Лекція 14.
Алгоритми кластеризації

Основні відомості та завдання кластеризації

Завданням кластеризації є поділ досліджуваної множини об’єктів на групи подібних об’єктів, які називаються кластерами. Слово кластер має англійське походження (cluster) і перекладається, як згусток, пучок, група. Часто поділ множини елементів на кластери називають кластер ним аналізом. Великою перевагою кластерного аналізу є те, що він дає змогу розбивати об’єкти не за однією з ознак, а за цілим їх набором. Крім того, кластер ний аналіз, на відміну від більшості математико-статистичних методів, не накладає ніяких додаткових обмежень на вид розглядуваних об’єктів, а дає можливість розглядати їх початкові дані. Це має велике теоретичне і практичне значення, наприклад, для прогнозування кон’юнктури за наявності різнорідних показників, через які складно застосувати традиційні економетричні підходи.

Кластерний аналіз дає змогу розглядати досить великий об’єм інформації і значно скорочує та стискує великі масиви інформації, робить їх компактними і наглядними. Є ряд особливостей, притаманних кластеризації. По-перше, кластеризація дуже залежить від природи об’єктів даних. Так, з одного боку, це можуть бути однозначно визначені, чітко кількісно окреслені об’єкти, а з іншого – об’єкти, які мають ймовірнісні або нечіткі множини. По-друге, кластеризація дуже залежить також від кластерів і припущених відношень об’єктів даних і кластерів. Так, необхідно брати до уваги такі властивості, як можливість – неможливість належності об’єктів декільком кластерам. Необхідно визначити саме розуміння належності кластеру: однозначність (належить – не належить), ймовірність (ймовірність належності), нечіткість (ступінь належності). В зв’язку з цим було розроблено декілька способів її вирішення. Одним з таких способів є аобудова набору характеристичних функцій класів, які показують, належить об’єкт до певного класу чи ні. Характеристична функція класів може бути двох типів:

•Дискретна функція, яка приймає одне з двох визначених значень, суть яких в належності – неналежності об’єкта заданому класу;
•Функція, яка набуває значення, наприклад, в інтервалі 0 – 1. Чим ближче значення функції до одиниці, тим більше об’єкт належить до заданого класу.
Спільний підхід до розв’язання задач кластеризації став можливим після розвитку теорії нечітких множин. Визначити нечіткі взаємозв’язки даних можна різними способами. Один це – визначення взаємозв’язків даних через їхнє відношення до деяких еталонних зразків, які називають центрами кластерів, але взаємозв’язки між даними в умовах невизначеності можна враховувати і за допомогою нечітких відношень між окремими зразками даних, не використовуючи при цьому поняття центра кластера. Другий підхід більш універсальний, тому його ми розглянемо детальніше на конкретному прикладі.

Приклад. Заданий набір даних із такими властивостями: кожний екземпляр даних задається чіткими числовими значеннями; клас для кожного конкретного екземпляра даних невідомий. Необхідно визначити: спосіб порівняння даних між собою; спосіб кластеризації; поділ даних за кластерами.

Розв'язання. Формально кластеризація описується так. Задана множина об'єктів даних *I*, кожний з яких представлений набором атрибутів. Необхідно побудувати множину кластерів *C* і відображення *F* множини *I* на множину *C*, тобто: $F : I \to C$

Відображення *F* задає моель даних, яка є розв'язком задачі. Якість розв'язання задачі визначається кількістю i_j правильно класифікованих об'єктів даних. Множину *I* визначають так: $I = \{i_1, i_2, ..., i_j, ..., i_n\}$

де – i_j

досліджуваний об'єкт. Прикладом такої множини може бути набір даних про іриси, з якими працював у середині 30-х років минулого сторіччя відомий статист Р.А. Фішер. Він розглядав три класи ірисів – Iris setosa, Iris versicolor, Iris virginica. Кожного з них було взято по 50 екземплярів із різними значеннями параметрів, а вибірково – по 5 екземплярів кожного класу наведено в таблиці.

№ об’єкта	Довжина чашолистика	Ширина чашолистика	Довжина пелючтки	Ширина пелюстки	Класи ірисів
1	5,1	3,5	1,4	0,2	Iris setosa
2	4,9	3,0	1,4	0,2	Iris setosa
3	4,7	3,2	1,3	0,2	Iris setosa
4	4,6	3,1	1,5	0,2	Iris setosa
5	5,0	3,6	1,4	0,2	Iris setosa
51	7,0	3,2	4,7	1,4	Iris versicolor
52	6,4	3,2	4,5	1,5	Iris versicolor
53	6,9	3,1	4,9	1,5	Iris versicolor
54	5,5	2,3	4,0	1,3	Iris versicolor
55	6,5	2,8	4,6	1,5	Iris versicolor
101	6,3	3,3	6,0	2,5	Iris virginica
102	5,8	2,7	5,1	1,9	Iris virginica
103	7,1	3,0	5,9	2,1	Iris virginica
104	6,3	2,9	5,6	1,8	Iris virginica
105	6,5	3,0	5,8	2,2	Iris virginica

Кожний із об’єктів характеризується набором параметрів $i_j = \{\chi_1, \chi_2, ..., \chi_k, ..., \chi_m\}$
В ірісів такими параметрами є довжина і ширина чашолистика та пелюстки. Кожна змінна може набувати значення з деякої множини
$\chi_k = \{\nu_k^1, \nu_k^2, ...\}$

,які є дійсними числами. Завдання кластеризації є побудова множини
$C = \{c_1, c_2, ..., c_h, ..., c_z\}$

, де c_h
-кластер, який має подібні об’єкти з множини I: $c_k = \{i_j, i_p : i_j \in I, i_p \in Iii\ (i_j - i_p) < \delta\}$
-, де δ
-- величина, яка визначає міру близькості для включення об’єктів в один кластер, а $d(i_j - i_p)$
-- міру близькості між об’єктами, яку називають відстанню. Невід’ємні значення $d(i_j - i_p)$
називають відстанню між елементами i_j
і i_p
, якщо виконуються такі умови:

1) $d(i_j - i_p) \geq 0$
для всіх $i_j \quad i_p$
2)
$$d(i_j - i_p) = 0$$
тоді і тільки тоді, коли $i_j = i_p$
3) $d(i_j - i_p) = d(i_j - i_p)$
4)
$$d(i_j - i_p) = d(i_j - i_r) + d(i_r - i_p)$$
Якщо відстань $d(i_j - i_p)$
менша від деякого значення δ
, то кажуть, що елементи перебувають близько один від одного і розміщуються в одному кластері. В протилежному випадку кажуть, що елементи відрізняються один від одного і їх розміщують у різні кластери. Більшість популярних алгоритмів, які розв'язують задачі кластеризації, використовують, як формат вхідних даних матрицю відмінності D. Прядки і стовпчики такої матриці відповідають елементам множини I. Елементами матриці є значення $d(e_j - e_p)$

в рядку j і стовпчику p. Очевидно, що на головній діагоналі такої матриці значення дорівнюватимуть нулю.

$$D = \begin{pmatrix} 0 & d(e_1 - e_2) & \dots & d(e_1 - e_n) \\ d(e_2 - e_1) & 0 & \dots & d(e_2 - e_n) \\ \dots & \dots & \dots & \dots \\ d(e_n - e_1) & d(e_n - e_2) & \dots & 0 \end{pmatrix}$$

Більшість алгоритмів працюють із симетричними матрицями, але, якщо матриця не симетрична, то її можна привести до симетричного вигляду шляхом перетворення

$$\left(D + D^{m}\right) / 2$$

Класифікація алгоритмів

Означення 1. **Алгоритмом кластеризації** називають алгоритм, за допомогою якого розділяють досліджувану множину об'єктів на групи подібних між собою.

При виконанні клоастеризації важливо знати, скільки кластерів має бути побудовано. Передбачається, що кластеризація повинна виявити природні локальні згущення об'єктів. Тому, кількість кластерів є параметром, який часто істотно ускладнює вибір виду алгоритму, якщо кількість кластерів невідома, та істотно впливає на якість результату, якщо воно відоме.

Проблема вибору кількості кластерів дуже нетривіальна. Досить сказати, що для отримання задовільного теоретичного розв'язання часто слід зробити дуже великі припущення про властивості деякої наперед заданоїх родини розподілів. Але про які припущення може йтися, якщо, особливо на початку дослідження, про властивості практично нічого невідомо? Тому алгоритми кластеризації зазвичай будуються як деякий спосіб перебору кількості кластерів і визначення їхніх оптимальних значень в процесі перебору. Кількість методів розбиття множини на кластери досить велика. Усі їх можна розподілити на ієрархічні та неієрархічні.

Означення 2. **Неієрархічним алгоритмом кластеризації** називають алгоритм, у якому характер роботи і умова припинення регламентована заздалегідь великою кількістю параметрів або значеннями завдань угрупування ознак при їхній великій кількості.

Означення 3. **Ієрархічним алгоритмом кластеризації** називають алгоритм, у якому характер його роботи регламентований побудовою повного дерева вкладених кластерів, тобто дендограмою алгоритму.

Ієрархічні алгоритми пов'язані з побудовою дендограм поділяють на агломеративні та диві зимні (ділимі).

Означення 4. **Агломеративним алгоритмом кластеризації** називають алгоритм, для якого характерне послідовне об'єднання початкових елементів і відповідне зменшення кількості кластерів (побудову кластерів виконують із низу вгору).

Означення 5. **Дивізимним алгоритмом кластеризації** називають алгоритм, у якого кількість кластерів зростає, починаючи з одного, внаслідок чого утворюється послідовність розщеплених груп (побудову кластерів виконують зверху вниз).