належать різні моделі кластеризації й відповідні методи кластеризаційного аналізу. Окрім того, здобуття аналітичних даних для подальшого їх використання у можливій реалізації сервісу для підбору й групування персоналу та подальших досліджень.

Методи і матеріали досліджень

Для початку варто визначитись із поняттям кластеризації. Кластеризація даних - це неконтрольована процедура аналізу арифметичних даних. Кластерний аналіз використовується для сегментації великого набору даних на підмножини, які називаються кластерами [1, 2]. Кожен кластер - це набір об'єктів даних, які подібні один до одного в одному кластері, але не схожі на об'єкти інших кластерів. Він використовується для класифікації одних і тих самих даних в однорідну групу. Він також використовується для роботи з великим

набором даних для виявлення прихованих закономірностей і взаємозв'язків, які допомагають швидко та ефективно приймати рішення.

Оскільки кластеризація - це вельми поширений методі у дослідженнях різних сфер і має багато варіацій, не можна дати однозначно вірного визначення даному процесу. Але кожен з них об'єднує той факт, що кластерний аналіз полягає "групування" неструктурованих об'єктів за спільними характеристиками.

Умовно, загальну ієрархію методів кластерного аналізу можна побачити на рис. 1 [2, 13] нижче.

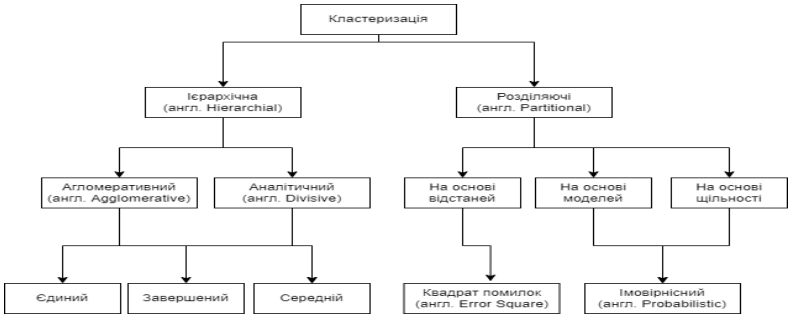


Рис. 1 Ієрархія кластерних методів [2]
Fig. 1 Hierarchy of cluster methods [2]

В методах ієрархічної кластеризації кластери формуються шляхом ітераційного поділу шаблонів за допомогою підходу зверху-вниз або знизу-вгору. У статті [3] згадується, що існує дві основні форми ієрархічного методу, а саме агломеративна та роздільна ієрархічна кластеризація [3, 11]. Ієрархічна кластеризація створює дерево кластерів (ієрархію кластерів), його часто називають дендрограмою (рис. 2). Агломераційна (англ. agglomerative) модель дотримується підходу знизу-вгору, який створює кластери, починаючи з одного об'єкта, а потім об'єднує ці атомарні кластери у все більші й більші кластери, допоки всі об'єкти нарешті не лежать в одному кластері або допоки не будуть виконані певні умови завершення. У свою чергу, розділяюча (аналітична, англ. divisible) модель виконується підходом зверху-вниз, який розбиває кластер, що містить усі об'єкти, на менші кластери, допоки кожен об'єкт не утворить кластер сам по собі або допоки не виконаються визначені умови завершення. Ієрархічні методи зазвичай призводять до формування вищезгаданих дендрограм [9, 13].

Незалежно від моделі, може використовуватись будь-який з основних ієрархічних методів: метод найближчого сусідів (метод одинарного зв'язку, англ. single linkage), метод дальнього сусіда (метод суцільного зв'язку, англ. complete linkage), метод середнього зв'язку (англ. average linkage) тощо.

У манускрипті [4, 12], автори згадують що таку кількість псевдонімів даний метод отримав не випадково, що цей тип кластеризації часто називають методом зв'язності, методом мінімуму чи методом найближчого сусіда. У однозв'язної кластеризації зв'язок між двома кластерами здійснюється за допомогою однієї пари елементів, а саме за допомогою двох елементів (по одному в кожному кластері), які знаходяться найближче один до одного. У цій кластеризації відстань між двома кластерами визначається найближчою відстанню від будь-якого члена одного кластера до будь-якого члена іншого кластера, це також визначає схожість. Якщо дані схожі, подібність між парою кластерів вважається рівною найбільшій подібності будь-якого члена одного кластера з будь-яким членом іншого кластера [5]. На рис. 2 показано відображення кластеризації одиночного зв'язку. Критерії між двома наборами кластерів A та B наступні (Формула 1):

$$min\{d(a,b): a \in A, b \in B\}. \quad (1)$$

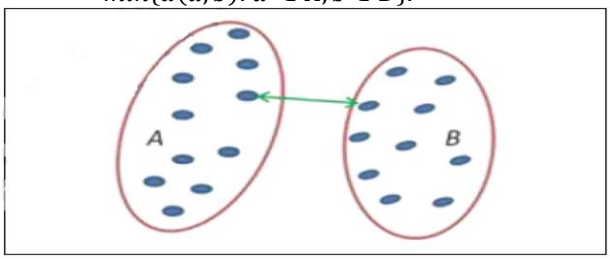


Рис.2. Маппинг кластеризації методом найближчого сусіда [5]
Fig. 2. Clustering mapping by the nearest neighbor method [5]

Кожен метод вимірювання зв'язку працює по-своєму. Незалежно від того, чи може один метод добре працювати для певного типу набору даних, він може не працювати в іншому типі набору даних. Загалом, метод найближчого сусіда (SNLK) залежить від найменшої відстані між двома точками, де кожна точка належить кожному кластеру з пари кластерів [7, 14]. Метод дальнього сусіда та середнього зв'язку бере найдальшу відстань та середнє значення відстаней відповідно. Наглядно це показано на рис. 3 нижче [6, 15].

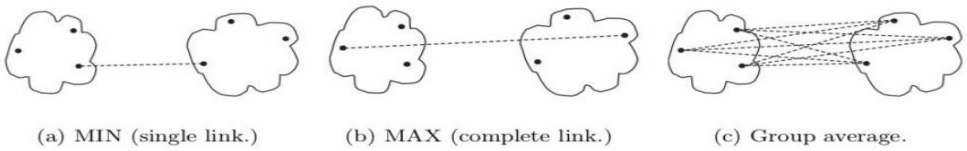


Рис. 3. Критерії зв'язку ієрархічних методів
Fig. 3. Communication criteria of hierarchical methods

Основна проблема, яка залишається в кластеризації методом найближчого сусіда, полягає в тому, що деякі кластери можуть об'єднуватися разом лише тому, що одна з їхніх точок даних знаходиться найближче до іншої точки в іншому кластері, тоді як більшість інших точок знаходяться на значно більшій відстані. Це називається ефектом ланцюжка, і цей ефект негативно впливає на загальний результат кластерного аналізу, якщо в наборі даних є шум [6, 16]. Те саме стосується й інших ієрархічних методів зв`язку. Результуюча дендрограма після застосування методу найближчого сусіда на тестовому наборі даних показана на рисунку 5 (a).

У методі повного зв'язку відстань між кластерами буде максимальною обчисленою відстанню, знайденою між наступними парами точок (Формула 2):

$$max\{d(a,b) : a \in A, b \in B\}. \quad (2)$$

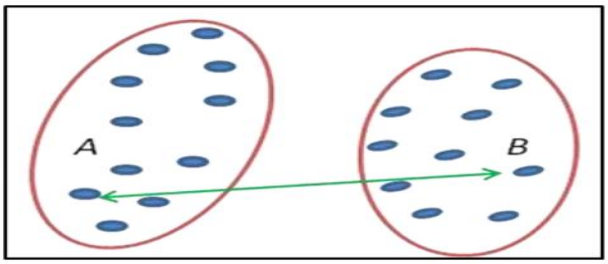


Рис. 4. Маппинг кластеризації методом дальнього сусіда [5]
Fig. 4. Clustering mapping using the far-neighbor method [5]

Результуюча дендрограма для даного методу зображена на рисунку 5 (b).
Для методу середнього зв'язку, де відстань обчислюється за наступною формулою 3:

$$\frac{1}{|A|\,|B|}\sum_{a\in A}\sum_{b\in B} d(a,b). \quad (3)$$

Дендрограму отриману після застосування методу середнього зв'язку зображено на рисунку 5 (c) вище.
Висота дендрограми - це відстань між скупченнями.

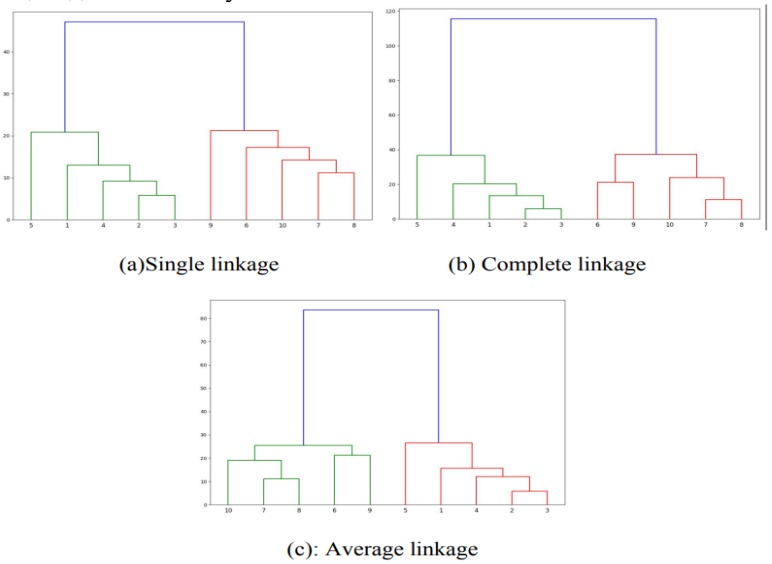


Рис. 5. Дендрограми отриманні після застосування методу: a) найближчого сусіда; b)дальнього сусіда; c) середнього зв'язку
Fig. 5. Dendrograms obtained after applying the method: a) nearest neighbor; b) distant neighbor; c) medium connection

У статті [7, 18] зазначається, що незважаючи на те, що принцип роботи FOREL алгоритму рекурентний, він

вирізняється своєю швидкодією. На кожному кроці випадковим чином вибирається об'єкт із вибірки, описується навколо нього сфера радіусу R, всередині цієї сфери вибирається центр тяжіння, і він стає центром нової сфери. Це продовжується допоки сфера не стабілізується чи не виконаються умови завершення. Тоді вихідну вибірку об’єктів, що належать сфері можна назвати кластером.

Також, автори згадують, що у лінійному просторі пошук центру відбувається за час $O(n)$, у метричному - $O(n^2)$, де n - обсяг вибірки, що кластеризується. Найкращих результатів алгоритм досягає вибірках з хорошим виконанням умов компактності [7, 17]. До переваг даного підходу можна віднести: точність мінімізації (при вдалому підборі початкових параметрів), можливість наглядної візуалізації процесу кластеризації (рис. 6), алгоритм сходиться тощо. До недоліків можна віднести відносно низьку швидкодію, низьку ефективність на даних з шумом, викидами чи просто поганою ділимістю, нездібність самостійно підібрати початкове значення радіусу на основі власних спостережень.

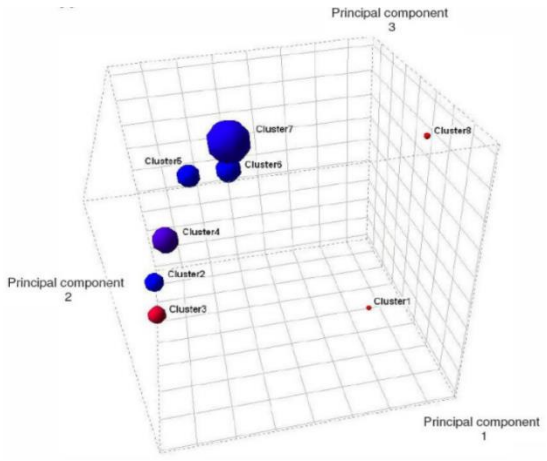


Рис. 6. Візуалізація роботи FOREL алгоритму у трьохвимірному просторі (3 компоненти)
Fig. 6. Visualization of the work of the FOREL algorithm in three-dimensional space (3 components)

Після опису попередніх методів та вступу не виникає сумнівів у необхідності дослідження для реалізації відбору співробітників методами машинного навчання, а саме методами кластерного аналізу. Для досягнення результату необхідної точності оптимальним шляхом було вирішено використовувати FOREL алгоритм під час дослідження.

Вирішення практичних проблем методами машинного навчання може здатися невигідним через кількість зусиль, часу та коштів витрачених на реалізацію. Та системи що управляються на “навченими” машинами існують для економії людських ресурсів. Після отримання готового продукту витрати завжди мінімізуються шляхом скорочення персоналу по вирішенні досліджуваної проблематики, зменшення впливу людського фактору і як результат - зменшення ризиків.

Завдання:

1. Здобути дані та нормалізувати;
2. Кластеризувати задану множину (знаходження центрів класів зокрема) та візуалізувати процес;
3. Обрати інші алгоритми кластеризації для порівняння результатів.

Для використання методів машинного навчання необхідно знайти дані. Датасет, що буде використовуватись у дослідженні яке базується на показниках працівників офісних компаній, повинен містити інформацію як про технічні, професійні показники співробітників, так і поведінкові, фізіологічні тощо. Така інформація може подаватись у вигляді таблиці, у якій кожен рядок - це запис про людину, а кожен стовпчик - характеристика чи ознака працівника. Для кластеризації методом багатовимірного неконтрольованого алгоритму FOREL, дані необхідно підготувати. У відкритому доступі можна знайти багато датасетів, що використовувались для кластеризації чи навчання вже існуючих систем. Їх можна використати у даному дослідженні, але більшість датасетів збирались незалежно, тому у кожного набір даних може відрізнятись описаними ознаками, критеріями, форматом чи кількістю даних. Для досягнення максимальної ефективності, необхідно вибрати датасети з найбільш схожими даними (за ознаками) та провести нормалізацію. Нормалізація полегшує процес обробки даних шляхом приведення різних даних до єдиного формату. Це дає можливість використовувати один інтерфейс при взаємодії з різними даними. Окрім того, під час нормалізації корисно позбутися шумів, викидів, неправильно заповнених даних та просто пустих значень , що можуть перешкодити процесам машинного навчання.

Для нормалізації було обрано лінійний метод (англ. Linear normalization, “Max-Min”). Нормування датасету даним методом полягає у приведенні значень вибірки до деяких необхідних меж (наприклад, від 0 до 1). Таке нормування є необхідним початковим етапом обробки особливо даних при використанні алгоритму, що виконується у багатовимірному просторі (як FOREL). Оскільки різні категорії даних з одного датасету можуть істотно різняться у величинах (наприклад, кілометри, роки чи кілограми). Якщо зв'язок між початковим і кінцевим діапазонами виражається лінійною функцією, то нормування є лінійним (значення ознак по всьому

діапазону змінюються на одну й ту саму величину), приклад зображено на рис. 7. Якщо використовується нелінійна функція (наприклад, експонента), то нелінійним (значення ознак, що розташовані у різних частинах діапазону, змінюються по-різному). У нашому випадку буде використано лінійну функцію.

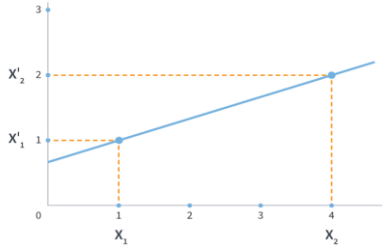


Рис. 7. Графічне представлення методу лінійної нормалізації
Fig. 7. Graphic representation of the linear normalization method

Лінійне нормування доречне, коли значення ознаки *X* рівномірно розподіленою на певний інтервал. Якщо в даних є аномалії, що сильно перевищують типове розподілення, то в цих випадках при нормуванні слід орієнтуватися не на екстремальні (граничні) значення (Формула 4):

$$\underline{x_i} = \frac{x_i - x_{i\,min}}{x_{i\,max} - x_{i\,min}}, \tag{4}$$

а на типові значення - середні (Формула 5):

$$x_{i\,avg} = \frac{1}{p}\sum_{k=1}^{p} X_I{}^k, \tag{5}$$

та дисперсію (Формула 6):

$$\underline{x_i} = \frac{x_i - x_{i\,avg}}{\sigma^i}, \tag{6}$$

Для початку у дослідженні використаємо публічний датасет “Employee Future Prediction” [8] (Рис. 8) з відкритого джерела Kaggle, що містить інформацію досвід роботи, вік, стать, компенсацію, досвід у конкретних областях, загальне враження, освіту тощо.

	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	Leave
0	Bachelors	2017	Bangalore	3	34	Male	No	0	0
1	Bachelors	2013	Pune	1	28	Female	No	3	1
2	Bachelors	2014	New Delhi	3	38	Female	No	2	0
3	Masters	2016	Bangalore	3	27	Male	No	5	1
4	Masters	2017	Pune	3	24	Male	Yes	2	1
...	...	...	...	...	...	...	...	...	...
4648	Bachelors	2013	Bangalore	3	26	Female	No	4	0
4649	Masters	2013	Pune	2	37	Male	No	2	1
4650	Masters	2018	New Delhi	3	27	Male	No	5	1
4651	Bachelors	2012	Bangalore	3	30	Male	Yes	2	0
4652	Bachelors	2015	Bangalore	3	33	Male	Yes	4	0

Рис. 9. Приклад даних досліджуваного датасету
Fig. 9. An example of data from the studied dataset

Представлено інформацію більш як чотирьох з половиною тисяч співробітників. Даний датасет ліцензований, доволі популярний на публічному ресурсі та має максимально можливий рейтинг.

Після підготовки даних можна приступити до кластеризації безпосередньо.

Неконтрольований алгоритм класифікації дає можливість провести більш “природну класифікацію” [9] за рахунок того, що на початку ми не знаємо кількість кластерів та не міняємо її самостійно під час процесу. Цей алгоритм виконує кластеризацію типу “дослідження вибірки” [10] без попередньої визначення невеликого набору точок в просторі для подальшого зменшення чи збільшення їх розмірності. Результат складається з набору готових кластерів, які можуть бути додатково проаналізовані та гіперкластеризовані, щоб встановити взаємозв'язок між природними класами. Тому й даний метод кластеризації відносять до “природних” [9]. Звідси слідує, що під час підбору працівників немає необхідності будувати команду довкола якоїсь одної людини - алгоритм сам підбере найбільш виважену комбінацію співробітників.

Важливу роль під час кластеризації працівників FOREL методом відіграє змінна R. R - це значення радіусу багатовимірної сфери центром якої є центр мас всіх об’єктів кластеру, де всі об’єкти містяться у сфері. Дане значення задається дослідником самостійно в залежності від “розрідженості” даних та бажаного результату. Очевидно, що чим більше значення радіусу, тим більше співробітників вийде в одному кластері (команді) як результат кластеризації. Визначати дану змінну варто в залежності від бажаних розмірів вихідної команди та різниці між досліджуваними даними.

У тестових цілях спробуємо провести й візуалізувати кластеризацію працівників у двовимірному просторі (дві ознаки). Для даного дослідження візьмемо рік приєднання працівника до компанії та коефіцієнт заробітної плати. Виберемо 100 випадкових співробітників. Під час нормалізації рік приєднання було приведено до кількості років проведених у компанії. Окрім того, після очищення шумів та позбавлення викидів, всі значення

кожної з ознак розміщено на проміжку від 0 до 1 (де нулю відповідає найменше, а 1 - найбільше значення вибірки) для наглядної візуалізації та спрощення процесу обрахунків.

Візуалізацію даних співробітників з тестової вибірки зображено на рисунку 10 нижче. Оскільки у даному датасеті існує тільки 3 ранги коефіцієнтів до заробітної плати (англ. payment tier) та максимум досвіду у даній вибірці представлено як 10 років, комбінації значень (позиції працівників у просторі відповідно) можуть повторюватись. Тому точок на рис. 10 менше ніж 100.

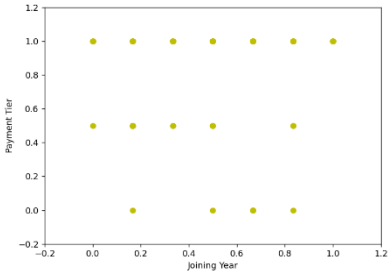
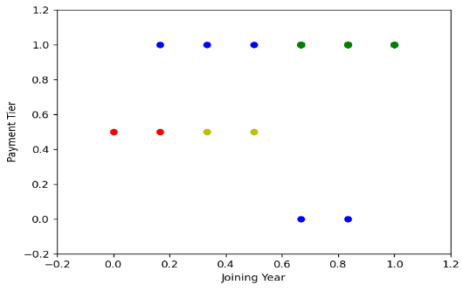


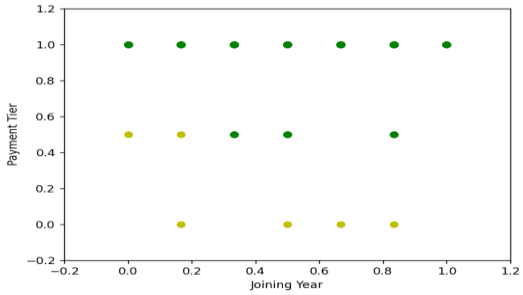
Рис. 10. Візуалізація двох ознак співробітників з тестової вибірки
Fig. 10. Visualization of two characteristics of employees from the test sample

Допустимо, нам треба зібрати декілька команд з наявних працівників розміром по 7-10 працівників, вважаючи, що одна точка на графіку - це один працівник.

Для початку можна спробувати запустити кластеризацію з радіусом *0,3*. Результат продемонстровано на рисунку 11a. Як можемо бачити, радіус *0,3* недостатній для того, щоб в результаті виникли кластери розмірність 7-10 осіб. Можемо спостерігати 5 груп по 2-3 співробітники у кожній. Варто зазначити, що даний алгоритм не завжди приводить до одного й того самого результату. Після чисельних тестових запусків проглядається певна закономірність, що дозволяє висунути припущення - з вхідним радіусом *0,3* результатом кластеризації для даної вибірки будуть кластери (групи) розміром в 1-3 об'єкти. Збільшимо радіус до *0,6*. Результати кластеризації з радіусом *0,6* продемонстровано на рисунках 11b, 12.



a)



b)

Рис. 11. Кластеризація працівників за двома ознаками: a) з радіусом 0,3; b) з радіусом 0,6
Fig. 11. Clustering of employees by two characteristics: a) with a radius of 0.3; b) with a radius of 0.6

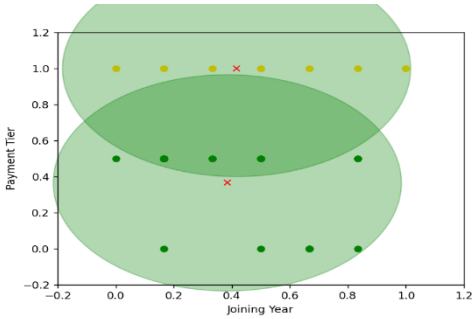


Рис. 12. Межі та центр кластерів приз радіусі 0,6
Fig. 12. Boundaries and centers of clusters with a radius of 0.6

На рисунку червоною міткою позначено центр мас кластеру, а зелена площа довкола охоплює всі елементи даного кластеру. Варто згадати, що під час виконання методу, об'єкти, які вже входять у якийсь кластер видаляються, а сам алгоритм виконується допоки не буде знайдено кластер для кожного. Тому, якщо у візуальному представленні є точки, що входять у зону декількох центрів мас, кожна з них все-одно належить тільки одному кластеру.

Спробуємо кластеризувати працівників по трьох ознаках та представити результати візуально у трьохвимірному просторі. Для дослідження до вже існуючих ознак (рік приєднання до компанії, коефіцієнт

доходу) додаймо більш персональний показник - вік. Спробуємо запустити алгоритм кластеризації одразу з радіусом *0,6*. Візуалізацію результатів кластеризації представлено на рисунку 13 (а). Подібні результати показують й інші запуски з даним радіусом. Спробуємо зменшити розмір радіусу до *0,3* одиниць. Результати продемонстровано на рисунку 13 (b).

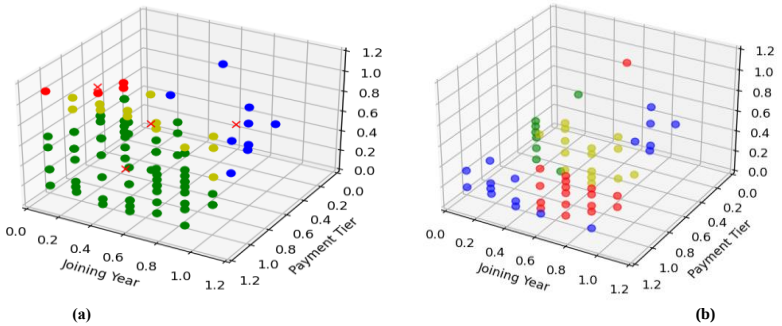


Рис. 13. Кластеризація з радіусом: a) 0,6 b) 0,3
Fig. 13. Clustering with a radius of: a) 0.6 b) 0.3

Як бачимо, результат краще підходить нашим вимогам (7-10 співробітників на кластер) та все ще занадто багато. Спробуємо зменшити радіус до *0,2*. Результат даної кластеризації зображено на рисунку 14.

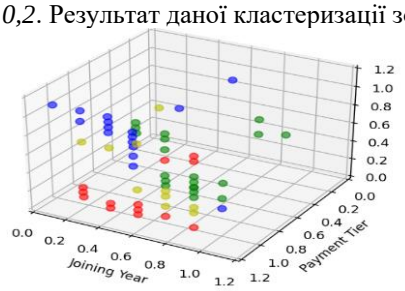


Рис. 14. Результат кластеризації з радіусом величиною 0,2
Fig. 14. The result of clustering with a radius of 0.2

Після декількох експериментів над вхідним значенням радіусу можна дійти до висновку, що для нашої тестової задачі найкраще підходить радіус розміром *0,2*. Результатом є кластери розмірністю 6-12 осіб, що зумовлено непередбачуваністю алгоритму (неконтрольований), неможливістю визначити кількість кластерів на початку та “розкидом” даних. Врахування віку внесло помітні корективи у процес кластеризації - тут значення можуть бути зовсім різні, тому сильно поменшало працівників, що за своїми ознаками знаходяться в тій самій точці.

Для порівняння результатів та якості виконаної роботи було вирішено використовувати метод *k*-середніх. Це один із найпопулярніших методів кластеризації. Процес починається з поділу *m* спостережень (з простору) на *k* кластерів, при цьому кожне спостереження відноситься до кластера, до центру (центроїду) якого воно найближче.

Як якості міри близькості використовується Евклідова відстань (Формула 7):

$$\rho(x,y) = ||x-y|| = \sqrt{\sum_{p=1}^{n}(x_p - y_p)^2}\text{, де } x,y \in R^n \quad (7)$$

Нехай, у нас є ряд спостережень $(x^{(1)}, x^{(2)}, \ldots, x^{(m)}), x^{(j)} \in R^n$.

Тоді, методом *k*-середніх розділяються *m* спостережень на *k* груп (кластерів), де $k \leq m$. Кожне з випадково обраних спостережень присвоюється до якогось *n* кластерів - до того, до якого відстань найменша, рисунок 15.

Після, проводиться обрахунок нового центра для кожного з кластерів. Сам центр визначається середнє арифметичне ознак точок даного кластеру, рисунок 15. Тоді, повторюються кроки 3-4, допоки при кожній ітерації об’єкти не потраплятимуть в той же кластер, тобто допоки кластери не стабілізуються (стануть стійкими), а дисперсія всередині кластера мінімізується, в той час як між кластерами навпаки - буде максимізована (6), візуалізацію даного кроку представлено на рисунку 15 (Формула 8).

$$min\left[\sum_{i=1}^{k}\sum_{x\in S_i}(x_i - \mu_i)^2\right], \quad (8)$$

де *k* - число кластерів, S_i - отримані кластери, $i=1,2,3\ldots,\mu_i$ - центр мас векторів $x_i \in S_i$;

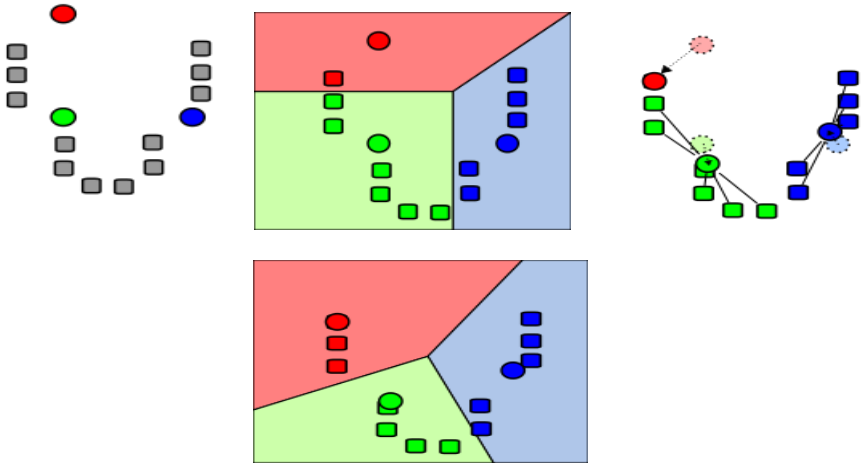


Рис. 15. Візуалізація процесу кластеризації методом k-середніх [11-12]
Fig. 15. Visualization of the clustering process using the k-means method [11-12]

Важливо зазначити, що в даному алгоритмі (на відміну від FOREL), неправильний вибір початкової кількості кластерів *k* може призвести до некоректних результатів. Саме тому при використанні методу *k*-середніх важливо спочатку провести перевірку відповідного числа кластерів для тестового набору даних.

Для експериментів буде використовуватися датасет “Employee Future Prediction” [8] (рис. 16) з відкритого джерела Kaggle. Досліджувані ознаки - досвід (у роках), коефіцієнт зарплати, вік, гендер (буде переведено у числове значення 0 або 1), освіта (у датасеті представлено 3 типи освіти: Bachelor, Master, PHD, які будуть умовно переведені у числовий формат), досвід у обраному відділі тощо. Завдяки великій кількості ознак мінімізується ймовірність того, що двоє співробітників знаходитимуться у одній точці багатовимірного простору на момент кластеризації. Оскільки ознак багато, простір у якому буде проводитись кластеризація - багатовимірний, що унеможливлює візуалізацію всього процесу.

	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	Leave
0	Bachelors	2017	Bangalore	3	34	Male	No	0	0
1	Bachelors	2013	Pune	1	28	Female	No	3	1
2	Bachelors	2014	New Delhi	3	38	Female	No	2	0
3	Masters	2016	Bangalore	3	27	Male	No	5	1
4	Masters	2017	Pune	3	24	Male	Yes	2	1
...	...	...	...	...	...	...	...	...	...
4648	Bachelors	2013	Bangalore	3	26	Female	No	4	0
4649	Masters	2013	Pune	2	37	Male	No	2	1
4650	Masters	2018	New Delhi	3	27	Male	No	5	1
4651	Bachelors	2012	Bangalore	3	30	Male	Yes	2	0
4652	Bachelors	2015	Bangalore	3	33	Male	Yes	4	0

Рис. 16. Приклад даних досліджуваного датасету
Fig. 16. An example of data from the studied dataset

У даному датасеті більш як 4 тисячі записів, тому, скоріш за все, там є дані що сильно між собою різняться.

Ціллю експерименту є дослідити можливість кластеризації офісних працівників у команди засобами машинного навчання, алгоритмом FOREL зокрема. У представленому датасеті міститься достатньо інформації для проведення експерименту. На основі цього, можна вивести практичну задачу - зібрати в команди 4652 працівники. Після очистки від шумів, викидів та нормалізації можна приступати до експерименту.

Для початку необхідно підібрати радіус R для проведення у відповідності з бажаними розмірами команд.

Спробуємо провести кластеризацію даних із заданого датасету FOREL методом з радіусом $R = 0{,}9$ (маємо на увазі, що після нормалізації, значення всіх ознак коливаються в межах $[0;\ 1]$). Кластеризація пройшла успішно. Побудуємо діаграму розмірів вихідних кластерів (рисунок 17).

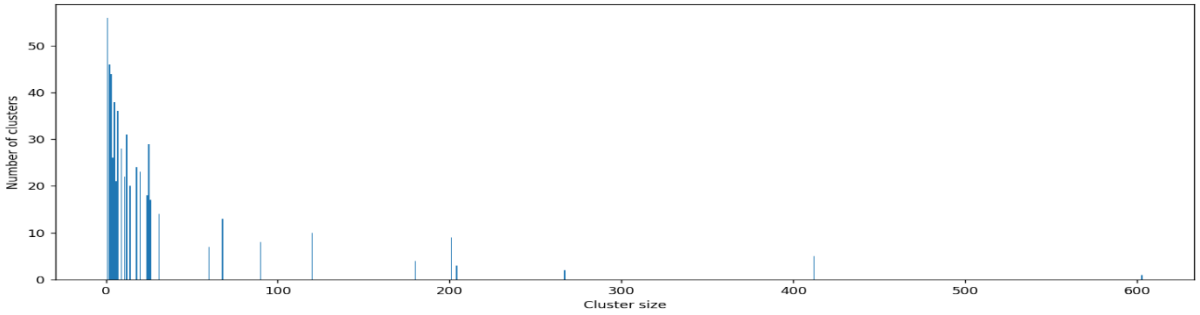


Рис. 17. Діаграма розмірів кластерів при $R = 0{,}9$
Fig. 17. Diagram of cluster sizes at R=0.9

Як бачимо, розмір більшості кластерів коливається від одного до 25-ти, хоча існують і кластери розмірністю в 410 чи 605. Це означає, що при радіусі $R = 0{,}9$ більшість даних увійшло до сфери, центр мас якої знаходиться у великому скупченні. Ймовірно, інші кластери також були б великі за розміром, якби під час алгоритму не видалялися б "закластеризовані" об'єкти. Очевидно, що радіус $R = 0{,}9$ не є доречним для вирішення задачі підбору персоналу з даними пропонованого датасету.

Проведемо експеримент використовуючи радіус розміром $R = 0{,}1$. Обчислення з невеликим значенням R займає значно довше часу, оскільки потребує більше ресурсів на нормалізацію кластерів. У даному випадку кількість кластерів вища, що представлено на рисунку 18. Оскільки дані доволі "розріджені", а значення R невелике, кількість кластерів розмірністю 1 найвища. На це варто звернути особливу увагу, так як малоймовірно, команди з однієї людини є ціллю при кластеризації співробітників. В той же час, у даній ситуації вже можна отримати користь, якщо у кадрових працівників є задача зібрати команди з 2-3 осіб. Дані групи (кластери) сотворились за спільними характеристиками і вже можуть бути допоміжними у процесі найму.

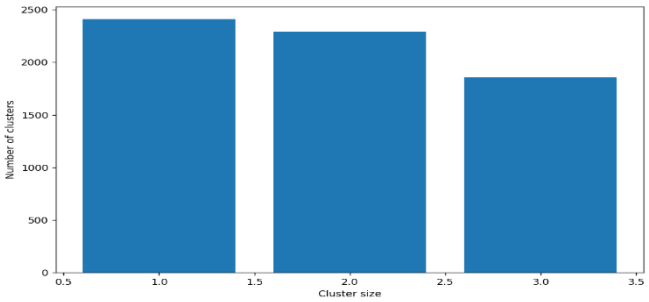


Рис. 18. Діаграма розмірів кластерів при R = 0, 1
Fig. 18. Diagram of cluster sizes at R=0.1

Хоча з результатів при $R = 1$ можна отримати корисну інформацію, це не є найбільш ефективним варіантом для підбору персоналу.

Кластеризуймо дані обраного датасету використовуючи радіус $R = 0{,}35$. Результати кластеризації представлено на рисунку 19 нижче. Можемо бачити, що даний розмір радіусу можна вважати найбільш оптимальним, оскільки розміри кластерів не розкидаються на неочікувані проміжки та є близькими до розмірів команд що трапляються в реальному житті. На основі діаграми умовний працівник кадрового відділу чи прожект менеджер може глянути на дані про працівників, що відповідають команді бажаного розміру. Як результат, отримано 622 кластери, їх можна використовувати як підказки на етапі формування нових команд.

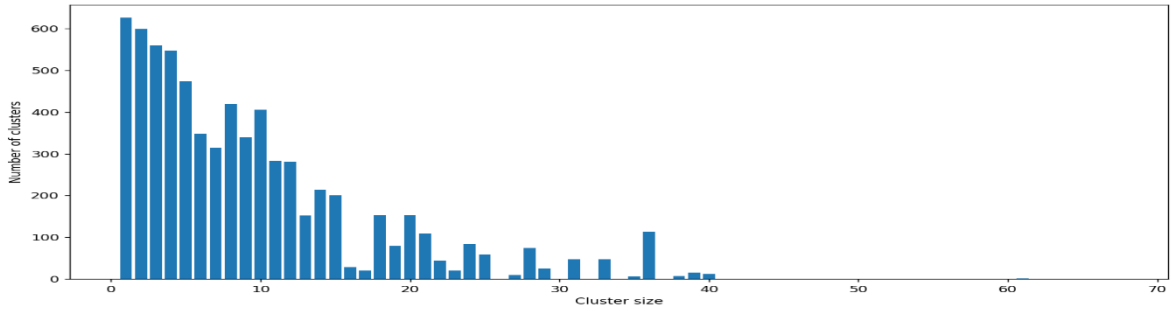


Рисунок 19 - Діаграма розмірів кластерів при R = 0, 35
Figure 19 - Diagram of cluster sizes at R=0.35

Обговорення результатів

На відміну від FOREL алгоритму, перед стартом методу k-середнє необхідно вказати очікувану кількість кластерів. Найбільш ефективних результатів було при використанні FOREL методу було досягнуто з радіусом $R = 0{,}35$, що призвело до утворення 622 нових кластерів. Для об'єктивної оцінки проведемо k-mean кластеризацію з вхідним параметром кластерів $N = 622$. Оскільки ознак працівника більше трьох, візуалізація процесу кластеризації не є доцільною. Розподілення кількості кластерів у відношенні до розмірностей продемонстровано на рисунку 20 нижче.

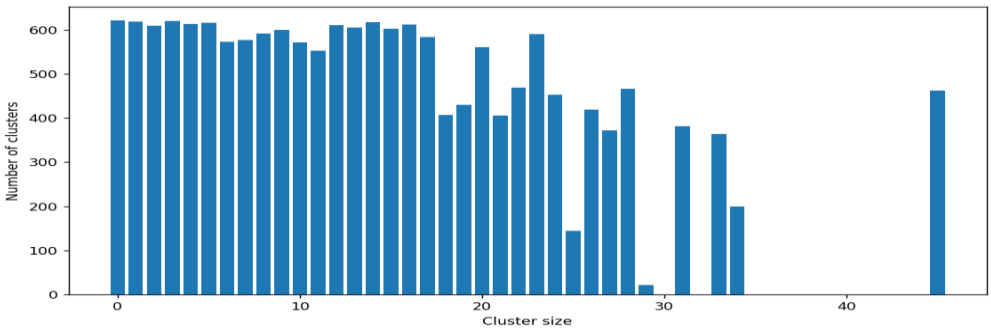


Рис. 20. Діаграма розмірів k-mean кластерів при N = 622
Fig. 20. Diagram of k-mean cluster sizes at N=622

Як бачимо, кластеризація методом *k*-середнє дає порівняно схожі результати. Недоліком тут є необхідність задавати очікувану кількість кластерів, що часто змушує бути прив'язаним до конкретної кількості працівників.

Висновки

Задача найму і розподілення працівників завжди є актуальною. У наш час внаслідок процесів, що невідворотно впливають на наше життя (COVID-19, російсько-українська війна), правила найму та кадрові рішення також змінюються й ускладняються. Окрім того, рекрутинг та тім-менеджмент - це галузь, де необхідно мати багато інформації. А де багато інформації, там доречним є використання методів машинного навчання, як мінімум. Це дозволило успішно використати алгоритми кластеризації для групування офісних працівників за їх характеристиками в даній роботі.

Для роботи було використано набір даних працівників "Employee Future Prediction" типовими ознаками спеціалістів. Варто відмітити, що кількість ознак можна розширити якими-завгодно даними (сімейний стан, впевненість, харизма, рівень IQ тощо).

Перспектива використання методів машинного навчання чи штучного інтелекту у повсякденній роботі, яку не прийнято автоматизувати стає все більш реальною. Це наближає нас до того самого майбутнього, де вся рутинна робота виконується спеціально навченими системами, а людям залишається творчість та спілкування. На основі методів з даної роботи можна створити систему, що слугуватиме помічником людям, які вирішують кадрові задачі.

Список використаних джерел

1. Sarker A. et al. Employee's performance analysis and prediction using K-means clustering & decision tree algorithm. Global Journal of Computer Science and Technology, 2018.
2. Fraley C., Raftery A. E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. Technical Report No. 329, Department of Statistics University of Washington, 1998.
3. Murtagh F. A survey of recent advances in hierarchical clustering algorithms which use cluster centers. Computer Journal, 2020. vol. 26. no. 4. pp. 354-359.
4. Saxena A., Prasad M., Gupta A., Bharill N., Patel O. P., Tiwari A. & Lin C. T. A review of clustering techniques and developments. neurocomputing, 2017. Vol. 267. p. 664-681.
5. Sneath P., Sokal R. Numerical Taxonomy. Freeman Co, San Francisco, CA.
6. Jarman Angur Mahmud. Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. Georgia Southern University, 2020.
7. Science portal Studeme. Study materials for students (info@studeme.org) © 2013 - 2022, 13.4.3. Алгоритм FOREL.
8. Dataset: Employee Future Prediction. Predict Employee Future In Company, https://www.kaggle.com/datasets/tejashvi14/employee-future-prediction. LICENSE CC0: Public Domain, Tejashvi (Owner), DOI (DIGITAL OBJECT IDENTIFIER).
9. Ptitsyn A., Hulver M., Cefalu W., York D., & Smith S. R. BMC Genomics, 2016. Vol. 7(1). p. 318. doi:10.1186/1471-2164-7-318.
10. Tung A.K., Hou J., Han J. Spatial clustering in the presence of obstacles // The 17th Intern. conf. on data engineering (ICDE'01). Heidelberg, 2001. p. 359–367. DOI: 10.1109/ICDM.2002.1184042
11. Boehm C., Kailing K., Kriegel H., Kroeger P. Density connected clustering with local subspace preferences // IEEE Computer Society. Proc. of the 4th IEEE Intern. conf. on data mining. Los Alamitos, 2004. p. 27–34. DOI: 10.1007/978-0-387-39940-9_605
12. Boyko N., Kmetyk-Podubinska K., Andrusiak I. Application of Ensemble Methods of Strengthening in Search of Legal Information. Lecture Notes on Data Engineering and Communications Technologies. 2021. Vol. 77. p. 188-200. URL: https://doi.org/10.1007/978-3-030-82014-5_13.
13. Boyko N., Hetman S., Kots I. Comparison of Clustering Algorithms for Revenue and Cost Analysis // Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Lviv, Ukraine. 2021, Vol.1. p. 1866-1877.

14. Procopiuc C.M., Jones M., Agarwal P.K., Murali T.M. A Monte Carlo algorithm for fast projective clustering // ACM SIGMOD Intern. conf. on management of data, Madison, Wisconsin, USA. 2002. p. 418–427.
15. Boyko N. Application of mathematical models for improvement of “cloud” data processes organization”. Mathematical Modeling and Computing, 2016. Vol. 3(2). p. 111–119. DOI: https://doi.org/10.23939/mmc2016.02.111
16. Hossain M. Z., Akhtar M. N., Ahmad R. B. and Rahman M. A dynamic K-means clustering for data mining. Indonesian Journal of Electrical Engineering and Computer Science, 2017. Vol. 13 (2). p. 521–526. DOI: http://doi.org/10.11591/ijeecs.v13.i2.pp521-526
17. Slamet C., Rahman A., Ramdhani M. A., and Darmalaksana W. Clustering the verses of the Holy Qur'an using K-means algorithm. Asian Journal of Information Technology, 2016. Vol. 15. no. 24. pp. 5159–5162.
18. Bekiros S., Nguyen D. K., Sandoval Junior L. and Uddin G. S. Information diffusion, cluster formation and entropy-based network dynamics in equity and commodity markets. European Journal of Operational Research, 2017. Vol. 256(3). p. 945–961. DOI: 10.1016/j.ejor.2016.06.052.

References

1. Sarker A. et al. Employee’s performance analysis and prediction using K-means clustering & decision tree algorithm. Global Journal of Computer Science and Technology, 2018.
2. Fraley C., Raftery A. E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. Technical Report No. 329, Department of Statistics University of Washington, 1998.
3. Murtagh F. A survey of recent advances in hierarchical clustering algorithms which use cluster centers. Computer Journal, 2020. vol. 26. no. 4. pp. 354-359.
4. Saxena A., Prasad M., Gupta A., Bharill N., Patel O. P., Tiwari A. & Lin C. T. A review of clustering techniques and developments. neurocomputing, 2017. Vol. 267. p. 664-681.
5. Sneath P., Sokal R. Numerical Taxonomy. Freeman Co, San Francisco, CA.
6. Jarman Angur Mahmud. Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. Georgia Southern University, 2020.
7. Science portal Studeme. Study materials for students (info@studeme.org) © 2013 - 2022, 13.4.3. Алгоритм FOREL.
8. Dataset: Employee Future Prediction. Predict Employee Future In Company, https://www.kaggle.com/datasets/tejashvi14/employee-future-prediction. LICENSE CC0: Public Domain, Tejashvi (Owner), DOI (DIGITAL OBJECT IDENTIFIER).
9. Ptitsyn A., Hulver M., Cefalu W., York D., & Smith S. R. BMC Genomics, 2016. Vol. 7(1). p. 318. doi:10.1186/1471-2164-7-318.
10. Tung A.K., Hou J., Han J. Spatial clustering in the presence of obstacles // The 17th Intern. conf. on data engineering (ICDE’01). Heidelberg, 2001. p. 359–367. DOI: 10.1109/ICDM.2002.1184042
11. Boehm C., Kailing K., Kriegel H., Kroeger P. Density connected clustering with local subspace preferences // IEEE Computer Society. Proc. of the 4th IEEE Intern. conf. on data mining. Los Alamitos, 2004. p. 27–34. DOI: 10.1007/978-0-387-39940-9_605
12. Boyko N., Kmetyk-Podubinska K., Andrusiak I. Application of Ensemble Methods of Strengthening in Search of Legal Information. Lecture Notes on Data Engineering and Communications Technologies. 2021. Vol. 77. p. 188-200. URL: https://doi.org/10.1007/978-3-030-82014-5_13.
13. Boyko N., Hetman S., Kots I. Comparison of Clustering Algorithms for Revenue and Cost Analysis // Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Lviv, Ukraine. 2021, Vol.1. p. 1866-1877.
14. Procopiuc C.M., Jones M., Agarwal P.K., Murali T.M. A Monte Carlo algorithm for fast projective clustering // ACM SIGMOD Intern. conf. on management of data, Madison, Wisconsin, USA. 2002. p. 418–427.
15. Boyko N. Application of mathematical models for improvement of “cloud” data processes organization”. Mathematical Modeling and Computing, 2016. Vol. 3(2). p. 111–119. DOI: https://doi.org/10.23939/mmc2016.02.111
16. Hossain M. Z., Akhtar M. N., Ahmad R. B. and Rahman M. A dynamic K-means clustering for data mining. Indonesian Journal of Electrical Engineering and Computer Science, 2017. Vol. 13 (2). p. 521–526. DOI: http://doi.org/10.11591/ijeecs.v13.i2.pp521-526
17. Slamet C., Rahman A., Ramdhani M. A., and Darmalaksana W. Clustering the verses of the Holy Qur'an using K-means algorithm. Asian Journal of Information Technology, 2016. Vol. 15. no. 24. pp. 5159–5162.
18. Bekiros S., Nguyen D. K., Sandoval Junior L. and Uddin G. S. Information diffusion, cluster formation and entropy-based network dynamics in equity and commodity markets. European Journal of Operational Research, 2017. Vol. 256(3). p. 945–961. DOI: 10.1016/j.ejor.2016.06.052.

Отримана в редакції 08.03.2023. Прийнята до друку 10.04.2023. Received 08 March 2023. Approved 10 April 2023. Available in Internet 12 April 2023.