УДК 004.738.5:005 **DOI: 10.37332/2309-1533.2019.7-8.10**
JEL Classification: M21, G14

Струтинська І.В.,
канд. екон. наук, доцент кафедри комп'ютерних наук,
Тернопільський національний технічний
університет імені Івана Пулюя

КЛАСТЕРИЗАЦІЯ БІЗНЕС-СТРУКТУР ЗА РІВНЕМ ЇХ ЦИФРОВОЇ ЗРІЛОСТІ З ВИКОРИСТАННЯМ ДВОХ ПІДХОДІВ: ІТЕРАТИВНОГО ТА ІЄРАРХІЧНОГО

Strutynska I.V.,
cand.sc.(econ.), associate professor at the
department of computer science,
Ternopil Ivan Puluj National Technical University

CLUSTERING OF BUSINESS STRUCTURES BY THE LEVEL OF THEIR DIGITAL MATURITY USING TWO APPROACHES: ITERATIVE AND HIERARCHICAL

Постановка проблеми. Найзручнішим та найефективнішим способом отримання будь-якої інформації про ту чи іншу проблематику чи респондентів у сьогоденні все ще залишається безпосереднє опитування цільової аудиторії на визначену тематику. Зі зростанням використання цифрових технологій такі анкетування все частіше переходять від особистого спілкування чи за допомогою телефону до онлайн-опитувальників. Це дозволяє охопити більшу аудиторію за коротший часовий проміжок та з затратою меншої кількості людських ресурсів. Як позитивні сторони проведення такого роду анкетувань можна відмітити: зручність висловлення думок; часткова або повна анонімність результатів; можливість проходження опитування у будь-який зручний для респондента спосіб та час; відсутність необхідності комунікації з працівниками організації, що проводить опитування тощо. Онлайн-опитування є особливо ефективним способом добування інформації у випадку, якщо уся цільова аудиторія є користувачами всесвітньої мережі в меншій чи більшій мірі. Збір даних є лише частиною комплексної задачі отримання потрібної інформації. Подальша обробка та аналіз даних з отриманням висновків та рекомендацій роблять цикл роботи з даними повним. Дані відповідних опитувань (досліджень) опрацьовують різноманітним чином, що передбачає витрату часу та залучення та врахування суб'єктивної думки людини, що опрацьовує відповідні результати. Проте, дану проблематику можна вирішити, використовуючи задачі аналізу даних, такі, як сегментація або кластеризація із застосуванням передових інформаційних технологій аналізу даних.

Аналіз останніх досліджень і публікацій. Задачу кластеризації числових даних як результатів певної серії вимірювань було описано у роботі Чєрєзова Д. С. та Тюкачєва Н. А. [1]. Схожу проблему групування респондентів було розглянуто в праці «Clustering Online Poll Data: Towards a Voting Assistance System» таких зарубіжних вчених, як: I. Katakis, N. Tsapatsoulis, C. Tziouvas and F. Mendes (2012) [2]. Основною метою роботи було надання рекомендацій щодо результатів визначення політичних вподобань респондентів та порівнянню методу кластеризації з іншими загальноприйнятими методиками надання таких рекомендацій. Проблематику кластеризації категорійних даних із використанням ймовірнісного підходу та алгоритму GACUC було досліджено у роботі J. McCaffrey «Machine Learning Using C#» (2014) [3]. Проте задача обробки та кластеризації даних змішаного типу, отриманих в результаті анкетування, на сьогоднішній день досліджена недостатньо.

***Постановка завдання*.** Метою статті є експериментальне знаходження оптимальної кількості кластерів та їх характерних рис для інтерпретованої (такої, що можна зрозуміти і пояснити) сегментації бізнес-структур за рівнем цифрової зрілості за допомогою декількох методів; порівняння результатів, отриманих різними методами, та визначення найбільш ефективного для конкретної задачі аналізу даних.

Завданнями виступають: дослідження та кластеризація респондентів опитування (бізнес-структур Тернопільської області) з використанням двох підходів: ітеративного та ієрархічного – до отримання стійких та зрозумілих результатів; пропонування алгоритму розв'язання задачі кластеризації респондентів за результатами онлайн-опитування, включаючи етапи збору, підготовки даних, отримання основних підсумків та вироблення майбутніх цілей.

Виклад основного матеріалу дослідження. Кластеризація або кластерний аналіз даних – одна із задач машинного навчання без вчителя, що полягає у розбитті множини об'єктів на підмножини (кластери) таким чином, щоб об'єкти, віднесені до одного кластера, були максимально схожими один на одного, а об'єкти, віднесені до різних, – максимально відмінними.

Одними із найбільш поширених сучасних задач, у яких використовується кластерний аналіз, є: аналіз текстів для потоків новин, групування зображень, сегментація споживачів, виокремлення спільнот в соціальних мережах тощо.

Варіативність завдань, видів наборів даних та очікуваних результатів призвела до утворення великої кількості методів та підходів до кластеризації, які відрізняються між собою як розумінням поняття «кластер», так і налаштуванням параметрів алгоритмів (кількість очікуваних кластерів, поріг щільності, метрики відстаней тощо) залежно від специфіки набору даних та подальшого використання результатів. У свою чергу, це спричиняє важкість однозначного вибору алгоритму роботи та його параметрів для кожного типу задач.

Зважаючи на це, кластеризацію також можна назвати інтерактивною задачею машинного навчання «з підкріпленням», що передбачає неодноразову експериментальну корекцію параметрів алгоритму задля отримання стійких та інтерпретовних результатів [4; 5].

Не існує єдиного загальноприйнятого способу класифікації методів та алгоритмів кластеризації. Один із підходів передбачає розрізняти методи кластеризації за моделями кластерів, що використовуються (підключення на основі зв'язку або ієрархічне, на основі центроїдів, на основі розподілу, кластеризація, що накладається на щільність, тощо). Інший – використовує групування методів за підходами, які лежать у їх основі (ймовірнісні, логічні, теоретико-графові, ієрархічні, нейронні, частотні алгоритми тощо) [5].

Одним із найпростіших підходів до методів кластеризації є їх поділ на дві групи: ієрархічні та неієрархічні. Ієрархічні методи кластерного аналізу поділяють на висхідні або низхідні та можуть бути представлені графічно у вигляді дендрограм. При цьому з кожним наступним кроком кількість кластерів збільшується або зменшується у залежності від вибраного методу: дивізивного або алгомеративного відповідно.

Найбільшою групою серед неієрархічних методів є ітеративні. При ітеративному підході визначають центри кластерів та перерозподіляють елементи набору даних за близькістю до вибраних центрів. До таких належать алгоритми k-means, Expectation-Maximization method, mean-shift та інші [6].

«Схожість» елементів кластерів та «близькість» кластерів визначають за обумовленими наперед метриками.

Етапи відповідного дослідження відображено на рис. 1.

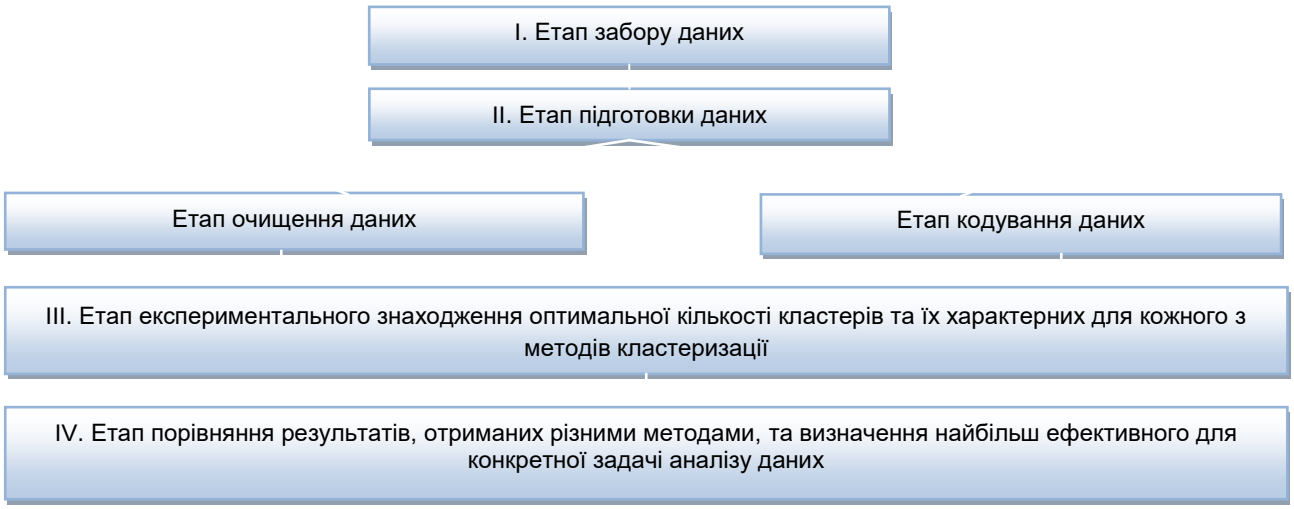

Рис. 1. Етапи кластеризації даних змішаного типу

Джерело: сформовано автором

Одним із найвідповідальніших моментів є *збір даних (I етап),* адже саме форма забору даних впливатиме на кількісні та якісні показники подальшої роботи.

Саме тому, для початку відповідного дослідження було розроблено інноваційну анкету-опитувальник «Цифрова трансформація бізнес-структур» [7], проведено опитування за допомогою сервісу Google Forms представників керівних позицій у компаніях та фірмах різних форм власності (ТОВ, ПП, ФОП) та сфер діяльності (будівництво, торгівля, ремонтні послуги, логістика, надання послуг тощо). Респонденти дослідження відповіли на чотири групи запитань, які стосувалися основних аспектів ведення їхнього бізнесу:

– I група запитань: інформативні (організаційна форма, сфери діяльності, ведення експортної-імпортної діяльності, розмір суб'єкта підприємництва та ін.). Дані в подальшому використовуватимуться для глибинного аналізу та вироблення рекомендацій;

– II група запитань: рівень цифрової грамотності людського капіталу організації;

– III група запитань: рівень цифрової зрілості організації (інформатизації бізнес-діяльності, використання цифрових інструментів у своїй роботі, робота з соціальними мережами, сервісами планування, аналітики чи реклами);

– IV група запитань: рівень забезпечення цифровою інфраструктурою (технологічне забезпечення: комп'ютери, мобільні пристрої та швидкість та якість Інтернету).

Зважаючи на специфіку запитуваної інформації та використання різних категорій запитань, було отримано відповіді у вигляді як якісних, так і категорійних (кількісних) даних. Наприклад, відомості щодо кількості працівників організації було отримано у вигляді натуральних чисел, а інформація щодо наявності бізнес-моделі представлялась бінарною відповіддю – «так» або «ні». Деякі запитання передбачали відкриті відповіді щодо ставлення анкетованого до тої чи іншої проблеми, пов'язаної з інформатизацією бізнес-структури. Такі відповіді були виключені із загального набору даних для кластеризації та не бралися для розрахунку кінцевого індексу цифрової трансформації бізнесу. Проте, відповідні дані були включені в групи інформативних факторів для подальшого опрацювання.

При заборі даних було виявлено численні механічні помилки та порожні значення у відповідях. Ці проблеми було вирішено з допомогою ручної обробки. Проте, при збільшенні кількості опитуваних така обробка вимагатиме уніфікації можливих варіантів відповідей для кожного із запитань або приведення усіх можливих відповідей лише до вибору із запропонованих.

Наступний, не менш важливий *II етап дослідження – підготовка даних*.

Підготовка даних складалась з етапів очищення та кодування даних, пропущених значень в даному дослідженні виявлено не було.

Очищення даних. Як ієрархічний агломеративний алгоритм, так і EM-метод було запущено для одного і того ж набору даних, тому попередню обробку через очищення даних було виконано однаково для обох методів.

Відповіді на відкриті запитання було приведено до визначеного шаблону, наприклад, лише «так» або «ні», або уніфіковано інакше. Атрибути виду «автоматично обчислений час проходження анкетування» чи «особисте ставлення респондента» було помічено як інформаційні і вилучено з вхідних даних задачі. Усі маніпуляції було виконано в ручному режимі завдяки невеликій розмірності задачі.

Кодування даних. Використання надбудови MS Excel не вимагає спеціальної підготовки даних та приймає на вхід звичайний аркуш електронної таблиці зі значеннями будь-якого типу. Тому для кластеризації за допомогою надбудови Data Mining в середовищі MS Excel кодування даних не проводилось.

Для роботи з бібліотеками машинного навчання Python кодування даних було необхідним, оскільки більшість алгоритмів використовують математичні операції над даними кількісного типу. Для тих запитань, де було можливим розставити відповіді у порядку «більше-менше» або «краще-гірше», було використано ранжоване кодування значень. Відповіді кодувалися від 0 до деякого додатного числа, де 0 означав однозначну відповідь «ні» або число, близьке до 0, а інші значення були проранжовані відповідно до зростання прояву ознаки.

Варіанти відповідей, не придатні до ранжування, є даними номінального типу і були позначено деякими числами-символами. Для подальшої обчислювальної роботи алгоритму з такими даними було використано метрику Говера, яка дає можливість працювати як з кількісними, так і з категорійними числовими даними одночасно.

У дослідженні використовувалось два способи кластеризації:

1) за допомогою надбудови Data Mining для електронних таблиць MS Excel. Можливості кластеризації в середовищі MS Excel представлено ітеративними алгоритмами: k-means та Expectation-Maximization. За опорний було визначено саме EM-алоритм [4; 8];

2) за допомогою функцій бібліотек для машинного навчання мови програмування Python [4].

Для опису роботи з двома алгоритмами уведено позначення: *N* респондентів $U = \{\overrightarrow{u_1}, \overrightarrow{u_2}, \dots, \overrightarrow{u_N}\}$ та *M* запитань $Q = \{q_1, q_2, \dots, q_M\}$. Кожен учасник $\overrightarrow{u_i} \in U$ $(i \in \overline{1,N})$ відповів на кожне із запитань $q_k \in Q$ $(k \in \overline{1,M})$, тому в результаті отримано матрицю відповідей розмірністю $(N \times M)$, у якій кожен респондент представлений у вигляді наступного кортежу: $\overrightarrow{u_i} = \{u_{i1}, u_{i2}, \dots, u_{ik}, \dots, u_{iM}\}$, де u_{ik} є відповіддю *i*-го опитаного на *k*-те запитання (Рис. 2). Надалі такий кортеж називатимемо точкою [9].

Респонденти	Запитання						
	q_1	q_2	q_3	…	q_k	…	q_M
$\overrightarrow{u_1} =$	u_{11}	u_{12}	u_{13}	…	u_{1k}	…	u_{1M}
$\overrightarrow{u_2} =$	u_{21}	u_{22}	u_{23}	…	u_{2k}	…	u_{2M}
…	…	…	…	…	…	…	…
$\overrightarrow{u_N} =$	u_{N1}	u_{N2}	u_{N3}	…	u_{Nk}	…	u_{NM}

Рис. 2. Матриця відповідей

Джерело: розраховано та структуровано автором

Для виконання III етапу дослідження необхідно розглянути принципи роботи вибраних методів кластеризації.

Метод ієрархічної агломерації. Детально принцип роботи модифікованого агломеративного методу описано у [9]. Згідно з агломеративним підходом до кластеризації, на початку кожна точка вважається окремим кластером. Під час роботи алгоритму на кожному кроці два найближчих кластери об'єднуються, в кінцевому результаті утворюючи наперед визначену кількість кластерів або зливаючись в один. Для початку роботи агломеративного алгоритму будують матрицю попарних відстаней між об'єктами кластеру. У контексті задачі для обчислення матриці відстаней було використано метрику Говера (1), запропоновану у [9]:

$$d\left(\overrightarrow{u_i}, \overrightarrow{u_j}\right) = \frac{1}{M}\sum_{k=1}^{M} d_{ijk}, \tag{1}$$

де $d_{ijk} = d(u_{ik},\ u_{jk})$ – відстань між відповідями в *k*-му запитанні;

M – кількість відповідей на запитання у кортежі.

Матриця відстаней D_k для *k*-го запитання є симетричною:

$D_k =$

0	d_{12k}	d_{13k}	…	d_{1Nk}
	0	d_{23k}	…	d_{2Nk}
		0	…	d_{3Nk}
			…	…
			…	0

Симетрична матриця D відстаней між окремими точками кластеру має вигляд:

$D =$

0	$d(\overrightarrow{u_1}, \overrightarrow{u_2})$	$d(\overrightarrow{u_1}, \overrightarrow{u_3})$	…	$d(\overrightarrow{u_1}, \overrightarrow{u_N})$
	0	$d(\overrightarrow{u_2}, \overrightarrow{u_3})$	…	$d(\overrightarrow{u_2}, \overrightarrow{u_N})$
		0	…	$d(\overrightarrow{u_3}, \overrightarrow{u_N})$
			…	…
			…	0

Елементами матриці D є усереднені значення попарних значень відстаней, обчислених за формулою (1). Всі ваги запитань опитувальника взято рівними 1.

Спосіб знаходження відстаней d_{ijk} залежить від типу даних у *k*-му запитанні. Якщо u_{ik} та u_{jk} кількісні, то відстань d_{ijk} виражається формулою:

$$d_{ijk} = \frac{|u_{ik} - u_{jk}|}{\max(u_k) - \min(u_k)} \tag{2}$$

Причому, в цьому випадку $d\left(\overrightarrow{u_i}, \overrightarrow{u_j}\right) \in [0;1]$. Якщо u_{ik} та u_{jk} – категорійні дані з неможливістю впорядкування (nominal data), то відстань обчислюється за формулою (3):

$$d_{ijk} = \begin{cases} 0, & u_{ik} = u_{jk}, \\ 1, & u_{ik} \neq u_{jk}. \end{cases} \tag{3}$$

В обох випадках $d_{ijk} = 0$ означає ідентичність відповідей респондентів u_i та u_j в *k*-му запитанні, а $d_{ijk} = 1$ – максимальну відмінність. Як наслідок, для усереднених відстаней, обчислених за формулою (1), усі значення $d\left(\overrightarrow{u_i}, \overrightarrow{u_j}\right) \in [0;1]$.

Відстань між окремими кластерами знаходилась за методом дальнього сусіда. Кластери, які є найближчими за обраною метрикою, об'єднуються, відстані від новоствореного до інших кластерів знову перераховуються, матриця відстаней автоматично оновлюється і об'єднання кластерів

продовжується. Метод дальнього сусіда дозволяє виділити досить компактні та стійкі структури, що відповідає поставленій задачі.

На противагу запропонованій модифікації агломеративного методу, серед ітеративних алгоритмів було вибрано метод Expectation-Maximization (EM-алгоритм нечіткої кластеризації), представлений у надбудові Data Mining для Microsoft Excel. У цьому випадку основною ідеєю методу є припущення, що елементи вхідної множини даних є незалежними випадковими величинами, розподіленими за тим чи іншим законом, у більшості випадків – нормальним Гаусівським розподілом [10; 11].

При використанні EM-методу вважається, що будь-який об'єкт з набору даних належить до всіх кластерів з різною ймовірністю. Перед початком роботи алгоритму задається число кластерів *K* та початкові наближені параметри для кожного з *K* розподілів вхідних даних. Під час виконання ітерацій відбувається покрокове покращення параметрів розподілів до заданого рівня точності моделі. Після завершення роботи алгоритму кожен об'єкт буде віднесено до кластеру, ймовірність належності до якого є максимальною. Таким чином на кожній з ітерації виконуються два послідовні кроки:

1) Expectation – обчислення ймовірності (міри правдоподібності) належності точок до кожного з кластерів;

2) Maximization – покращення значень параметрів розподілів з метою максимального збільшення ймовірностей належності точок до кластерів.

Налаштування параметрів алгоритмів. Майстер задачі «Cluster» в MS Excel дозволяє вибрати потрібні параметри та налаштувати їхні значення [10]. Для даної задачі у налаштуваннях алгоритму було змінено перелік запитань, що впливатимуть на результат, задано значення кількості кластерів та cluster seed EM-методу кластеризації.

Виклик методу sklearn.cluster.AgglomerativeClustering для створення моделі кластеризації з допомогою Python у найбільш загальному випадку передбачає задання 3-х параметрів: кількості кластерів, intra-cluster distance та inter-cluster distance metrics. Матриця метрики відстаней між елементами може бути однією із запропонованих у [9] або бути обчисленою інакше.

Виклик функції створення моделі кластеризації:

```
Model = AgglomerativeClustering(number_of_clusters = m,
metric = "precomputed", linkage = complete)                    (4)
labels = model.fit_predict(distances),
```

де m – наперед визначена кількість кластерів;

distances – матриця відстаней, попередньо обчислена за метрикою Говера (1) – (3).

Розглянемо результати кластеризації кожним із методів та порівняємо отримані результати (IV етап). На виході кластеризації у надбудові Data Mining середовища MS Excel отримується розбиття набору даних на кластери з можливістю візуалізації, перегляду статистичної інформації та профілів кластерів.

При використанні методів кластеризації бібліотеки sklearn для аналізу даних на Python на виході отримується одновимірний числовий масив, що позначає до якого кластеру належить кожен кортеж із вхідних даних. Подальший аналіз та візуалізація отриманого результату здійснюється додатково. Алгоритм підбору оптимального числа кластерів описано в [9].

Результати ієрархічно-агломеративної кластеризації.

Матриця відстаней *D* між точками вхідної множини набула вигляду відображеного на рис. 3.

	1	2	3	4	5	…	34
1	0	0.34	0.26	0.46	0.36	…	0.57
2	0.34	0	0.15	0.33	0.36	…	0.49
3	0.26	0.15	0	0.24	0.27	…	0.52
4	0.46	0.33	0.24	0	0.37	…	0.41
5	0.36	0.36	0.27	0.37	0	…	0.47
…	…	…	…	…	…	…	…
34	0.57	0.49	0.52	0.41	0.47	…	0

Рис. 3. Матриця відстаней *D* між точками вхідної множини

Джерело: розраховано та структуровано автором

Порівняльний аналіз отриманих кластерів за основними характеристиками структуровано у табл. 1. Значення у відсотках позначають частку респондентів кожного кластеру, які однаково відповіли на певне запитання.

Таблиця 1

Порівняльна характеристика кластерів, утворених за допомогою агломеративної кластеризації засобами Python

Запитання	Кластер 1 (16)		Кластер 2 (5)		Кластер 3 (2)		Кластер 4 (10)		Кластер 5 (1)	
Організаційна структура	ФОП	50, 0 %	ТОВ	60,0 %	ФОП	100%	ТОВ	100%	ТОВ	100%
Наявність сайту	Ні	62,5 %	Так	100%	Так	100 %	Так/Ні	по 50%	Так	100%
Робота ланцюжка корзина-покупець	Не працює	62,5 %	Так	60,0 %	Так	100%	Ні	90%	Так	100%
Оптимізація сайту	Ні	87,5 %	Частково оптимізовано	60,0 %	Ні/спеціаліст	по 50%	Не оптимізовано	70%	Оптимізовано спеціалістом	100%
Наявність бізнес-сторінки ФБ/Інстаграм	Не має	81,25 %	Є	80,0 %	Так, частково	по 50,0 %	Так	80,0 %	Так	100%
Ведення сторінки ФБ/Інстаграм	Не ведуть	50,0 %	Ведеться самостійно	60,0 %	Маркетолог	100%	Ведеться самостійно	60,0 %	Маркетолог	100%
Використання реклами ФБ	Не використовували	93,75 %	Використовували без результату / не використовували	по 40,0 %	Використовували, хороший результат	100%	Не використовували	60,0 %	Використовували, хороший результат	100%
Використання Google Ads	Не використовували	100%	Використовували, хороший результат / не використовували	по 40,0 %	Використовували, хороший/поганий результат	по 50,0 %	Не використовували	100%	Використовували, поганий результат	100%
Використання Google Analytics	Не використовували	87,5 %	Використовують та аналізують / не використовують	по 40,0 %	Використовують та аналізують / не використовують	по 50,0 %	Не використовували	80,0 %	Використовують та аналізують	100%
Спеціалізована ERP-система	Не використовували	50,0 %	Не використовували	100%	Так	100%	Не використовували	90,0 %	Так	100%
Спеціалізована CRM-система	Не використовували	100%	Не використовували	60%	Так	100%	Не використовували	100%	Так	100%
Спеціалізовані додатки	Не використовували	100%	Не використовували	100%	Ні	100%	Не використовували	90,0 %	Так	100%
Чи можна замовити товар/послугу через Інтернет?	Не можна	81,25 %	Так	100%	Так	100%	Ні	50,0 %	Так	100%

Джерело: розраховано та структуровано автором

Як видно із табл.1, число респондентів, що відповіли однаково на вибрані запитання варіюється від 40 до 100%.

Згідно з результатами, найбільшу кількість респондентів (16) було віднесено до першого кластера, основними характеристиками якого є:

– відсутність досвіду роботи з будь-якими цифровими інструментами;

– відсутність компаній в Інтернет-середовищі;

– неефективність сайту за його наявності.

Другий кластер було сформовано 5-ма компаніями, основні характеристики якого визначаються наступним чином:

– наявність компаній в Інтернет-просторі;

– використання простих інструментів у неповному обсязі або самостійно.

Ще 2 респонденти утворили третій кластер, характеристиками якого є:

– ефективне функціонування сайту та ланцюжка покупки;

– використання більшості цифрових інструментів включно з рекламою;

– робота маркетолога над просуванням бренду чи продукту.

Четвертий кластер складається з 10 опитаних і його характерними ознаками є:

– відсутність функціонування ланцюжка покупки на сайті;

– невикористання складних цифрових інструментів;

– наявність лише соціальних мереж.

П'ятий кластер складається лише з однієї компанії, яка успішно використовує практично усі цифрові інструменти за допомогою спеціалістів.

Достатня ступінь відмінностей між кластерами та достатня ступінь схожості елементів всередині кластеру (80–100%) дає змогу чітко виділити наступні групи та ранжувати їх за рівнем використання цифрових технологій та інструментів у бізнес-діяльності. У табл. 2 показано ранжування типів бізнес-структур за спаданням цифрової зрілості.

Щодо кластеризації, здійсненої EM-методом у середовищі MS Excel, то вона виявилась нестійкою. Оскільки EM-алгоритм представляє групу ітеративних методів нечіткої кластеризації, такий результат є нормальним та придатним для використання у певному класі задач.

Таблиця 2

Ранжування кластерів бізнес-структур за рівнем цифрової зрілості

№ кластеру	Характеристика кластеру
I	компанії, що використовують майже усі передові цифрові технології, включно з технологіями аналізу даних
II	компанії, що використовують складніші цифрові інструменти, такі як ERP та CRM
III	компанії, що використовують деякі цифрові інструменти самостійно (SEO, соціальні мережі, реклама)
IV	компанії з обмеженим використанням лише одного інструменту – соціальних мереж
V	компанії, що не використовують цифрові технології

Джерело: сформовано автором

Порівняльну характеристику отриманих кластерів відображено у табл. 3.

Таблиця 3

Порівняльна характеристика кластерів отриманих EM-методом

Запитання	**Кластер 1 (8)**		**Кластер 2 (9)**		**Кластер 3 (9)**		**Кластер 4 (6)**		**Кластер 5 (2)**	
1	2	3	4	5	6	7	8	9	10	11
Організаційна структура	ФОП	87,5 %	ТОВ	66,6 %	ФОП	55,5 %	ТОВ	66,6 %	ТОВ	100%
Наявність сайту	Так	62,5 %	Ні	55,5 %	Ні	55,5 %	Так	66,6 %	Так	100%
Робота ланцюжка корзина-покупець	Ні	62,5 %	Ні	88,8 %	Ні	88,8 %	Ні	100%	Так	100%
Оптимізація сайту	Ні	50,0 %	Ні	77,7 %	Ні	77,7 %	Ні	100%	Оптимізовано спеціалістом	100%
Наявність бізнес-сторінки ФБ/Інстаграм	Так	62,5 %	Ні	77,7 %	Так, частково	55,5 %	Частково	50,0 %	Так	100%

продовження табл. 3

1	2	3	4	5	6	7	8	9	10	11
Ведення сторінки ФБ/Інстаграм	Працює добре (самості йно або маркето лог)	50,0 %	Не використ товує-ться / працює погано	44,4 %	Працю є без плану	55,5 %	Працює без плану	55,5 %	Марке-толог	100%
Використання реклами ФБ	Не викори стовув али	50,0 %	Не викори стовув али	100%	Не викори стовув али	77,7 %	Не використ овували	50,0 %	Викори стовува ли (добре/ погано)	по 50,0 %
Використання Google Ads	Не викори стовув али	75,0 %	Не викори стовув али	100%	Не викори стовув али	88,8 %	Не використ овували	83,0 %	Викори стовува ли (не спрацю вало)	по 50,0 %
Використання Google Analytics	Не викори стовув али	62,5 %	Не викори стовув али	77,7 %	Не викори стовув али	100 %	Не використ овували	66,6 %	Викори стовую ть та аналізу ють	100%
Спеціалізована ERP-система	Не викори стову-вали	75,0 %	Не викори стову-вали	77,7 %	Не викори стову-вали	55,5 %	Не викори-стову-вали	66,6 %	Так / частко во	по 50%
Спеціалізована CRM-система	Не викори стову-вали	75,0 %	Не викори стову-вали	100%	Не викори стову-вали	88,8 %	Не викори-стову-вали	100%	Так	100%
Спеціалізовані додатки	Не викори стову-вали	100%	Не викори стову-вали	100%	Не викори стову-вали	88,8 %	Не викори-стову-вали	100%	Так / Ні	по 50%
Чи можна замови-ти товар/послугу через Інтернет?	Так	62,5 %	Ні	55,5 %	Ні	77,7 %	Так / Ні	50,0 %	Так	100%

Джерело: розраховано та структуровано автором

На противагу до агломеративної кластеризації, ступінь однаковості відповідей на запитання усередині кластерів є набагато нижчою, і в середньому коливається у межах 60–70%. Як бачимо, невисокою є також міра відмінності кластерів між собою. Багатократне повторення застосування ЕМ-методу не покращило якості результатів.

Аналогічно до попереднього методу кластеризації, виділено кластер, до якого ввійшли дві бізнес-структури, що активно використовують цифрові технології у своєму бізнесі. Відмінності між іншими кластерами є незначними, різниці у відсотковому співвідношенні відповідей на запитання є мінімальними, тому неможливо виділити характерні риси кожної підмножини. Неможливість виокремити кластери з чітко вираженими особливостями не відповідає поставленій у дослідженні задачі.

Висновки з проведеного дослідження. У даному дослідженні було проведено експериментальне порівняння застосування двох підходів до кластеризації респондентів за результатами онлайн-анкетування з допомогою сервісу Google Forms – hard and soft кластеризації. Hard кластеризацію було реалізовано засобами Python із застосуванням ієрархічного агломеративного методу, soft – з допомогою надбудови Data Mining середовища MS Excel та застосуванням ітеративного ЕМ-методу.

Порівняльний аналіз результатів, отриманих двома методами показав такі результати:

– з допомогою ієрархічної агломеративної кластеризації отримано 5 кластерів, достатньо відмінних один від одного та з високим ступенем схожості між елементами кластеру (60–100% залежно від запитання). Виокремлено характерні риси кластерів (використання соціальних мереж, рекламних кабінетів та послуг, аналітичних інструментів, пошукової оптимізації сайтів тощо);

– використання ЕМ-методу не дозволило отримати хороші результати кластеризації та досягнути мети задачі, результати виконання ЕМ-методу змінювалися з кожним запуском алгоритму.

Дослідним шляхом встановлено, що метод агломеративної ієрархічної кластеризації є ефективним методом для вирішення задачі кластеризації даних змішаного типу, отриманих в результаті опитування респондентів.

Окрім вдосконалення параметрів роботи алгоритму, задачами до подальшого вирішення залишаються: усунення механічних помилок при введенні відповідей та наявність порожніх значень;

різнотипність даних, що спричиняє складність їх уніфікації та коректного впорядкування; підбір математичних метрик, що використовуються як аргументи функцій виконання кластеризації та обчислення її якості.

Література

1. Черезов Д. С., Тюкачев Н. А. Обзор основных методов классификации и кластеризации данных. *Вестник ВГУ. Серия: системный анализ и информационные технологии.* 2009. № 2. С. 25-29. URL: http://www.vestnik.vsu.ru/pdf/analiz/2009/02/2009-02-05.pdf (дата звернення: 14.08.2019).

2. Clustering Online Poll Data: Towards a Voting Assistance System / F. Mendes, I. Katakis, N. Tsapatsoulis, C. Tziouvas, V. Triga. *Seventh International Workshop on Semantic and Social Media Adaptation and Personalization*. 2012. URL: https://www.researchgate.net/publication/261486679_Clustering_Online_Poll_Data_Towards_a_Voting_Assistance_System (дата звернення: 14.10.2019).

3. McCaffrey J. Machine Learning Using C#Succinctly. Syncfusion. 2014. 148 p. URL: https://www.syncfusion.com/ebooks/machine/k-means-clustering (дата звернення: 10.10.2019).

4. Python. URL: https://www.python.org (дата звернення: 18.10.2019).

5. Scikit. Clustering documentation / Scikit learn. URL: https://scikit-learn.org/stable/modules/clustering.html (дата звернення: 12.10.2019).

6. Зацерковний В. І., Бурачек В. Г., Железняк О. О., Терещенко А. О. Геоінформаційні системи і бази даних : монографія. Ніжин : НДУ ім. М. Гоголя, 2017. Кн. 2. 237 с.

7. Google Forms. URL: https://www.google.com/intl/uk_ua/forms (дата звернення: 17.10.2019).

8. Microsoft. Cluster Wizard (Data Mining Add-ins for Excel), Microsoft Docs. (2017 Dec). URL: https://docs.microsoft.com/en-us/sql/analysis-services/cluster-wizard-data-mining-add-ins-for-excel?view=sql-server-2014 (дата звернення: 13.10.2019).

9. Small and Medium Business Structures Clustering Method Based on Their Digital Maturity / I. Strutynska, H. Kozbur, L. Dmytrotsa, I. Bodnarchuk and O. Hlado. *International Scientific-Practical Conference Problems of Infocommunications. Science and Technology,* (October 10–11, 2019). 2019. P. 278-282. URL: http://www.dut.edu.ua/uploads/n_7589_67076384.pdf (дата звернення: 15.10.2019).

10. Cluster analysis / Wikipedia. URL: https://en.wikipedia.org/wiki/Cluster_analysis (дата звернення: 17.10.2019).

11. Expectation-maximization algorithm / Wikipedia. URL: https://en.wikipedia.org/wiki/Expectation-maximization_algorithm (дата звернення: 12.10.2019).

References

1. Cherezov, D.S. and Tiukachev, N.A. (2009), "Overview main methods of data classification and clustering", *Vestnik VGU*. *Seria: Sistemnyy analiz i informatsionnye tekhnologii,* no. 2, pp. 25-29, available at: http://www.vestnik.vsu.ru/pdf/analiz/2009/02/2009-02-05.pdf (access date August 14, 2019).

2. Mendes, F., Katakis, I., Tsapatsoulis, N., Tziouvas, C. and Triga, V. (2012), "Clustering Online Poll Data: Towards a Voting Assistance System", *Seventh International Workshop on Semantic and Social Media Adaptation and Personalization*, available at: https://www.researchgate.net/publication/261486679_Clustering_Online_Poll_Data_Towards_a_Voting_Assistance_System (access date October 14, 2019).

3. McCaffrey, J. (2014), Machine Learning Using C#Succinctly, Syncfusion, 148 p., available at: https://www.syncfusion.com/ebooks/machine/k-means-clustering (access date October 10, 2019).

4. Python, available at: https://www.python.org (access date October 18, 2019).

5. Scikit. Clustering documentation, Scikit learn, available at: https://scikit-learn.org/stable/modules/clustering.html (access date October 12, 2019).

6. Zatserkovnyi, V.I., Burachek, V.H., Zhelezniak, O.O. and Tereshchenko, A.O. (2017), *Heoinformatsiini systemy i bazy danykh* [Geoinformation systems and databases], monograph, NDU im. M. Hoholia, Nizhyn, Ukraine, Book 2, 237 p.

7. Google Forms, available at: https://www.google.com/intl/uk_ua/forms (access date October 17, 2019).

8. Microsoft (2017), Cluster Wizard (Data Mining Add-ins for Excel), Microsoft Docs, available at: https://docs.microsoft.com/en-us/sql/analysis-services/cluster-wizard-data-mining-add-ins-for-excel?view=sql-server-2014 (access date October 13, 2019).

9. Strutynska, I., Kozbur, H., Dmytrotsa, L., Bodnarchuk, I. and Hlado, O. (2019), "Small and Medium Business Structures Clustering Method Based on Their Digital Maturity", *International Scientific-Practical Conference Problems of Infocommunications. Science and Technology*, (October 10–11), pp. 278-282, available at: http://www.dut.edu.ua/uploads/n_7589_67076384.pdf (access date October 15, 2019).

10. Cluster analysis, Wikipedia, available at: https://en.wikipedia.org/wiki/Cluster_analysis (access date October 17, 2019).

11. Expectation-maximization algorithm, Wikipedia, available at: https://en.wikipedia.org/wiki/Expectation-maximization_algorithm (access date October 12, 2019).

Стаття надійшла до редакції 30.10.2019 р.