Загальні поняття кластеризації

Кластеризація, або кластерний аналіз — це статистична процедура, задача якої полягає в розбитті вибірки об'єктів на підмножини, що не перетинаються і називаються кластерами. Кожен кластер має складається зі схожих об'єктів, а об'єкти різних кластерів мають істотно відрізнятися один від одного.

Задача кластеризації відноситься до статистичної обробки, а також до широкого класу задач навчання без вчителя. Ще її можна описати через задачу класифікації. Задача кластеризації — це по факту задача класифікації, бо в обох випадках ми ділимо об'єкти на основі їх подібності між собою, але у випадку кластеризації приналежність навчальних об'єктів будь-яким класам не задається. Така задача — загальна, тому для її розв'язання використовуються різні підходи.

Алгоритми побудови кластерів можуть дуже відрізнятися у підходах до того, що відносити в один кластер і як їх взагалі ефективніше шукати. Кластери можна утворювати ґрунтуючись на відстані між ними, на щільності ділянок у просторі даних, інтервалах або на конкретних статистичних розподілах. Це все залежать від конкретного набору даних та мети використання результатів. Кластерний аналіз не є автоматизованим, це скоріше ітераційний процес, тому що часто доводиться змінювати метод опрацювання даних та параметри моделі, поки не буде отримано з результат з заданими властивостями.

Розв'язок неоднозначний, і на це є кілька причин. По-перше, не існує найкращого критерію якості кластеризації. Відомий цілий ряд досить ефективних критеріїв, а також ряд алгоритмів, які не мають чітко вираженого критерію, але все одно здійснюють досить якісну кластеризацію по побудові. Всі вони можуть давати різні результати. По-друге, число кластерів, як правило, не відомо заздалегідь і встановлюється відповідно до деякого суб'єктивного критерія. По-третє, результат кластеризації істотно залежить від метрики ρ, вибір якої, як правило, також суб'єктивний і визначається спеціалістом.

Задача групування набору об'єктів полягає в тому, що об'єкти в одному кластері більш схожі один на одного, ніж об'єкти в інших кластерах. Подібність — це буквально кількість, яка собою відображає міцність взаємозв'язку між двома об'єктами. Кластеризація використовується в основному для видобутка даних, а також в інших

областях, таких як машинне навчання, розпізнавання образів, аналіз зображень, пошук інформації, біоінформатика, стиснення даних і комп'ютерна графіка. Існує два типа кластеризації: жорстка та м'яка. У жорсткій кластеризації кожен об'єкт даних або повністю належить кластеру чи взагалі не належить. В м'якій кластеризації точка чи об'єкт даних може з певною ймовірністю належати більш ніж одному кластеру.

Дані, що використовуються в кластерному аналізі, можуть мати інтервальний, порядковий або категоріальний тип. Однак наявність суміші різних типів змінної зробить аналіз більш складним. Це пояснюється тим, що в кластерному аналізі необхідно мати певний спосіб вимірювання відстані між спостереженнями, тобто тип використовуваної міри залежить від типу даних.

Кластеризація — це досить суб'єктивна задача, для розв'язання якої може існувати більше одного правильного алгоритму. Кожен алгоритм слідує своєму набору правил для визначення «подібності» між об'єктами даних. Найбільш відповідний алгоритм кластеризації для конкретної проблеми часто потрібно вибирати експериментально, якщо немає математичної причини віддати перевагу одному алгоритму кластеризації над іншим. Алгоритм може добре працювати на певному наборі даних, але не працювати для іншого.

Ієрархічна кластеризація

Ієрархічні алгоритми кластеризації, або алгоритми таксономії, будують не одне розбиття вибірки на непересічні класи, а систему вкладених розбиттів. Результат таксономії зазвичай представляється у вигляді таксономічного дерева — дендрограми. Класичним прикладом такого дерева є ієрархічна класифікація тварин і рослин.

Дендограми дозволяє уявити кластерну структуру у вигляді плаского графіка незалежно від того, яка розмірність початкового простору. Існують і інші способи візуалізації багатовимірних даних, такі як багатовимірне шкалювання або карти Кохонена, але вони привносять в картину штучні спотворення, вплив яких досить важко оцінити. Є два типи методів:

1. Агломератні методи: нові кластери утворюються шляхом об'єднання дрібніших кластерів, і таким чином дерево створюється від листя до стовбура.

2. Дивізійні методи: нові кластери створюються шляхом ділення більших кластерів на більш дрібні, і таким чином дерево створюється від стовбура до листя.

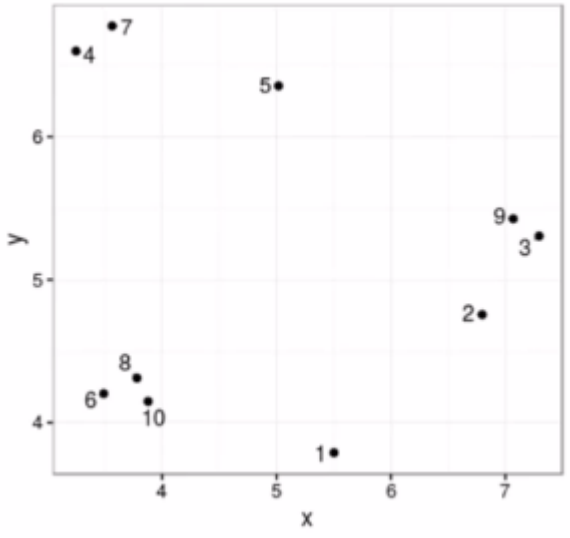
Подібність кластерів часто розраховується через «неподібність», наприклад, евклідова відстань між двома кластерами. Отже, чим більше відстань між двома кластерами, тим краще. Ключовою операцією в ієрархічній агломераційній кластеризації є неодноразове об'єднання двох найближчих кластерів у один кластер, але дуже важливо спочатку відповісти на три питання: як ви представити кластер з більш ніж однією точкою, як визначити «близькість» кластерів та коли перестати поєднувати кластери.

Алгоритми ієрархічної кластеризації припускають, що аналізована множина об'єктів характеризується певним ступенем зв'язності. За кількістю ознак іноді виділяють монотетічені та політетічні методи класифікації. Як і більшість візуальних способів подання залежностей графи швидко втрачають наочність при збільшенні числа кластерів. Найчастіше використовуються агломеративні методи.

Злиття кластерів припиняється в залежності від доступної інформації про дані, які ми маємо. Якщо групувати футболістів на полі на основі їхніх позицій на полі, яке представлятиме їх координати для розрахунку відстані між гравцями, очевидно, що треба зупинитися на лише двох кластерах, оскільки можуть бути тільки дві команди, які грають у футбольний матч.

Алгоритм методу виглядає таким чином:

1. Припустимо, що множина точок представлена в наступному просторі:



2. Обчислити матрицю близькості, що містить відстань між кожною парою шаблонів. Розглядати кожен зразок як окремий кластер.
3. Знайти найбільш схожу пару кластерів за допомогою матриці близькості. Об'єднати ці два кластера в один більший кластер. Центр даного кластеру визначити як середньоарифметичне по всім параметрам. Оновити матрицю близькості, щоб відобразити цю операцію злиття.

```
      1    2    3    4    5    6    7    8    9
2  1.62
3  2.35 0.74
4  3.60 4.00 4.25
5  2.61 2.39 2.51 1.79
6  2.05 3.35 3.96 2.41 2.64
7  3.56 3.81 4.01 0.36 1.51 2.57
8  1.80 3.05 3.65 2.35 2.39 0.31 2.47
9  2.26 0.72 0.26 4.00 2.25 3.78 3.75 3.47
10 1.66 2.98 3.61 2.53 2.48 0.39 2.65 0.19 3.44
```

3. Якщо всі шаблони знаходяться в одному кластері, зупинитися. В іншому випадку перейти до кроку 2. Графічно процес об'єднання (дерево) можна зобразити у наступному вигляді:

Або у вигляді матриці об'єднання:

A	B	C	D
Об'єкт 1	Об'єкт 2	Відстань між кластерами	Нова назва
8	10	0.19	10
3	9	0.26	9
10	6	0.32	6
4	7	0.36	7
9	2	0.41	2
7	5	0.46	5
2	1	0.58	1
1	6	1.12	6
6	5	4.0	5

Ієрархічний алгоритм дає дендограму, що представляє собою складене групування шаблонів і рівні схожості, на яких змінюються самі групування. Більшість ієрархічних алгоритмів кластеризації є варіантами однозв'язного і повнозв'язного алгоритму, а також алгоритму мінімальної дисперсії.

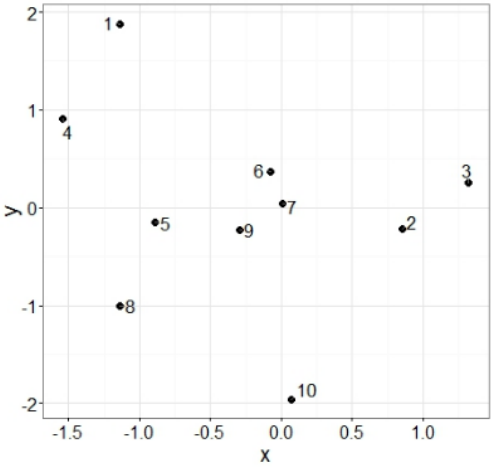
Найбільш популярними з них є однозв'язний та повнозв'язний алгоритми. Вони відрізняються тим, як вони характеризують подібність між парою кластерів.

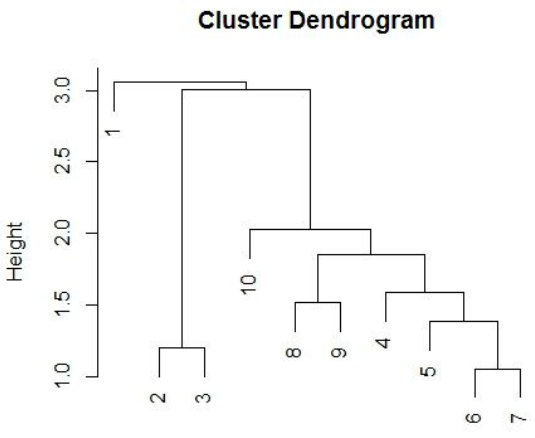
У методі з однозв'язним алгоритмом відстань між двома кластерами — це мінімум відстаней між усіма парами шаблонів, взятих з двох кластерів (один з першого кластера, другий — з другого).

У повнозв'язному алгоритмі відстань між двома кластерами — це максимум усіх попарних відстаней між шаблонами в двох кластерах.

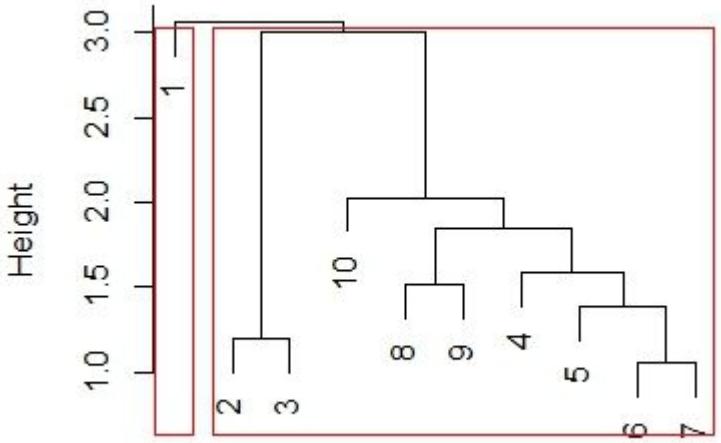
В обох випадках два кластери об'єднуються для формування більшого кластера на основі мінімальних критеріїв відстані. Повнозв'язний алгоритм створює щільні та компактні кластери, а однозв'язний алгоритм, навпаки, страждає від ланцюгового. Він має тенденцію виробляти кластери, які є занадто громіздкими або подовженими.

Особливу увагу варто приділити тому, як відбувається вибір числа кластерів при ієрархічній кластеризації. Розглянемо приклад кластеризації наступного набору даних:

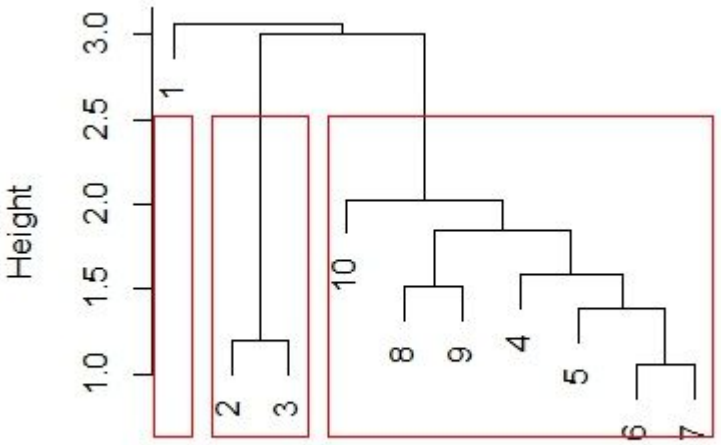




Уявіть, що ми подумки продовжили кожну "гілку" до самого низу дендограмм. А тепер починаємо робити зрізи на графіку зверху вниз. Першим зрізом ми перетнемо дві "гілки". Це і буде варіант поділу на два кластери.



Тепер давайте рухатися далі вниз, поки ми не зріжемо вже три "гілки".



Рішення для 4, 5 і 6 кластерів буде виглядати наступним чином: