Климчук В.О. Кластерний аналіз: використання у психологічних дослідженнях // Практична психологія та соціальна робота. – 2006. – №4. – С. 30-36.

КЛАСТЕРНИЙ АНАЛІЗ: ВИКОРИСТАННЯ У ПСИХОЛОГІЧНИХ ДОСЛІДЖЕННЯХ

Анотація.

У статті вивчаються можливості використання кластерного аналізу у психологічних дослідженнях. Робота проводиться на матеріалі реальних психологічних досліджень.

Постановка проблеми.

Ми продовжуємо роботу над методами багатомірного математичного аналізу результатів психологічних досліджень. Темою цієї статті є кластерний аналіз – незамінний метод, коли необхідно знайти внутрішню структуру даних, побудувати емпіричну таксономію, виявити ієрархічну будову даних. Як же працює кластерний аналіз? Зараз ми ці питання детально і розглянемо. Знову ж таки, будемо спиратися на роботу з програмою Statistica 5.5.

Мета кластерного аналізу. Достатньо багато досліджень ставлять за мету організацію отриманих даних у наглядні структури. Так, в біології часто метою є розбиття сукупності тварин на види і підвиди, у психології – класифікація видів поведінки, у педагогіці – таксономія виховних цілей тощо. Допомогти це зробити може *кластерний аналіз*. Фактично, кластерний аналіз є набором різноманітних алгоритмів класифікації. Техніка кластеризації застосовується в дуже різноманітних сферах діяльності. Так, автори підручника Statsoft наводять приклади застосування кластерного аналізу в медицині – кластеризації піддаються симптоми захворювання чи види лікування, і отримуються достатньо цікаві класифікації [9]. Теж саме стосується психіатрії та психотерапії. Відомі також ряд прикладів застосування кластерного аналізу в сфері маркетингу. Загалом, коли необхідно класифікувати великі масиви інформації на групи, які придатні для подальшого аналізу – кластерний аналіз є незамінним інструментом. Цікаві можливості кластерного аналізу у сфері психологічних досліджень групових процесів та явищ.

Кластерний аналіз має одну суттєву особливість – він не є звичайним статистичним методом, оскільки до нього у більшості випадків незастосовні процеси перевірки статистичної значимості. Кластерний аналіз дає найбільш

можливо-значиме рішення. Саме тому досить часто його використовують тоді, коли дослідник має набір даних, але не має жодної апріорної гіпотези про класи цих даних.

Особливості. Перш ніж перейти до безпосередньо алгоритмів кластеризації, виділимо декілька **зауважень**, які слід враховувати, використовуючи кластерний аналіз:

a. Більшість методів кластерного аналізу є доволі таки простими евристичними процедурами, які, як правило, не мають статистичного обґрунтування.
b. Різні методи кластеризації можуть породжувати різні кластерні рішення для одних і тих же даних. Це звичне явище у більшості прикладних дослідженнях, у тому слід по-перше обирати найбільш осмислене рішення, по-друге – завжди вказувати, який саме метод кластеризації було використано.
c. Використовуючи кластерний аналіз, дослідник має на меті виявлення структури даних. В той же час дія кластерного аналізу полягає у привнесені структури у аналізовані дані. Тобто, кластеризація може призвести до появи артефактів (виявлення структури в даних, які її не мають).
d. Осмислене рішення при кластерному аналізі можна обрати лише тоді, коли є базис для його осмислення – теорія. Без теоретичної моделі, без гіпотези стосовно структури даних з'являється небезпека наївного емпіризму, коли результати кластеризації приймаються на істину у кінцевій інстанції.

Етапи проведення кластерного аналізу. Яка ж загальна структура проведення кластерного аналізу? Можна виділити такі етапи роботи:

1. Проведення дослідження.
2. Підготовка даних до кластерного аналізу.
3. Вибір методу кластерного аналізу.
4. Вибір міри відстані між об'єктами та її обчислення.
5. Вибір стратегії кластеризації.
6. Застосування обраної стратегії для утворення кластерів.
7. Перевірка результатів кластерного аналізу на осмисленість і їх інтерпретація.

Підготовка даних до кластерного аналізу. Підготовка даних до кластерного аналізу відбувається так же, як і при факторному аналізі, із збереженням усіх поставлених вимог (див. попередні номери журналу).

Ми провели дослідження позитивного ставлення студентів психологічного факультету спеціальності "Практична психологія. Соціальна педагогіка" до студентів інших факультетів та до майбутніх професійних ролей, і нас цікавить об'єднання студентів у групи на основі схожого ставлення. Для цього ми створили рольовий перелік факультетів та можливих спеціальностей випускників психологічного факультету. Потім ми запропонували студентам оцінити своє ставлення до всіх ролей за 10-бальною шкалою. В результаті було отримано масив даних (табл. 1)[1].

Таблиця 1.

Масив даних, підготовлений до кластерного аналізу

		Об'єкти (студенти-учасники дсолідження)										
		А.О.	О.В.	М.П.	К.П.	Р.А.	К.В.	В.Д.	П.Р.	Д.К.	Е.О.	З.А.
Випадки (результати оцінювання)	Студент психологічного факультету	5	9	9	5	5	10	3	5	1,5	1,5	1,5
	Студент педагогічного факультету	2	0,5	3	1	2	3	4	2	9	9	9
	Студент математичного факультету	3	1	2	4	3	4	6	3,5	10	10	10
	Студент природничого факультету	9	2	1,5	3	4	2	7	7	9,5	9,5	9,5
	Студент історичного факультету	4	3	2,5	3,5	6	5	9	9	7	7	7
	Студент філологічного факультету	10	4	1	2	1	6	1,5	1	8	8	8
	Психолог	4,5	9,5	10	10	10	9	2,5	5,5	2	2	2
	Соціальний педагог	5,5	10	8	9,5	9,5	9,5	3,5	4	1	1	1
	Вчитель	8	5	5	7	5	7	5	3	6	6	6
	Викладач	6	6	7	9	8	5,5	8	8	5	5	5
	Керівник	3,5	7	6	8	7	4,5	10	10	3	3	3
	Консультант	7	8	4	6	6	8	2	6	4	4	4

[1] *Висловлюю подяку кандидату психологічних наук В.В. Горбуновій за надані результати дослідження.*

При проведенні кластерного аналізу слід враховувати, що традиційно кластеризація проводиться по об’єктах (стовпчиках масиву даних).

Методи кластерного аналізу. Дані для аналізу підготовлені. Тепер слід визначитися з методами. Виділяють три основних методи кластерного аналізу: деревоподібна кластеризація, метод К-середніх та двовходове об’єднання.

С. 30

Метод деревоподібної кластеризації (ієрархічна кластеризація, tree clustering) дозволяє побудувати ієрархічне кластерне дерево, що має такий вигляд (рис. 1):

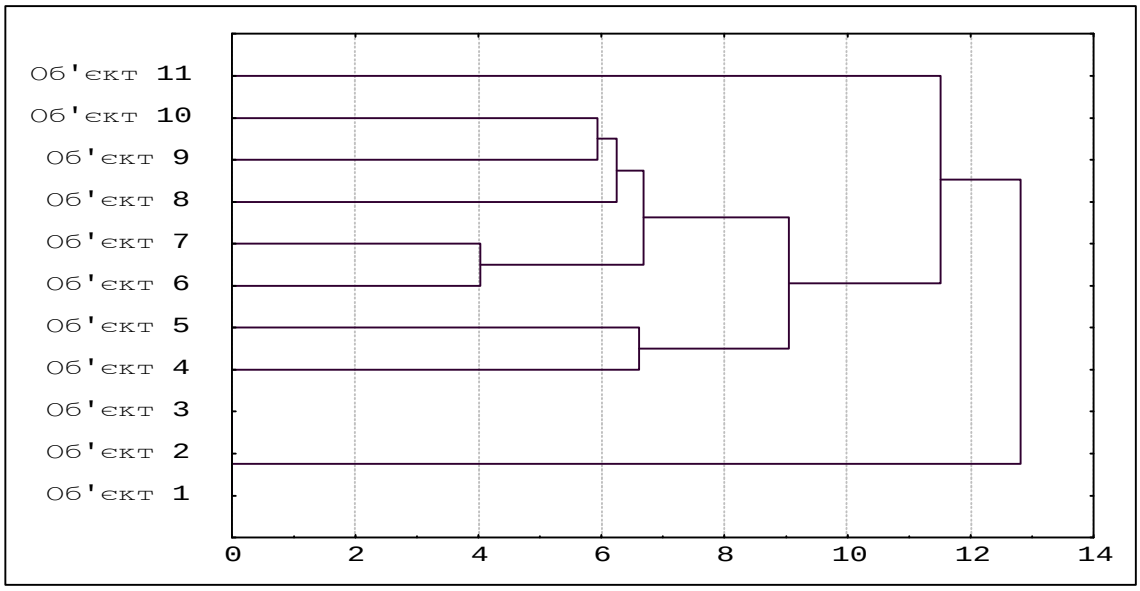

Рис. 1. Ієрархічне кластерне дерево

Метод К-середніх (K-means clustering) використовується тоді, коли у дослідника вже є певні апріорні гіпотези стосовно кількості кластерів. В межах цього методу дослідник має наперед задати кількість кластерів, і алгоритм кластеризації дозволить знайти ці кластери так, щоб вони максимально різнилися один від одного. Перевагою цього методу є можливість перевірки статистичної значимості відмінностей між виділеними кластерами. На рис. 2 наведено графічне зображення результатів кластерного аналізу методом К-середніх.

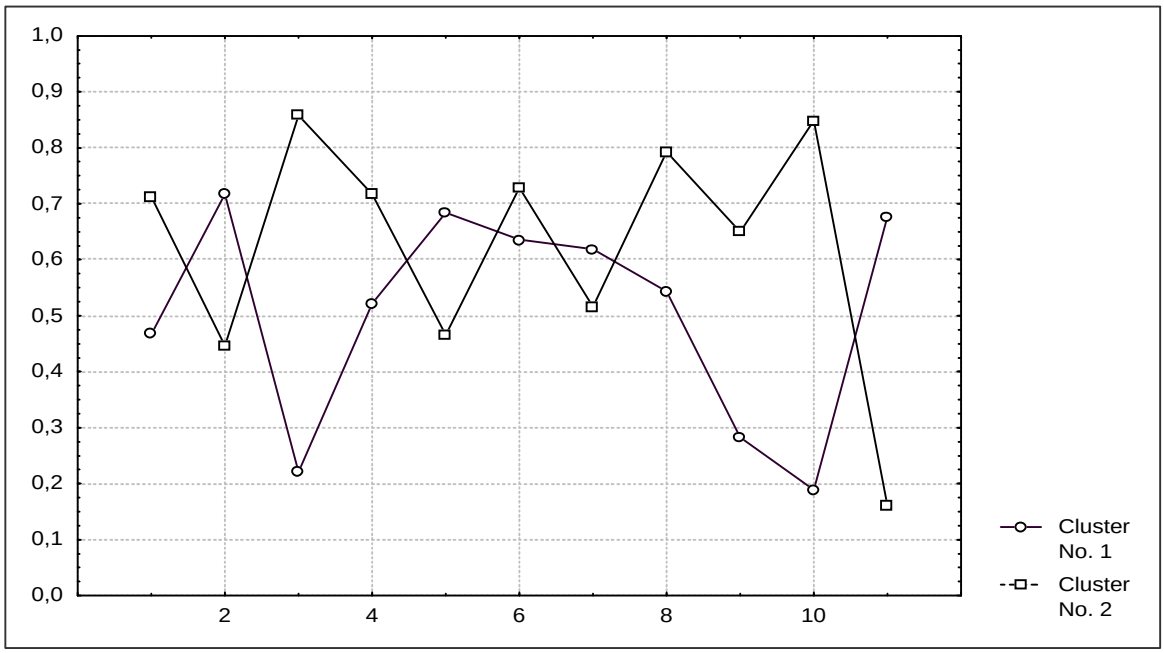

Рис. 2. Кластерний аналіз методом К-середніх.

Метод двовходового об'єднання (two-way joining) використовують у випадках, коли хочуть провести одночасну кластеризацію об'єктів (стовпчиків) та спостережень (рядків). На рис. 3 наведено графічне зображення результатів використання цього методу.

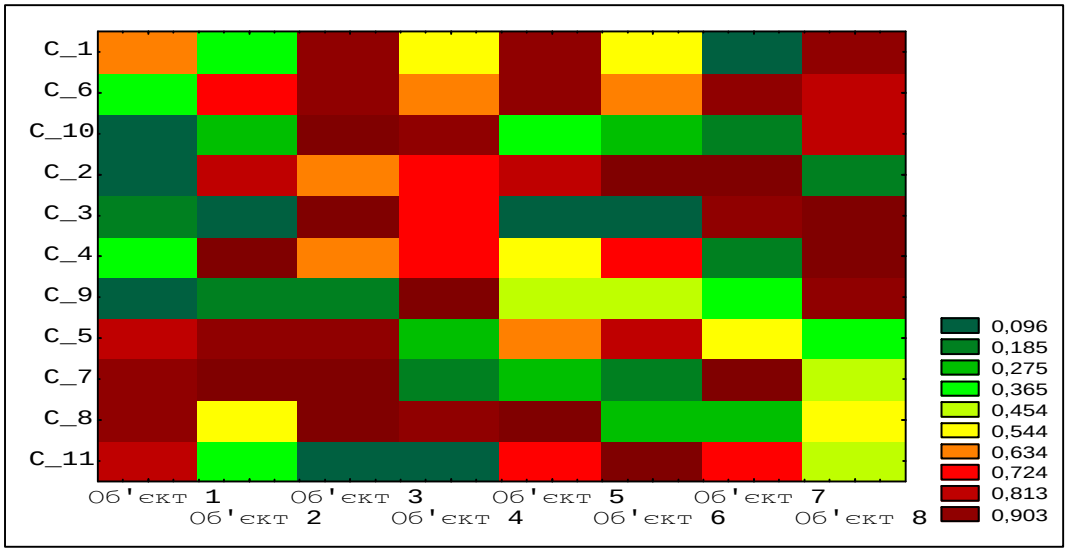

Рис. 3. Результати кластерного аналізу методом двовходового об'єднання

Деревоподібна кластеризація: пошук відстаней між об'єктами. Ми з'ясували, що для проведення кластерного аналізу перш за все необхідно визначитися із методом. Оберемо метод деревоподібної кластеризації як найбільш показовий і зрозумілий (більшість початківців вперше знайомляться саме з ним).

Після визначення з методом необхідно обрати міру відстані між об'єктами та обчислити ці відстані. Розглянемо, які бувають ***міри відстаней*** між об'єктами.

Евклідова відстань (Euclidian distances). Це найуживаніша міра відстані між об'єктами, яка являє собою геометричну відстань між об'єктами у багатомірному просторі. Формула для обчислення Евклідової відстані має такий вигляд (1):

$$L = \sqrt{\sum_i (x_i - y_i)^2} \qquad (1)$$

Евклідова відстань обчислюється по вихідним не стандартизованим даним, а тому всі змінні повинні бути виміряні в одному масштабі (якщо це сантиметри, то всі змінні повинні бути виміряні в сантиметрах тощо).

Квадрат Евклідової відстані (Squared Euclidian distances). Ця міра використовується тоді, коли хочуть на порядок збільшити значення відстаней між дуже віддаленими між собою об'єктами. Формула має такий вигляд (2):

$$L = \sum_i (x_i - y_i)^2 \qquad (2)$$

Манхеттенівська відстань (відстань міських кварталів – City-block (Manhattan) distances). Ця міра у більшості випадків призводить до таких же результатів, як і Евклідова відстань, але зменшується вплив окремих великих різниць (викидів) через те, що відстань обчислюється по простим різницям координат (3):

$$L = \sum_i |x_i - y_i| \qquad (3)$$

Відстань Чебишева (Chebychev distance metric). Використовується тоді, коли хочуть позначити два об'єкти як різні, якщо вони відрізняються якимсь одним виміром (однією координатою). Формула має такий вигляд (4):

$$L = Максимум|x_i - y_i| \qquad (4)$$

Відсоток невідповідності (Percent disagreement). Міра використовується у випадках, коли дані належать номінативній шкалі, і обчислюється за такою формулою (5):

$$L = \frac{Кількість\ x_i \neq y_i}{i} \qquad (5)$$

1-коефіцієнт кореляції Пірсона (1-Pearson r). Використовується у випадку, абсолютні значення та різниці між об'єктами несуттєві, а більш важливим є

наявність зв'язку між ними – $L = 1 - r$. Ця міра чутлива лише до схожості профілів об'єктів. Призводить до результатів, близьких до факторного аналізу – кластери можуть наближатися до факторів.

Оберемо Евклідову відстань, і проведемо обчислення. Для прикладу (табл. 2) обчислимо відстань між першим та другим об'єктами (досліджувані А.О. та О.В. з вихідного масиву даних табл. 1).

Таблиця 2.

Обчислення відстані між об'єктами А.О. і О.В.

	Об'єкти		$(x_i - y_i)$	$(x_i - y_i)^2$	$L = \sqrt{\sum_i (x_i - y_i)^2}$
	А.О.	О.В.			
Студент психологічного факультету	5	9	-4	16	
Студент педагогічного факультету	2	0,5	1,5	2,25	
Студент фізико-математичного факультету	3	1	2	4	
Студент природничого факультету	9	2	7	49	
Студент історичного факультету	4	3	1	1	
Студент філологічного факультету	10	4	6	36	$L = \sqrt{175,75} = 13,26$
Психолог	4,5	9,5	-5	25	
Соціальний педагог	5,5	10	-4,5	20,25	
Вчитель	8	5	3	9	
Викладач	6	6	0	0	
Керівник	3,5	7	-3,5	12,25	
Консультант	7	8	-1	1	
$\sum_i (x_i - y_i)^2$				175,75	

Провівши послідовні обчислення відстаней між усіма об'єктами будують **таблицю відстаней (distances matrix)** (табл. 3). Процес обчислень відстаней вручну є достатньо громіздким і тривалим, а тому доцільніше для цього використати комп'ютерні програми.

Таблиця 3.

Таблиця відстаней між усіма об'єктами

	А.О.	О.В.	**М.П.**	К.П.	Р.А.	К.В.	В.Д.	П.Р.	**Д.К.**	**Е.О.**	**З.А.**
А.О.	0,00	13,26	14,81	13,41	13,51	11,51	14,37	13,63	12,82	12,82	12,82
О.В.	13,26	0,00	**6,24**	7,00	7,26	5,94	16,76	12,94	**21,66**	**21,66**	**21,66**
М.П.	14,81	6,24	0,00	6,82	6,69	8,02	14,94	12,37	**20,41**	**20,41**	**20,41**
К.П.	13,41	7,00	6,82	0,00	4,03	8,82	13,44	10,95	19,62	19,62	19,62
Р.А.	13,51	7,26	6,69	4,03	0,00	8,86	12,35	9,04	18,89	18,89	18,89
К.В.	11,51	5,94	8,02	8,82	8,86	0,00	16,43	13,83	18,69	18,69	18,69
В.Д.	14,37	16,76	14,94	13,44	12,35	16,43	0,00	6,61	12,85	12,85	12,85
П.Р.	13,63	12,94	12,37	10,95	9,04	13,83	6,61	0,00	15,97	15,97	15,97
Д.К.	12,82	21,66	20,41	19,62	18,89	18,69	12,85	15,97	0,00	0,00	0,00
Е.О.	12,82	21,66	20,41	19,62	18,89	18,69	12,85	15,97	0,00	0,00	0,00
З.А.	12,82	21,66	20,41	19,62	18,89	18,69	12,85	15,97	0,00	0,00	0,00

Чим менше значення у комірці таблиці, тим ближче знаходяться між собою відповідні об'єкти. Так, з табл. 3 видно, що найбільша відстань – між студентами О.В. та Д.К., Е.О., З.А. (L=21,66), а також між М.П. та Д.К., Е.О., З.А. (L=20,41). Очевидно, що вже на цьому етапі аналіз можна зробити висновок про наявність принаймні двох груп студентів, які різняться за свої ставленням до навчання та майбутньої професії – група О.В., М.П., та група Д.К., Е.О., З.А. Подивимося на відстані всередині кожної з груп. Дійсно, виявляється, що дуже близькі між собою студенти О.В. та М.П. (L=6,24), а також ідентичні між собою (знаходяться на нульовій відстані) студенти Д.К., Е.О. та З.А. (L=0).

Деревоподібна кластеризація: стратегії кластеризації. Можна продовжити цей аналіз, шукаючи близькі та віддалені групи студентів, орієнтуючись виключно на таблицю відстаней, а можна продовжити деревоподібний кластерний аналіз, використавши одну із ***стратегій кластеризації***. Стратегії кластеризації являють собою правила об'єднання об'єктів (змінних) у кластери. Вони переглядають таблицю схожостей об'єктів, і на кожному кроці послідовно об'єднують пару найбільш схожих об'єктів (змінних чи кластерів). Завершується процес утворенням одного кінцевого великого кластера, який включає в себе всі об'єкти. Основна різниця між стратегіями – це спосіб вимірювання відстаней. Однак, тут уже мова йдеться не про безпосередні відстані між об'єктами – на першому кроці

кластеризації кожен об’єкт являє собою окремий кластер, і відстані між ними визначаються обраною мірою (Евклідова відстань, відстань Чебишева тощо). Мова вже йде про той випадок, коли декілька об’єктів зв’язуються разом – як тепер визначити відстані між утвореними кластерами? Розглянемо для цього найпоширеніші ***стратегії кластеризації***.

Стратегія найближчого сусіда (Nearest neighbor) або стратегія одиночного зв’язок (Single lincage). Тут відстань між двома кластерами визначається як відстань між двома найближчими об’єктами (найближчими сусідами). Стратегія ніби нанизує об’єкти один на один, і в результаті кластери представляються у вигляді довгих “ланцюжків”. Стратегія пов’язує два кластери разом, коли будь-які два об’єкти в цих кластерах ближче один до одного, ніж усі інші.

Стратегія найвіддаленішого сусіда (Furthest neighbor) або стратегія повного зв’язку (Complete lincage). При використанні цієї стратегії відстань між кластерами визначається найбільшою відстанню між двома об’єктами з різних кластерів (між найвіддаленішими сусідами). Стратегія добре працює, коли об’єкти реально належать різним класам. Якщо є природним типом кластерів в отриманих даних є ланцюжки, то ця стратегія є непридатною. Стратегія утворює в основному “кущі” об’єктів.

Стратегія незваженого попарного середнього (Unweighted pair-group average). Відстань між двома кластерами визначається як середня відстань між всіма парами об’єктів у них. Метод ефективний випадку реального об’єднання об’єктів як у “кущі”, так і в “ланцюжки”.

Стратегія зваженого попарного середнього (Weighted pair-group average). Стратегія відрізняється від попередньої тим, що при обчисленнях розмір відповідного кластера (кількість об’єктів, які він містить) використовується в якості вагового коефіцієнта. Тому цю стратегію використовують тоді, коли передбачають появу кластерів нерівного розміру.

Стратегія Варда (Ward’s method). Ця стратегія суттєво відрізняється від попередніх, оскільки використовує методи дисперсійного аналізу для оцінки відстаней між кластерами. Ця стратегія мінімізує суму квадратів (SS) для двох гіпотетичних кластерів, які можуть бути сформовані на кожному кроці процесу

кластеризації. Метод вважається ефективним, але намагається створювати кластери малого розміру.

Які ж загальні властивості описаних стратегій? Якщо уявити таблицю схожостей як багатомірний простір, а об'єкти – як точки цього простору, то можна описати, що роблять з точками цього простору розглянуті нами стратегії.

Перший тип стратегій – *стискаючі*. Стратегії цього типу ніби стискають простір об'єктів, зменшуючи відстані між усіма групами даних. Коли черговий об'єкт піддається такій стратегії, він, швидше за все, буде віднесений до вже існуючого кластера, ніж стане джерелом утворення нового. Сюди можна віднести стратегію найближчого сусіда.

Другий тип стратегій – *розширюючі*. Ці стратегії ніби розширюють простір об'єктів, збільшуючи відстані між ними. Точки ніби розступаються, утворюючи дрібніші, але чіткіші групи (кластери). Ці стратегії створюють кластери гіперсферичної форми, приблизно однакові за об'ємом. До цього типу належать стратегія найвіддаленішого сусіда та стратегія Варда.

Третій тип стратегій – *зберігаючі*. Стратегії, які належать до цього типу, залишають вихідний простір об'єктів без змін. Це стратегії зваженого та незваженого попарного середнього.

Давайте до даних таблиці відстаней (табл. 3) застосуємо **стратегію найближчого сусіда**. Найменша відстань у таблиці – між об'єктами Д.К., Е.О., З.А. (L=0). Очевидно, вони утворюють **перший** кластер. Друга по величині відстань – між об'єктами К.П. і Р.А. (L=4,03). Їх варто об'єднати у **другий** кластер. **Третій** кластер утворять об'єкти О.В. і К.В. (L=5,94). До третього кластера також слід приєднати об'єкт М.П., оскільки його відстань від О.В. рівна L=6,25. Таким чином, **четвертий кластер** буде мати свою структуру – первинний кластер з об'єктів О.В. і К.В., і вторинний кластер, який включає в себе ще об'єкт М.П. **П'ятий кластер** буде утворений об'єктами В.Д. і П.Р. (L=6,61).

На цьому етапі кластеризації майже кожен об'єкт увійшов до якогось кластера, і тепер слід зв'язати між собою вже утворені кластери. Наступна за величиною є відстань L=6,69 – це відстань між об'єктами М.П. (четвертий кластер) і Р.А. (другий кластер). Очевидно, що другий та третій кластери слід об'єднати у кластер вищого

порядку – **шостий кластер**. На наступному кроці варто вже шукати відстані між п’ятим і шостим кластерами. Нею буде L=9,04 між об’єктами П.Р. і Р.А. і утвориться **сьомий кластер**. Далі аналізуючи відстані можна побачити, що до сьомого кластеру приєднується об’єкт А.О. з мінімальною відстанню L=11,51 від об’єкта К.В. з сьомого кластера. Так утворюється **восьмий кластер**. І нарешті **дев’ятий кластер** утворений об’єднанням першого та восьмого з відстанню L=12,81 (об’єкти Д.К., Е.О., З.А та А.О.)

Можна узагальнено представити описаний процес кластеризації у ***таблиці об’єднань (amalgamation schedule)*** – табл. 4.

Таблиця 4.

Таблиця об’єднань

Кро	Відстані між найближчими сусідами	ОБ’ЄКТИ										
		1	2	3	4	5	6	7	8	9	10	11
1	0,00	Д.К.	Е.О.									
2	0,00	Д.К.	Е.О.	З.А.								
3	4,03	К.П.	Р.А.									
4	5,94	О.В.	К.В.									
5	6,25	О.В.	К.В.	М.П.								
6	6,61	В.Д.	П.Р.									
7	6,69	О.В.	К.В.	М.П.	К.П.	Р.А.						
8	9,04	О.В.	К.В.	М.П.	К.П.	Р.А.	В.Д.	П.Р.				
9	11,51	А.О.	О.В.	К.В.	М.П.	К.П.	Р.А.	В.Д.	П.Р.			
10	12,82	А.О.	О.В.	К.В.	М.П.	К.П.	Р.А.	В.Д.	П.Р.	Д.К.	Е.О.	З.А.

В наведеній таблиці представлено кожен описаний вище крок, біля кожного кроку стоїть відстань між найближчими сусідами, а справа – виділені на кожному з кроків кластери. На цьому процедура кластеризації завершена, і можна побудувати графічне зображення отриманих кластерів – ***кластерне дерево (hierarchical tree plot)*** – рис. 4.

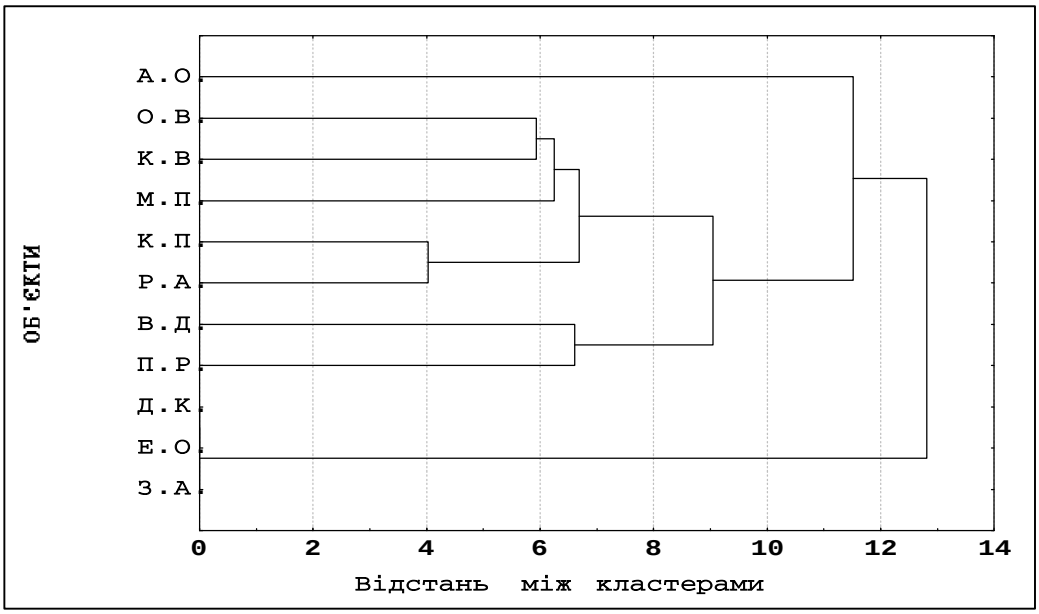

Рис. 4. Кластерне дерево

Дійсно, тепер наочно і чітко видно утворену кластерну структуру – кластери студентів, що мають однакове ставлення до своїх рольових позицій і до студентів інших факультетів. Перша група студентів – Д.К., В.О., З.А., друга група – К.П. та Р.А., третя група – О.В. та К.В., до якої приєднується М.П., четверта група – В.Д. та П.Р. Далі від усіх знаходиться студент А.О., який не входить у жоден з первинних кластерів, а значить, найбільше відрізняється від усіх інших.

Інтерпретація. Останній етап кластерного аналізу – інтерпретація. Інтерпретуючи кластерні дерева, слід намагатися в межах кожного із виділених кластерів знайти певний смисловий інваріант, який би пояснив причину об'єднання об'єктів у цей кластер. Для цього додатково можна використати: 1) описові статистики об'єктів, 2) описові статистики спостережень (значень об'єктів), 3) вихідний масив даних. В результаті такого аналізу необхідно знайти певну психологічну характеристику, яка одночасно зможе слугувати назвою кластера.

Так, до першого кластера потрапили студенти Д.К., Е.О., З.А. Звернувшись до таблиці вихідних даних (табл. 1), можна побачити що це студенти, які найпозитивніше ставляться до студентів інших факультетів, і одночасно – не дуже позитивно до майбутніх рольових позицій. Студенти другого кластеру (К.П. та Р.А.) навпаки, позитивно ставляться до майбутніх професійних рольових позицій, і не дуже позитивно – до студентських колективів. Студенти третього кластеру (О.В., К.В., М.П.) позитивно ставляться до професійних позицій психолога та до студентів

своєї спеціальності. Студенти четвертого кластеру (В.Д. та П.Р.) найпозитивніше ставляться до професійних позицій педагога та до студентів такої ж орієнтації. Ставлення студента А.О. доволі таки важко визначити – очевидно, його професійна ідентифікація ще не відбулася.

Узагальнюючи результати проведеного аналізу, можна припустити наявність такої основи для класифікації студентів, як професійна ідентифікація. Студенти, у яких **несформована професійна ідентифікація** поділяються на два класи – ті, які більш позитивно ставляться до студентського життя (Д.К., Е.О., З.А.), і ті, які позитивно ставляться до майбутньої діяльності, не диференціюючи її на види (К.П., Р.А.). Студенти, у яких **професійна ідентифікація сформована**, також діляться на дві класи – орієнтовані на психологічну діяльність (О.В., К.В., М.П.) та орієнтовані на педагогічну діяльність (В.Д., П.Р.).

Таблиця 5.

Інтерпретація результатів кластерного аналізу

Несформована професійна ідентифікація	позитивно ставляться до студентського життя	Д.К., Е.О., З.А.
	позитивно ставляться до майбутньої діяльності, не диференціюючи її на види	К.П., Р.А.
Сформована професійна ідентифікація	орієнтовані на психолого-педагогічну діяльність	О.В., К.В., М.П.
	орієнтовані на керівну діяльність	В.Д., П.Р.

Очевидно, що не всі аспекти емпірично виявленої класифікації повністю накладаються на структуру кластерного дерева, але, власне, у психології такої абсолютно точності навряд чи коли вдасться досягти. Зроблені узагальнення можуть стати джерелом подальших більш ґрунтовних досліджень і висновків з них. Кластерний же аналіз на цьому етапі дослідницької роботи свою задач виконав – виявив структуру студентської групи на основі рівня професійної ідентифікації. Зразки інтерпретацій результатів кластерного аналізу можна знайти в класичних роботах В.Ф. Петренка [6, 7], В.В. Століна [5, 8].

Представлення результатів кластерного аналізу. Більшість результатів досліджень науковці представляють у формі наукових звітів, наукових робіт

(дисертаційних, конкурсних), наукових статей та тез доповідей. Якщо йдеться мова про представлення результатів кластерного аналізу, то слід визначитися із наборами показників, які варто включати в текст та у додатки до роботи. Розглянемо це питання з точки зору способу представлення.

Науковий звіт. Готується по завершенню виконання певного етапу наукової роботи, і подається замовнику цієї роботи. Науковий звіт є технічною документацією, і його оформлення регулюється вимогами ДСТУ 3008-95. Відповідно до вимог, у науковому звіті повинні бути представлені ВСІ числові результати дослідження. Основна суть цих вимог – щоб будь-хто міг, взявши ваші дані, провести повторні обчислення і отримати ті ж самі результати. Якщо мова йдеться про кластерний аналіз, то необхідно навести *таблиці вихідних даних, таблиці відстаней, таблиці об'єднань і кластерні графіки*. При цьому *таблиці вихідних даних* та *таблиці об'єднань* можна навести у додатках, а *таблиці відстаней* та *кластерні графіки* – у тексті звіту. З метою економії місця в основному тексті (деякі замовники обмежують обсяг звіту) в додатки можна винести і *таблиці відстаней*.

Наукова робота. Готується по завершенню цілісного наукового дослідження. Наукові роботи можуть бути дисертаційними та конкурсними. Оформлення *дисертаційної роботи* регулюється вимогами ДСТУ 3008-95 та ВАК України. Відповідно до вимог, у дисертаційній роботі теж мають бути представлені ВСІ числові результати дослідження. Вимоги до представлення результатів кластерного аналізу – ті ж самі, що і у попередньому пункті. Вимоги до *конкурсної роботи* диктуються організацією, яка оголошує конкурс наукових робіт, однак, у більшості випадків, вони залишаються стандартними.

Наукова стаття. Вимоги до оформлень наукових статей формулюються редакторати тих чи інших видань. Однак, якщо говорити про результати кластерного аналізу, то для їх представлення у статті достатньо *кластерного графіка*, а для більшої доказовості – *таблиці відстаней*.

Тези доповіді. В тезах доповіді не прийнято представляти ні графічної, ні табличної інформації. Для представлення результатів кластерного аналізу зазвичай описують гіпотезу дослідження, метод та отримані результати.

У всіх випадках обов'язково слід вказати: метод кластеризації; міру відстані між об'єктами; стратегію кластеризації. Наприклад: "*Кластерний аналіз було проведено методом деревоподібної кластеризації, мірою відстані слугувала відстань Чебишева, а кластеризація проводилася з допомогою стратегії Варда*".

Кластерний аналіз за методом В.Ю. Крилова та Т.В. Острякової. В.Ю. Крилов та Т.В. Острякова запропонували декілька нових цікавих стратегій кластеризації, які досить просто можна реалізовувати вручну [3]. Цікаво також те, що ґрунтуються ці стратегії на виявлених Л.С. Виготським етапах розвитку понять [1]: асоціативний комплекс; ланцюговий комплекс; асоціативно-ланцюговий комплекс; комплекс-колекція тощо. Виявляється, на основі уявлень Л.С. Виготського можна легко будувати кластери відповідної структури. Розглянемо основі стратегії кластеризації.

Простий асоціативний кластер. Для побудови асоціативного кластера обирається один елемент, який буде ядром. Далі підбираються усі елементи, які мають найближчу відстань до ядра – рис. 5. Таким чином, ядер може бути стільки, скільки є елементів у матриці. Для вибору того чи іншого елементу ядром, слід враховувати зміст задачі, які вирішує дослідник, особливості матриці відстаней.

Нехай нас у матриці відстаней (табл. 2) цікавить студент ***О.В.*** *(можливо, він найбільш активний або успішний у навчанні).*

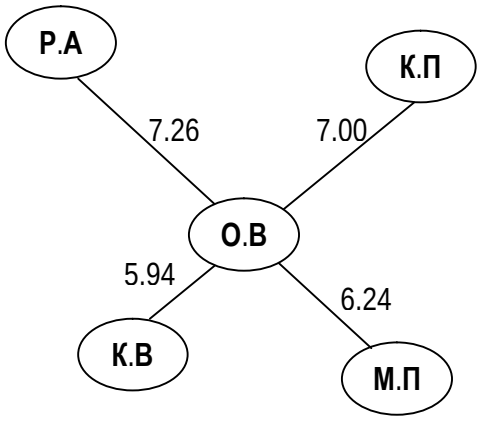

Рис. 5. Простий асоціативний кластер. Ядро – О.В.

Складний асоціативний кластер. Від простого відрізняється тим, що в якості ядра виступають декілька елементів. Для побудови такого кластеру обирають складне ядро (як правило, декілька близьких між собою елементів), потім для кожного елемента підбирають його найближчих сусідів. Після цього обирається певна мінімальна відстань і для залишаються у кластері лише ті елементи, які мають цю мінімальну відстань від ядра (рис. 6).

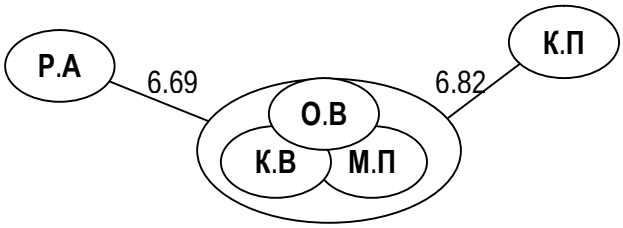

Рис. 6. Складний асоціативний кластер. Ядро – О.В., К.В., М.П.

Ланцюговий кластер. Для його побудови обирається з матриці один початковий елемент (№1), далі шукається найближчий до нього елемент (№2) і т.д. Побудова ланцюгового кластеру завершується тоді, коли вичерпані усі близькі елементи – рис. 7.

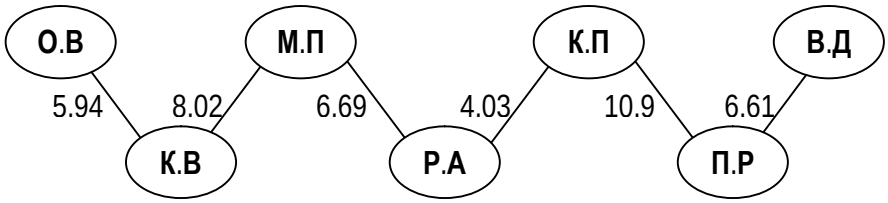

Рис. 7. Ланцюговий кластер

Побудова асоціативного та ланцюгового кластерів переслідує різні цілі. При побудові асоціативного кластера метою є пошук елементів, найближчих до ядра, а при побудова ланцюгового кластера – пошук послідовного зв'язку початкового елементу з усіма іншими елементами матриці.

Прикладом використання ланцюгового кластера може бути дослідження циркуляції чуток на приватному підприємстві. Для цього слід створити матрицю, у якій мірою близькості між працівниками організації є кількість щоденних *контактів між ними. Тоді, побудувавши ланцюгові кластери для кожного працівника, можна буде виявити найімовірніші шляхи розповсюдження чуток.*

Однак, було б цікаво розробити таку стратегію, яка б поєднувала у собі переваги як асоціативного, так і ланцюгового кластерів. Таким буде асоціативно-ланцюговий кластер.

Асоціативно-ланцюговий кластер. Для його побудови спочатку використовується стратегія побудови асоціативного кластера, а потім до кожного елемента цього кластера використовується стратегія побудови ланцюгового кластера – рис. 8.

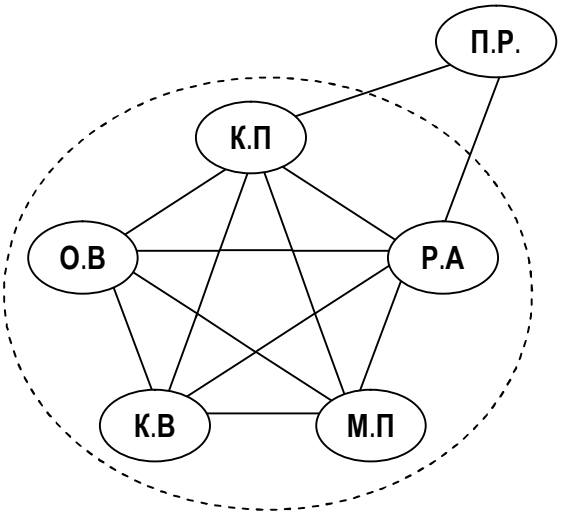

Рис. 8. Асоціативно-ланцюговий кластер

Висновки.

Перед психологами часто постають задачі класифікації, узагальнення, пошуку структури отриманих у дослідженні даних. Типова ситуація – аспірант провів дослідження з вибіркою 5 000 осіб за 15 параметрами, зібрав усі необхідні дані і ... робота зупинилася, тому що він просто втопився у морі цифр. Саме в цей момент на допомогу може прийти кластерний аналіз. З його допомогою можна виявити класи даних, їх структуру і прийняти рішення про перспективи подальшої роботи. Також цей метод може стати безцінним засобом для побудови різноманітних емпіричних класифікацій, таксономій, ієрархічних структур.

Єдине застереження – не зловживати. Кластерний аналіз – не панацея, і при необдуманому використанні може показати те, чого насправді не існує або просто викривити реальність.

Література.

1. Выготский Л.С. Собрание сочинений: В 6-ти т. Т.2. Проблемы общей психологии / Под ред. В.В. Давыдова. – М.: Педагогика, 1982. – С. 136-148.
2. Дронов С.В. Многомерный статистический анализ. – Барнаул: Изд-во

Алтайского гос. ун-та, 2003. – 213 с.

3. Крылов В.Ю., Острякова Т.В. Новые методы кластерного анализа на основе психологической теории развития понятий Л.С. Выготского // Психологический журнал. – 1995. – Т16. – №1. – С. 130-137.
4. Наследов А. Д. SPSS: Компьютерный анализ данных в психологии и социальных науках. – СПб.: Питер, 2004. – 416 с.
5. Общая психодиагностика. Основы психодиагностики, немедицинской психотерапии и психологического консультирования /Под ред. А.А. Бодалева, В.В. Столина. – М.: Изд-во Моск.ун-та, 1987. – 304 с.
6. Петренко В.Ф. Основы психосемантики: Учеб. пособие. – М.: Изд-во Моск. ун-та, 1997. – 400 с.
7. Петренко В.Ф. Психосемантика сознания. – М.: Изд-во Моск. ун-та, 1988. – 208 с.
8. Столин В.В. Самосознание личности. – М.: Изд-во Моск. ун-та, 1983. – 285 с.
9. http://www.statsoft.ru/home/textbook/default.htm
10. http://www.codenet.ru/progr/alg/ai/htm/gl3_10.php