Technical sciences
BULLETIN OF POLTAVA STATE AGRARIAN ACADEMY
ISSN: 2415-3354 (Print)
2415-3362 (Online)
https://journals.pdaa.edu.ua/visnyk

original article | UDC 004.4'2: 631.526.3 | **doi: 10.31210/visnyk2019.02.35**

COMPARATIVE ANALYSIS OF CLUSTERING METHODS SUITABLE FOR PLANT VARIETIES MORPHOLOGICAL CHARACTERISTICS DATA PROCESSING

N. S. Orlenko,
ORCID ID: 0000-0003-4103-7806, E-mail: sops@sops.gov.ua,
K. M. Mazhuha,
ORCID ID: 0000-0002-1434-8687,
M. B. Dushar,
ORCID ID: 0000-0002-2601-5564,
V. V. Maslechkin,
ORCID ID: 0000-0002-6246-4287,
Ukrainian Institute of Plant Variety Expert Examination, 15, Henerala Rodymtseva str., Kyiv, 03041, Ukraine

Despite the fact that clustering is uncontrolled classification of multi-dimensional data in corresponding clusters, the clustering problem has been addressed in many contexts and by researchers in many subjects. One of the research areas, where clustering is useful, is morphological analysis of plant variety characteristics, which helps to identify new varieties more accurately. That is why it is important to compare the results of clustering and the using of other methods and measure intervals in order to determine the most suitable methods for morphological characteristics analysis. The following methods were used during the research: analytical, mathematical, statistical, and graphic. This paper presents a comparative analysis of clustering methods using the famous Fisher's Iris data set and also the classification methods, which are the most suitable for analyzing morphological characteristics of plant varieties. As a result, this paper presents a survey of better plant varieties clustering results influenced by different hierarchical agglomerative classification methods (Between-Groups Linkage, Within Groups, Nearest Neighbor, Furthest Neighbor, Centroid Method, Median Method, Wards Method) using Euclidean and non-Euclidean measure intervals. Clustering results were evaluated by using descriptive statistics methods (cross-tables). Some clustering algorithms and technologies, which we used during the research, were also described. The article considers possible measure interval which is used in algorithms, and presents the most popular clustering algorithms and shows their role in the Data mining. Numerous techniques and clustering algorithms were suggested earlier to assist clustering of time series data streams. The clustering algorithms and their effectiveness in various applications are compared to recognize the most suitable method to solve the existing problem of morphological analysis and new plant varieties identification. The best results were obtained using Average Linkage (Between Groups) with Pearson Correlation measure interval, Average Linkage (Within Group) with Cosine measure interval, Average Linkage (Within Group) with Pearson Correlation measure interval, Ward Method with Cosine measure interval. Frequency statistics (cross-tables) to evaluate the quality of classification results was suggested. Thus, the conducted testing proved that there is no universal algorithm that would ideally distribute the set of Fisher's Irises to clusters. Therefore, clustering of plant varieties should be carried out iteratively, consistently applying the most common clustering algorithms and carefully evaluating clustering results in order to select the method and measure interval, which classify plant varieties most optimally and enable to interpret the classification results correctly.

Key words: *hierarchical agglomerative methods, measure interval, Fisher's Iris data set, classification, cross-tables.*

ПОРІВНЯЛЬНИЙ АНАЛІЗ ІЄРАРХІЧНИХ МЕТОДІВ КЛАСТЕРІЗАЦІЇ, ПРИДАТНИХ ДЛЯ ОБРОБЛЕННЯ ДАНИХ МОРФОЛОГІЧНИХ ОЗНАК СОРТІВ РОСЛИН

Н. С. Орленко, К. М. Мажуга, М. Б. Душар, В. В. Маслечкін,
Український інститут експертизи сортів рослин, вул. Генерала Родимцева, 15, м. Київ, 03041, Україна

Незважаючи на те, що кластеризація є безконтрольною класифікацією багатовимірних даних у відповідні кластери, застосування кластерного аналізу під час дослідження морфологічних характеристик сортів рослин дозволяє зменшити розмірність вибірки даних, що сприяє більш точній ідентифікації нових сортів. Саме тому важливим питанням є порівняння результатів кластеризації із застосуванням різних методів і метрик та виявлення найбільш придатних для аналізу морфологічних характеристик. Методи: аналітичний, математичний, статистичний, графічний. Під час виконання досліджень використано широко відомий набір даних, що має назву Іриси Фішера. Результати. Досліджено вплив на результат кластерного аналізу різних ієрархічних агломеративних методів класифікації (ближнього сусіда, дальнього сусіда, середнього зв'язку, середнього сусіда (центроїда) та метода Варда) із застосуванням евклідових та не евклідових метрик. Оцінено результати кластерізації з використанням засобів описової статистики (методу перехресних таблиць). Встановлено, що найбільш придатними для проведення кластеризації за морфологічними характеристиками для наборів даних, які описуються метричними шкалами є методи: середнього зв'язку (між групами) із застосуванням кореляції Пірсона, середнього зв'язку (всередині групи) із застосуванням метрик Косінус та кореляції Пірсона, а також методу Варда із застосуванням метрики Косінус. Запропоновано використовувати апарат частотної статистики (перехресні таблиці) для оцінювання якості результатів класифікації. Висновки. Проведене тестування довело, що не існує жодного універсального алгоритму, який би ідеально розподілив набір Ірисів Фішера на кластери. Не зважаючи на те, що встановлено методи й метрики, які є найбільш вдалими для класифікації протестованого набору даних, ці методи не можна рекомендувати для використання під час тестування морфологічних ознак усіх ботанічних таксонів. Кластеризацію сортів рослин потрібно проводити ітераційно, послідовно застосовуючи найбільш поширені алгоритми кластеризації та ретельно оцінювати результати кластеризації з метою вибору метода та метрики, які найбільш оптимально класифікують сорти рослин та дозволять правильно інтерпретувати результати класифікації. Результати такої кластеризації рекомендовано оцінювати з використанням методу перехресних таблиць та обирати кращий за якістю кластерів.

Ключові слова: *ієрархічні агломеративні методи, метрика, набір даних Іриси Фішера, класифікація, перехресні таблиці.*

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ИЕРАРХИЧЕСКИХ МЕТОДОВ КЛАСТЕРИЗАЦИИ, ПРИГОДНЫХ ДЛЯ ОБРАБОТКИ ДАННЫХ МОРФОЛОГИЧЕСКИХ ПРИЗНАКОВ СОРТОВ РАСТЕНИЙ

Н. С. Орленко, К. Н. Мажуга, М. Б. Душар, В. В. Маслечкин,
Украинский институт экспертизы сортов растений, ул. Генерала Родимцева, 15, г. Киев, 03041, Украина

Несмотря на то, что кластеризация является неконтролируемой классификацией многомерных данных в соответствующие в кластеры, применение кластерного анализа позволяет уменьшить размерность выборки данных при исследовании морфологических характеристик сортов растений. Это, в свою очередь, способствует более точной идентификации новых сортов. Именно поэтому важным вопросом является сравнение результатов кластеризации с применением различных методов и метрик и выявление наиболее подходящих для анализа морфологических характеристик. В ходе исследования использованы аналитический, математический и статистический методы. Во время выполнения исследований использован широко известный набор данных – Ирисы Фишера. Выполнен анализ влияние на результат кластерного анализа различных иерархических методов классификации с использованием евклидовых и неевклидовых метрик. Установлено, что наиболее пригодными для проведения кластеризации по морфологическим характеристикам для наборов данных, которые описываются метрическими шкалами являются методы: средней связи (между группами) с применением корреляции Пирсона, средней связи (внутри группы) с применением метрик косинус и корреляции Пирсона, а также ме-

тод Варда с применением метрики Косинус. Предложено использовать аппарат частотной статистики (перекрестные таблицы) для оценивания качества результатов классификации. Вывод. Кластеризацию сортов растений следует проводить итерационно, последовательно применяя наиболее распространенные алгоритмы кластеризации и тщательно оценивая результаты кластеризации с целью выбора метода и метрики, которые наиболее оптимально классифицируют сорта растений, что и позволят верно интерпретировать результаты классификации.

Ключевые слова: *иерархические агломеративные методы, метрика, набор данных Ирисы Фишера, классификация, перекрестные таблицы.*

Вступ

Застосування багатовимірного статистичного аналізу дозволяє більш ефективно реалізовувати систему заходів з охорони сортів рослин шляхом більш точної ідентифікації нових сортів рослин за їх морфологічними ознаками під час проведення кваліфікаційної експертизи на відмінність однорідність та стабільність (ВОС).

Методи класифікації, які придатні для обробки даних польових та лабораторних досліджень сортів рослин, варіюють від дуже простих до надзвичайно складних. Тому варто обережно підходити до вибору статистичних методів, зокрема кластерного. Зазвичай класифікацію проводять за такими етапами: відбирають набір даних об'єктів кластеризації (тут – це набір даних Іриси Фішера), визначають множину змінних для оцінювання об'єктів у вибірці та, у разі необхідності, нормалізації значень змінних (у контексті цього дослідження – це довжина зовнішньої частки оцвітини, ширина зовнішньої частки оцвітини, довжина внутрішньої частки оцвітини та ширина внутрішньої частки оцвітини, обчислення значень міри схожості між об'єктами.

Відмітимо, що після отримання результатів кластеризації необхідно оцінювати якість утворених кластерів. Зазвичай, таке оцінювання проводять інтуїтивно, спираючись на досвід дослідників. Але у практиці кваліфікаційної експертизи сортів на ВОС інтуїтивне судження є недостатньо обґрунтованим. Тому інтуїтивне інтроспективне оцінювання може бути лише застосовано лише для невеликих наборів об'єктів, але великомасштабні експерименти вимагають застосування більш об'єктивного методу.

Під час дослідження застосовано найбільш поширені методи ієрахічного кластерного аналізу. Методи ієрархічної кластеризації докладно описані в наукових статтях Хуі Дінгі, Гоцем Трейчевським, Пітером Шеустернером, Сяоюе Ванем та Еамонном Кеохом [8]. Особливості агломеративних методів наведено авторами: Джайн А., Мурті М., Флінт П. Педро Перева Родрігуз та його співавтори в [13] проаналізували систему кластеризації потокового часового ряду. Сян Ліан та інші в [17, 18] запропонували поліноміальний підхід, на підставі якого виконується прогнозування на основі апроксимованої кривої останніх значень ознак. Використання класичного набору даних Іриси Фішера під час тестування роботи алгоритмів машинного навчання, застосування теорії нечітких множин та кластеризації розглянуті в роботах [1–3, 5, 8–13, 16, 19, 20].

Кластерний аналіз в аграрній сфері широко застосовується вітчизняними та іноземними дослідниками [4, 23]. Але потрібно відмітити, що окремі автори зовсім не згадують або не обґрунтовують у свої роботах, який метод та метрику використано під час кластеризації.

Методи об’єктивного оцінювання результатів розглянуто в публікаціях [4, 21–23].

Авторами статті запропоновано власний підхід до оцінювання якості отриманих кластерів сортів рослин.

Тому *метою* роботи є виявлення методів кластеризації та метрик, які найбільш придатні для аналізу морфологічних ознак сортів рослин та засобів оцінювання результатів кластеризації.

Для досягнення цієї мети було поставлено такі *завдання:*

– визначити вплив обраних методів та метрик на результати кластеризації сортів рослин за морфологічними ознаками на прикладі загальновідомого набору даних «Іриси Фішера»;

– встановити метод оцінювання результатів кластеризації та обґрунтувати доцільність його застосування.

Матеріали і методи досліджень

Під час дослідження був використаний широковідомий набір даних ботанічних таксонів півників, який також називають набором даних Іриси Фішера [15, 16]. Цей набір даних називають Ірисами Андерсона, оскільки Едгар Андерсон зібрав ці дані для кількісного визначення морфологічних ознак

різних сортів півників трьох ботанічних таксонів [7]. Два з трьох видів були зібрані на півострові Гаспе з одного й того самого пасовища однією людиною, для вимірювання застосовувався один й той самий приклад [6]. Набір даних складається з 50 зразків кожного з трьох видів ірису, а саме: півників щетинистих, півників строкатих та півників віргінiка. Набір даних містить п’ять стовпчиків (рис. 1).

	A	B	C	D	E
1	Довжина зовнішньої частки оцвітини	Ширина зовнішньої частки оцвітини	Довжина внутрішньої частки оцвітини	Ширина внутрішньої частки оцвітини	Вид Ірису
2	5,10	3,50	1,40	0,20	Півник щетинисті
3	4,90	3,00	1,40	0,20	Півник щетинисті
4	4,70	3,20	1,30	0,20	Півник щетинисті
5	4,60	3,10	1,50	0,20	Півник щетинисті
50	5,30	3,70	1,50	0,20	Півник щетинисті
51	5,00	3,30	1,40	0,20	Півник щетинисті
52	7,00	3,20	4,70	1,40	Півник строкатий
53	6,40	3,20	4,50	1,50	Півник строкатий
54	6,90	3,10	4,90	1,50	Півник строкатий
55	5,50	2,30	4,00	1,30	Півник строкатий
56	6,50	2,80	4,60	1,50	Півник строкатий
57	5,70	2,80	4,50	1,30	Півник строкатий
58	6,30	3,30	4,70	1,60	Півник строкатий
59	4,90	2,40	3,30	1,00	Півник строкатий
60	6,60	2,90	4,60	1,30	Півник строкатий
103	5,80	2,70	5,10	1,90	Півник віргінiка
104	7,10	3,00	5,90	2,10	Півник віргінiка
105	6,30	2,90	5,60	1,80	Півник віргінiка
106	6,50	3,00	5,80	2,20	Півник віргінiка
107	7,60	3,00	6,60	2,10	Півник віргінiка

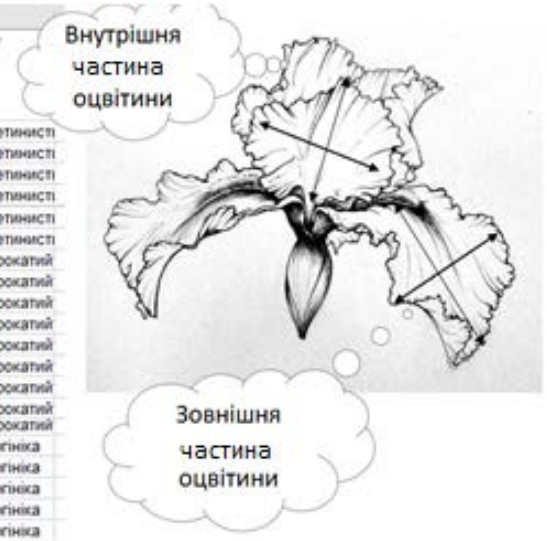

Рис. 1. Набір даних Іриси Фішера

У перших чотирьох стовпчиках записані морфологічні ознаки для кожного зразка півників, а саме: довжина і ширина зовнішньої частини оцвітини та внутрішньої частки оцвітини в сантиметрах. У п'ятому стовпчику зазначено вид півника.

Під час кластеризації набору даних, із використанням агломераційних методів, міри схожості між кластерами можуть бути написані за допомогою формули Ленса-Вільямса (1):

$$d(i\,j,k) = a_i\, d(i,k) + a_j\, d(j,k) + b\,d(i,j) + c\,|d(i,k) - d(j,k)|. \qquad (1)$$

Одиночний зв'язок (Найближчий сусід)	$ai = aj = 0.5\ ;\ b = 0\ ;\ c = -0.5;$ $d(i+j,k) = \min\{d(i,k), d(j,k)\}.$
Повний зв'язок (Найбільш віддалений сусід)	$a_i = a_j = 0.5\ ;\ b = 0\ ;\ c = 0.5;$ $d(i+j,k) = \max\{d(i,k), d(j,k)\}.$
Зважений центроїдний метод (медіана).	$a_i = a_j = 0.5\ ;\ b = -0.25\ ;\ c = 0.$
Незважене попарне середнє	$a_i = n_i / (n_i+n_j)\ ;\ a_i = n_j / (n_i+n_j)\ ;\ b = c = 0;$ $D(C_i+C_j) = 1 / (n_i n_j) \sum d(a,b).$
Метод Варда	$a_i = (n_i + n_k) / (n_k + n_i + n_j);\ a_j = (n_j + n_k) / (n_k + n_i + n_j);$ $b = (n_k) / (n_k + n_i + n_j);\ c = 0.$

Графічне зображення, що пояснює застосування метрик наведено на рисунку 2.

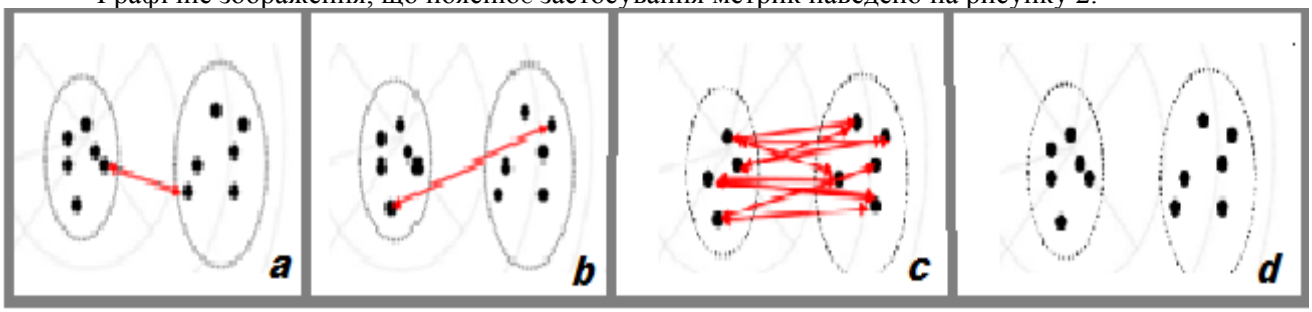

Рис. 2. Графічне відображення застосування метрик

Одиночний зв'язок (рис. 2 a). У цій метриці відстанню між кластерами вважається відстань між двома найбільш близькими об'єктами в різних кластерах.

Повний зв'язок (рис. 2 b.). Під час використання цього метода відстань між кластерами обчислюється аналогічно методу одиночного зв'язку, але замість мінімальної відстані обчислюється максимум.

Середній зв'язок (рис. 2 c). Відстань між двома відмінними один від одного кластерами можна визначити як середню відстань між усіма парами об'єктів з різних кластерів.

Зважений центроїдний метод (рис. 2 d). Цей метод ідентичний попередньому, за винятком того, що під час обчислення використовують вагу під час обчислення різниці між розмірами кластерів.

Метод Уорда: мінімізує суму квадратів критерію (міри неоднорідності), розрахунок проводиться за формулою (2):

$$ESS = \sum_{k=1}^{K} \sum_{x_i \in Ck} \sum_{j=1}^{d} (x_{ij} - x_{kl})^2 \quad (2)$$

Під час тестування ієрархічних методів були використані такі метрики: *Евклідова відстань.* Класична метрика Евкліда, що є геометричною відстанню в багатовимірному просторі, обраховується за формулою (3):

$$D(x,y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2} \quad (3)$$

Квадрат евклідової відстані. Сума квадратів різниці між значеннями для предметів. Ця метрика описується формулою (4):

$$D(x,y) = \sum_{i}^{n} (x_i - y_i)^2 \quad (4)$$

Відстань кореляція Пірсона. Співвідношення між двома векторами значень, що описується формулою (5):

$$r_{x,y} = \frac{\sum_{i=1}^{k} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{k} (x_i - \bar{x})^2 \sum_{i=1}^{k} (y_i - \bar{y})^2\right]^{\frac{1}{2}}} \quad (5)$$

Відстань Косинус. Косинус кута між двома векторами значень визначається за формулою (6):

$$D(x,y) = \frac{\sum_{i=1}^{k} x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}} \quad (6)$$

Відстань Чебишева є максимальною абсолютною різницею між характеристиками двох об'єктів й обчислюється за формулою (7):

$$D(x,y) = \max(|x_i - y_i|) \quad (7)$$

Відстань Мінковського – це корінь суми абсолютних відмінностей між значеннями елементів, що обраховується за формулою (8):

$$D(x,y) = \sqrt{\sum_i^n (x_i - y_i)^2} \quad (8)$$

Розрахунки було проведено з використанням тестової версії статистичного пакету IBM SPSS Statistics 22 (trial version) [9, 21].

Результати досліджень та їх обговорення

Тестування набору даних Іриси Фішера проводилось з використання методів: ближнього сусіда, дальнього сусіда, середнього зв'язку, центроїда, медіани та метода Уорда у комбінаціях з метриками евклідова відстань, квадрат евклідової відстані, кореляція Пірсона, Косинус, Чебишєва та Мінковського (таб. 1).

1. Результати тестування

Метод кластеризації	Метрикаl	Кількість зразків		
		Кластер 1	Кластер 2	Кластер 3
Середнього зв'язку (між групами)	Евклідова відстань	50	64	36
	Квадрат евклідової відстані	50	88	12
	Косінус	49	1	100
	Кореляція Пірсона	50	54	46
	Чебишєва	50	90	10
	Мінковського	50	64	36
Середнього зв'язку (всередині групи)	Евклідова відстань	50	72	28
	Квадрат евклідової відстані	50	72	28
	Косінус	50	53	47
	Кореляція Пірсона	50	54	46
	Чебишєва	50	64	36
	Мінковського	50	72	28
Одинарного зв'язку (ближнього сусіда)	Евклідова відстань	50	98	2
	Квадрат евклідової відстані	50	98	2
	Косінус	49	1	100
	Кореляція Пірсона	49	1	100
	Чебишєва	50	98	2
	Мінковського	50	98	2
Повного зв'язку (даль-нього сусіда)	Евклідова відстань	50	72	28
	Квадрат евклідової відстані	50	72	28
	Косінус	50	74	26
	Кореляція Пірсона	50	28	72
	Чебишєва	50	65	35
	Мінковського	50	72	28
Середнього сусіда (центроїда)	Евклідова відстань	50	98	2
	Квадрат евклідової відстані	50	64	36
	Косінус	50	32	68
	Кореляція Пірсона	50	32	68
	Чебишєва	50	96	4
	Мінковського	50	98	2
Метод Медіани	Евклідова відстань	50	63	37
	Квадрат евклідової відстані	50	87	13
	Косінус	49	1	100
	Кореляція Пірсона	50	47	53
	Чебишєва	50	86	14
	Мінковського	50	63	37
Метод Уорда	Евклідова відстань	50	64	36
	Квадрат евклідової відстані	50	64	36
	Косінус	50	52	48
	Кореляція Пірсона	50	39	61
	Чебишєва	50	65	35
	Мінковського	50	64	36

Результати кластерізації були збережені у вихідному файлі з метою подальшого аналізу якості кластеризації даних. Авторами статті проведено апробацію використання статистичного оцінювання якості кластеризації з використанням методу перехресних таблиць. Найбільш точні результати під час розрахунків за методом середнього зв'язку було досягнуто із застосуванням метрики кореляція Пірсона. Як свідчить рисунок 3 а, у перший кластер потрапили всі 50 зразків півників щетинистих, у другий кластер потрапили 48 півників строкатих та шість півників віргініка. У третьому кластері 34

півники віргініка та два півники строкатих. Аналогічний результат отримано з використанням методу середнього зв'язку із застосуванням метрики кореляція Пірсона. Близький до цього результат отримано з використанням метрики Косінус.

Під час застосування методу одинарного зв’язку та метрик Евклідова відстань, квадрат Евклідової відстані, відстані Чебишєва та Мінковського у першому кластері 50 зразків, у другому 98 зразків, у третьому два зразки. У разі застосування метрик Косінус та Пірсона у першой кластер потрапили 49 зразків, у другий один зразок, у третій 100 зразків, що зовсім не відповідає співвідношенню ботанічних таксонів у вихідному наборі даних. Методи повного зв’язку та метод центроїда із використання всіх вище названих метрик також неправильно розподіляють зразки у кластери.

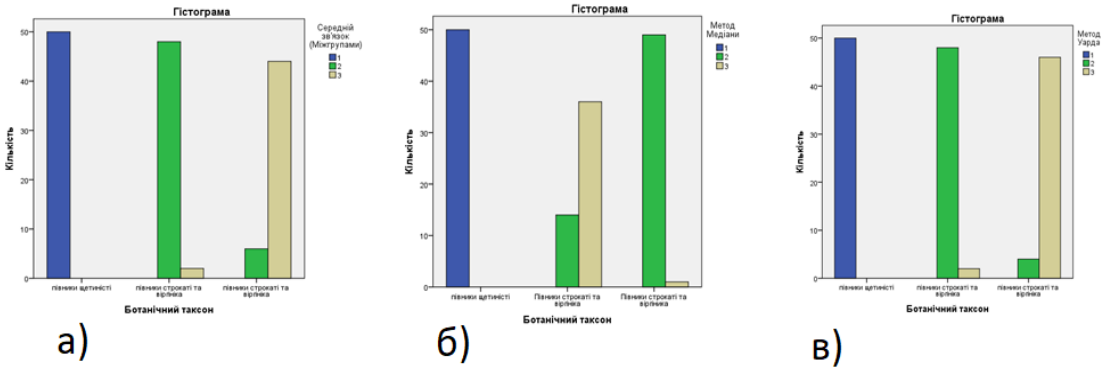

Рис. 3. а) метод середнього зв’язку та метрика кореляція Пірсона, б) метод Медіана та метрика кореляції Пірсона, в) методом Уарда та метрика Косінус

Під час використання методу Медіана у першому кластері 50 зразків півників щетинистих, у другому 14 зразків півників строкатих та 49 півників віргініка, у третьому 36 півників строкатів та один зразок півників віргініка.

Під час використання методу Уорда та метрики Косинус (рис. 3 в) у першому кластері 50 зразків півників щетиністих, у другому 48 зразків півників строкатих та чотири зразка півників віргініка, у третьому два зразки півників строкатих та 46 зразків півників віргініка.

Висновки

Результати тестування свідчать, що не існує жодного універсального алгоритму, який би ідеально розподілив набір Ірисів Фішера на кластери. Найкращий результат було досягнуто під час кластерізації з використанням методів: середнього зв'язку між групами із застосуванням метрики кореляція Пірсона, а також методу середнього зв'язок усередині групи із застосуванням метрики кореляція Пірсона та метрики Косінус, й методу Уорда з використанням метрики Косінус.

Використання статистичного оцінювання якості результатів кластеризації апаратом перехресних таблиць дозволило більш точно виокремити найбільш придатні методи та метрики агломеративного ієрархічного кластерного аналізу. Ураховуючи, що колекція даних про сорти рослин постійно зростає, рекомендовано використовувати класифікацію як інструмент інтелектуального аналізу даних для полегшення розпізнання нових сортів рослин за їх морфологічною ознакою. Проте кластеризацію сортів рослин потрібно проводити ітераційно, послідовно застосовуючи найбільш поширені алгоритми кластеризації та метрики. Результати кластеризації рекомендовано оцінювати з використанням апарату частотної статистики, зокрема перехресних таблиць.

Перспективи подальших досліджень. У базі даних Українського інституту експертизи сортів рослин зберігаються дані щодо 46 303 сортів рослин 661 ботанічного таксону. Перелік морфологічних ознак для кожного з ботанічних таксонів визначається методичними рекомендаціями UPOV та вітчизняними методиками. Дослідженню методів та метрик кластеризації, які придатні для аналізу морфологічних ознак кожного з ботанічних таксонів буде приділятися увага у найближчим часом.

Reference

1. Balan, H. O., Tkachyk, S. O., Orlenko, N. O., & Bushulian, O. V. (2018). Analysis of the phytosanitary state of crops of various soybean varieties in the conditions of the Southern Steppe of Ukraine. *Plant*

Varieties Studying and Protection, 14 (3), 295–301. doi:10.21498/2518-1017.14.3.2018.145300 [In Ukrainian].

2. Melnyk, A. V. (2013). Vykorystannia klasternoho analizu za pidboru sortiv i hibrydiv ripaku yaroho dlia vyroshchuvannia v Livoberezhnomu Lisostepu Ukrainy. *Visnyk Poltavskoi Derzhavnoi Ahrarnoi Akademii*, (4), 6–11. doi:10.31210/visnyk2013.04.01 [In Ukrainian].

3. Prysyazhnyuk, O. I., & Dymytrov, S. G. (2014). Ocinka reakciyi novyx gibrydiv sonyashnyku na umovy vyroshhuvannya. *Naukovi dopovidi Nacional`nogo universytetu bioresursiv i pryrodokorystuvannya Ukrayiny*, 7, 8–14 [In Ukrainian].

4. Stekh, Yu. V., Faisal Sardikh, M. E., Kernytskyi, A. B., & Dombrova, M. S. (2010). Alhorytmichna otsinka optymalnosti rezultativ klasteryzatsii za kryteriiem vidstani. *Visnyk Natsionalnoho universytetu "Lvivska politekhnika"*, 685, 131–134 [In Ukrainian].

5. Tyshchenko, V. N., Panchenko, P. M., & Chernыsheva, O. P. (2013). Ydentyfykatsyia sortov y selektsyonnыkh lynyi pshenytsы ozymoi po sbalansyrovannosty kolychestvennыkh pryznakov s yspolzovanyem klasternoho analyza. *Visnyk Poltavskoi Derzhavnoi Ahrarnoi Akademii*, (3), 28–35. doi:10.31210/visnyk2013.03.04 [In Ukrainian].

6. Anderson, E. (1935). The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, 59, 2–5.

7. Anderson, E. (1936). The Species Problem in Iris. *Annals of the Missouri Botanical Garden*, 23 (3), 457. doi:10.2307/2394164.

8. Bagnall, A., & Janacek, G. (2005). Clustering Time Series with Clipped Data. *Machine Learning*, 58 (2-3), 151–178. doi:10.1007/s10994-005-5825-6.

9. Bisson, G., Nedellec, C., & Canamero, L. (2000). Designing clustering methods for ontology building - The Mo'K workbench. *Proceedings of the ECAI Ontology Learning Workshop*, 13–19.

10. Bryman, A. (2012). *Quantitative Data Analysis with IBM SPSS, 17, 18 & 19: A Guide for Social Scientists*. New York: Routledge. doi:10.4324/9780203180990.

11. Crossman, A. (2014). Analyzing quantitative data: Statistical software programs for use with quantitative data. Retrieved from: http://sociology.about.com/od/ Research- Tools/a/Computer-programs-quantitative-data.htm.

12. Crossman, A. (2014). Analyzing Quantitative Data: Statistical Software Programs for Use with Quantitative Data. Retrieved from: http://sociology.about.com/od/ResearchTools/a/Computer-programs-quantitative-data.htm.

13. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. (2008). Querying and mining of time series data. *Proceedings of the VLDB Endowment*, 1 (2), 1542–1552. doi:10.14778/1454159.1454226.

14. Dutta, D., Roy, A., & Choudhury, K. (2013). Training Aritificial Neural Network Using Particle Swarm Optimization Algorithm. *International Journal on Computer Science And Engineering(IJCSE)*, 3 (3).

15. Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2), 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

16. Fisher, R. A. (1936). UCI Machine Learning Repository: Iris Data Set. Retrieved from: http://archive.ics.uci.edu/ml/datasets/Iris. Consulted 10 AUG 2013.

17. Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Englewood Cliffs.

18. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM Computing Surveys, 31(3), 264–323. doi:10.1145/331499.331504.

19. Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). Data Mining in Agriculture. *Springer Optimization and Its Applications*. doi:10.1007/978-0-387-88615-2.

20. Robbins, S. (2012). How Does SPSS Differ from a Typical Spreadsheet Application. Retrieved from: https://publish.illinois.edu/commonsknowledge/2012/06/07/how-does-spss-differ-from-atypical-spreadsheet-application.

21. Rodrigues, P. P., Gama, J., & Pedroso, J. P. (2008). Hierarchical Clustering of Time-Series Data Streams. *IEEE Transactions on Knowledge and Data Engineering*, 20 (5), 615–627. doi:10.1109/tkde.2007.190727.

22. Stekh, Y., Fajsal, M. E., Sardieh Kernytskyy A., & Nykyforchyn, R. (2008). Dialog graphical system of classification with the help of distance function. *Proceedings of the XVI Ukrainian-Polish Conference on "CAD in Machinery Design. Implementation and Education Problems*". Lviv: CADMD.

23. Sunaga, D. Y., Nievola, J. C., & Ramos, M. P. (2007). Statistical and Biological Validation Methods in Cluster Analysis of Gene Expression. *Sixth International Conference on Machine Learning and Applications.* USA: ICMLA. doi:10.1109/icmla.2007.55.

Стаття надійшла до редакції 03.04.2019 р.

Бібліографічний опис для цитування:

Орленко Н. С., Мажуга К. М., Душар М. Б., Маслечкін В. В. Порівняльний аналіз ієрархічних методів кластерізації придатних для оброблення даних морфологічних ознак сортів рослин. *Вісник ПДАА*. 2019. № 2. С. 261–269.