## *Лекція 7. Застосування кластерного аналізу у моделюванні соціально-економічних систем*

**Мета:** ознайомитися з поняттям кластерний аналіз, навчитися застосовувати методи кластерного аналізу при моделюванні соціально-економічних систем.

**План**

7.1. Загальна характеристика кластерного аналізу (КА) як наукового методу пізнання.
7.2. Історія виникнення та застосування КА.
7.3. Формальна постановка та основні етапи задачі кластеризації.
7.4. Особливості проведення кластерного аналізу.
7.5. Аналіз результатів: причини неоднозначності та інтерпретація.

**Перелік ключових термінів і понять:** *кластерний аналіз, кластер, класифікація, метрика, метод k-середніх, евклідова відстань.*

### 7.1. Загальна характеристика кластерного аналізу (КА) як наукового методу пізнання

Класифікація – це основний процес в інтелектуальній діяльності людини. Зустрічаючись з новим явищем, ми намагаємося знайти йому аналог у відомій нам області. Розглядаючи групу яких-небудь об’єктів, ми мимоволі розділяємо їх на підгрупи близьких один одному елементів. Класифікація присутня при упорядкуванні відомих нам фактів, явищ, предметів.

Отже, можна зробити висновок, що класифікація – це фундаментальне поняття науки і практики.

***Кластерний аналіз*** – це сукупність методів, які дозволяють класифікувати багатомірні спостереження, кожне з яких описується набором вихідних змінних.

***Метою кластерного аналізу*** є утворення груп схожих між собою об’єктів, що прийнято називати ***кластерами***.

Слово ***кластер*** англійського походження (cluster), переводиться як згусток, пучок, група. Споріднені поняття, використовувані в науковій літературі, – клас, таксон, згущення.

***Кластерний аналіз*** – це загальна назва множини обчислювальних процедур, які використовують *при створенні класифікації*. У результаті роботи з процедурами утворюються класи чи групи подібних об’єктів.

Більш точно, ***кластерний аналіз*** – це багатомірна статистична процедура, що виконує збір даних, які містять інформацію про вибірку об’єктів, і потім упорядковує об’єкти у порівняно однорідні групи.

Кластерний аналіз виконує такі ***основні завдання***:

- розробка типології або класифікації.
- дослідження корисних концептуальних схем групування об’єктів.

- породження гіпотез на основі дослідження даних.
- перевірка гіпотез або дослідження для визначення, чи дійсно групи, виділені тим чи іншим способом, присутні в наявних даних.

*Зауваження !!!:* Необхідно *не плутати* процедури ***кластерізації*** та ***класифікації***.

***Класифікація*** – віднесення елемента (об'єкта) до певного класу *із заздалегідь відомими параметрами*, отриманими на етапі навчання. Кількість класів при класифікації – строго обмежена.

***Кластеризація*** – це розбиття множини даних на кластери – підмножини однорідних одиниць сукупності, *параметри яких заздалегідь невідомі*.

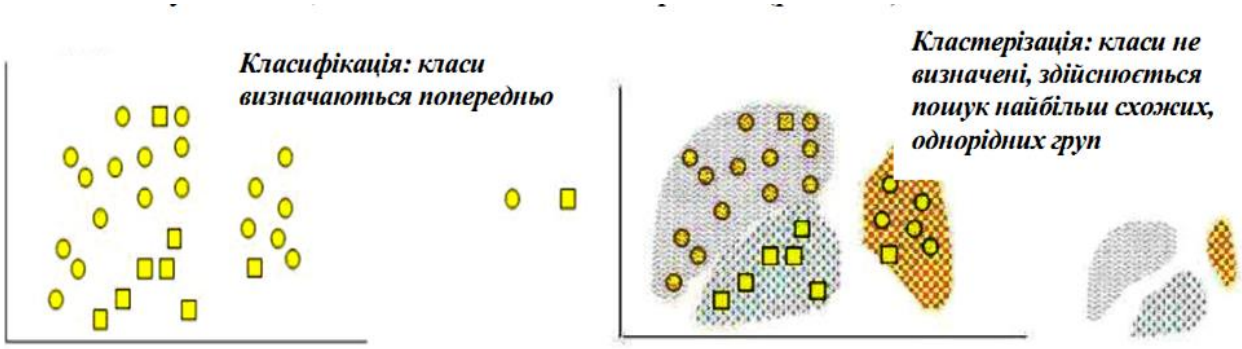


Рис. 7.1. Відмінності классифікації та кластерізації

Кластерний аналіз має низку **переваг перед іншими методами класифікації** даних:

1) він дозволяє виконувати розбиття об'єктів *як за однією ознакою, так і за цілим набором ознак*. Причому вплив кожного з параметрів може бути доволі просто підсилений або послаблений шляхом внесення в математичні формули відповідних коефіцієнтів;

2) кластерний аналіз *не накладає обмежень на вид об'єктів групування* і дозволяє розглядати множину вихідних даних практично довільної природи;

3) особливістю кластеризації є те, що більшість алгоритмів здатні *самостійно визначити кількість кластерів*, на які потрібно розбити дані, а також виділити характеристики цих кластерів *без участі людини*, тільки за допомогою алгоритму, що використовується.

На рис. 7.2 наведено приклад кластерізації об'єктів. Наведені об'єкти доволі прості і мають обмежену кількість характеристик: координати, форма, колір.

Залежно від того, які характеристики використовуються для групування, кластеризація може дати абсолютно різні результати. Реальні об'єкти мають значно більший набір властивостей і, отже, більше варіантів компонування.

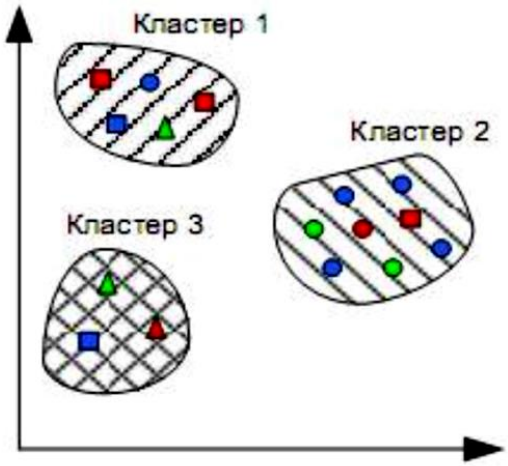


Рис. 7.2. Приклад кластерізації об’єктів

Для того щоб дати точне визначення кластеру, потрібно знати не тільки умови конкретної задачі, але й те, які саме характеристики використовуються в процесі групування.

Характеристиками кластера можна назвати дві ознаки:

– внутрішня однорідність;

– зовнішня ізольованість.

## 7.2. Історія виникнення та застосування кластерного аналізу

Термін *кластерний аналіз* уперше ввів у 1939 р. Р. Тріон (Tryon) , проте активний розвиток цих методів і їхнє широке використання почався наприкінці 60-х – початку 70-х років.

Техніка кластеризації застосовується в різноманітних сферах. Хартиган (Hartigan, 1975) дав прекрасний огляд багатьох опублікованих досліджень, що містять результати, отримані методами кластерного аналізу. Він корисний, коли потрібно класифікувати велику кількість інформації.

***Наприклад:***

в економіці – для сегментації ринку, вивчення поведінки покупців, визначення конкурентоспроможності нового товару, скорочення розмірності даних тощо,

у медицині – кластеризація захворювань, їх симптомів, а також таксономія пацієнтів, препаратів тощо;

у маркетингу – задача сегментації конкурентів і споживачів,

у менеджменті – розбивка персоналу на різні групи, класифікація споживачів і постачальників, виявлення схожих виробничих ситуацій, при яких виникає брак;

у соціології задача кластеризації – розбивка респондентів на однорідні групи;

в екології – виявлення схожих екологічних умов існування людини та виробництва (стан земель, повітря, води тощо).

У цілому кластерний аналіз виявляється дуже корисним і ефективним тоді, коли необхідно класифікувати «гори» інформації до придатного для подальшої обробки вигляду.

## 7.3. Формальна постановка та основні етапи задачі кластеризації

***Кластерний аналіз*** (англ. *Data clustering*) – задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини (***кластери)*** таким чином, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися.

***Формальна постановка задачі.***

Нехай $X$ – множина об'єктів, $Z$ – множина номерів кластерів.

Задано функцію відстані між об'єктами $\rho(x, x')$.

Є скінченна вибірка об'єктів $X^m = \{x_1, ..., x_m\} \subset X$.

Потрібно розбити вибірку на підмножини ***(кластери)***, що не перетинаються, так, щоб кожен кластер складався з об'єктів, близьких по метриці $\rho$ а об'єкти з різних кластерів істотно відрізнялися. При цьому кожному об'єкту $x_i \in X^m$ приписується номер кластеру $z_i$.

***Алгоритм кластеризації*** – це функція $a: X \to Z$, яка будь-якому об'єкту $x \in X$ ставить у відповідність номер кластера $z \in Z$.

Множина $Y$ у деяких випадках відома заздалегідь, проте частіше ставиться завдання визначити оптимальне число кластерів, з погляду деякого критерію якості кластеризації.

Задача кластеризації відноситься до статистичної обробки, а також до широкого класу завдань інтелектуального аналізу даних *навчання без вчителя*.

***Основні етапи кластерноо аналізу.***

Незалежно від конкретної сфери застосування кластерного аналізу передбачає такі етапи:

- відбір вибірки для кластеризації;
- визначення множини характеристик, за якими будуть оцінюватися об'єкти у вибірці;
- обчислення значень тієї чи іншої міри схожості (метрики) між об'єктами;
- застосування одного з методів кластерного аналізу для створення груп схожих об'єктів;
- перевірка достовірності результатів кластеризації.

Якщо кластерному аналізу передує факторний аналіз, то вибірка не потребує корегування – викладені вимоги виконуються автоматично самою процедурою факторного моделювання. В іншому випадку вибірку потрібно корегувати.

## 7.4. Особливості проведення кластерного аналізу

***Основні методи кластеризації.*** Об'єднання схожих об'єктів у групи може бути здійснене різними способами. Саме для цього етапу існує ціла низка методів:

- K-середніх (K-means);
- нечітка кластеризація C-середніх (C-means);
- графові алгоритми кластеризації;
- статистичні алгоритми кластеризації;
- алгоритми сімейства FOREL;
- ієрархічна кластеризація або таксономія;
- нейронна мережа Кохонена;
- ансамбль кластеризаторів;
- алгоритми сімейства KRAB;
- EM-алгоритм;
- метод просіювання

та ін.

***Типи вхідних даних та вимоги до вхідних даних.***

Вхідними даними кластерного аналізу є набір об'єктів. Залежно від способу представлення цих об'єктів розрізняють такі типи вхідних даних:

- *вектор характеристик*. Кожен об'єкт описується набором своїх характеристик; ці характеристики можуть бути числовими або нечисловими;
- *матриця відстаней*. Кожен об'єкт описується відстанями до всіх інших об'єктів вибірки.

Кластерний аналіз висуває такі *вимоги до даних:*

- об'єкти не повинні корелювати (бути залежними) між собою;
- об'єкти мають бути безрозмірними;
- розподіл об'єктів має бути близьким до нормального;
- об'єкти повинні відповідати вимозі стійкості, під якою розуміється відсутність впливу на їх значення випадкових чинників;
- вибірка повинна бути однорідна.

***Найбільш розповсюджені метрики.***

При проведенні кластерного аналізу для визначення відстаней між двома об'єктами $x=(x_1,...,x_i,...,x_n)$ та $y=(y_1,...,y_i,...,y_n)$ використовуються, зокрема, такі метрики:

- метрика Евкліда (евклідова метрика) $d_E(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$;
- відстань Хемінга: $d_{XEM}(x,y)=\sum_{i=1}^{n}(x_i-y_i)$;

- метрика Чебишева: $d_{SUP}(x,y) = SUP|x_i - y_i|$;
- ступенева відстань: $d_S(x,y) = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{r}}$;
- відстані Джеффріса-Матусіти та ін.

Застосування кожної метрики має свої особливості при проведенні кластерного аналізу, а тому її вибір повинен бути ретельно обґрунтовано. Особливості та основні характеристики деяких з наведених метрик представлено в табл. 7.1.

Таблиця 7.1. Характеристика метрик для вимірювання відстаней

| Назва | Недоліки | Переваги |
|---|---|---|
| 1 | 2 | 3 |
| Метрика Евкліда $d_e(x_i, y_i) = \sqrt{\sum_{i=1}^{Nf}(x_i - y_i)^2}$ | Не враховує знакові розходження | 1. Пропорційно збільшує відстань між об'єктами у випадку різних абсолютних значень показників. 2. Збільшується розмірність кластерного поля, об'єкти штучно віддаляються один від одного. 3. Межі між кластерами стають чіткими і точними. |
| Метрика Хемінга $d_{\text{õåì}}(x_i, y_i) = \sum_{i=1}^{Nf}(x_i - y_i)$ | Втрачаються важливі знакові характеристики розходжень | 1. Використовується, коли знакові розходження характеристик об'єктів мають принципове значення. 2. За рахунок нівелювання знакових розходжень показників об'єкти сконцентруються навколо області ядра кластера. |
| Метрика L-норма $d_L(x_i, y_i) = \sum_{i=1}^{Nf}\|x_i - y_i\|$ | Не враховуються знакові розходження | 1. Збільшується розмірність кластерного поля. 2. Об'єкти штучно віддаляються один від одного. 3. Межі між кластерами стають чіткішими і точнішими. |
| Метрика Чебишева $d_{\sup}(x_i, y_i) = SUP\|x_i - y_i\|$ | Неправомірно змінює картину класифікації через ігнорування усіх факторів, крім одного | З усіх різниць значень факторів, взятих по модулю, обирається одна – найбільша, саме вона буде характеристикою відстані між об'єктами. Отже, чітко формується однакова відстань між об'єктами. |

## 7.5. Аналіз результатів: причини неоднозначності та інтерпретація.

Розв'язок задачі кластеризації принципово неоднозначний, і цьому є декілька причин:

- не існує однозначно найкращого критерію якості кластеризації. Відома ціла низка евристичних критеріїв, а також низка алгоритмів, що не мають чітко вираженого критерію, але здійснюють достатньо розумну кластеризацію «за побудовою». Усі вони можуть давати різні результати;
- кількість кластерів, як правило, невідома заздалегідь і встановлюється відповідно до деякого суб'єктивного критерію;
- результат кластеризації істотно залежить від метрики, вибір якої, як правило, також суб'єктивний і визначається експертом.

***Інтерпретація результатів***.

Результатом кластеризації є групи об'єктів, об'єднані за певною характеристикою чи характеристиками. Однак ці результати можуть бути інтерпретовані по-різному. Зокрема, при аналізі результатів *соціологічних* досліджень рекомендується здійснювати аналіз ієрархічними методами, наприклад методом Уорда, при якому всередині кластерів оптимізується мінімальна дисперсія, і в результаті створюються кластери приблизно рівних розмірів. Як міра відмінності між кластерами використовується квадратична евклідова відстань, що сприяє збільшенню контрастності кластерів.

Тепер виникає питання стійкості знайденого кластерного розв'язку. По суті, перевірка стійкості кластеризації зводиться до перевірки її достовірності. Тут існує емпіричне правило – ***стійка типологія зберігається при зміні методів кластеризації***. Результати ієрархічного кластерного аналізу можна перевіряти ітеративним кластерним аналізом методом k-середніх. Якщо при порівнянні групи збігаються більше, ніж на 70 % (понад 2/3 збігів), то кластерне рішення приймається.

Перевірити адекватність рішення, не вдаючись до допомоги інших видів аналізу, не можливо. Принаймні, в теоретичному плані ця проблема не вирішена.

### Питання для самоконтролю:

1. У чому полягає відмінність між кластерізацією та класифікацією?
2. Які переваги має кластерізація в порівнянні з класифікацією?
3. Задачу кластерізації можно віднести до задачі навчання з вчителем чи навчання без вчителя?
4. Яким чином визначається кількість кластерів?
5. У чому полягає причина неоднозначності розв'язку задачі кластерізації?
6. Як перевірити достовірність результатів кластерного аналізу?