КЛАСТЕРНИЙ АНАЛІЗ МЕТОДОМ К-СЕРЕДНІХ

Скіцан Ольга Дмитрівна, студентка групи 355а
Національний аерокосмічний університет ім. М.Є. Жуковського «ХАІ»

Кластерний аналіз (англ . Data clustering) — задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, звані кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних істотно суттєвих кластерів відрізнялися.

Задача кластеризації відноситься до статистичної обробки. Кластерний аналіз — це багатовимірна статистична процедура, що виконує збір даних, що містять інформацію про вибірку об'єктів, і яка упорядковує об'єкти в порівняно однорідні групи (кластери) (Q-кластеризація, або Q-техніка, власне кластерний аналіз). Кластер — група елементів, що характеризуються загальною спільною властивістю, головна ціль кластерного аналізу — знаходження груп схожих об'єктів у вибірці.

Сфера використання кластерного аналізу, через його універсальності, дуже широка. Кластерний аналіз застосовують в економіці, маркетингу, археології, медицині, психології, хімії, біології, державному управлінні, філології, антропології, соціології та інших областях.

Ось кілька прикладів застосування кластерного аналізу:

1) медицина - класифікація захворювань, їх симптомів, способів лікування, класифікація груп пацієнтів;
2) маркетинг - завдання оптимізації асортиментної лінійки компанії, сегментація ринку по групах товарів або споживачів, визначення потенційного споживача;
3) соціологія - розбиття респондентів на однорідні групи;
4) психіатрія - коректна діагностика груп симптомів є вирішальною для успішної терапії;
5) біологія - кластеризація організмів по групі;
6) економіка - кластеризація суб'єктів РФ по інвестиційній привабливості.

Існує величезна безліч алгоритмів для кластеризації даних.

Загальноприйнятою класифікації методів кластеризації не існує, але можна виділити ряд груп підходів

1) імовірнісний підхід
2) підходи на основі систем штучного інтелекту
3) логічний підхід
4) теоретико-графовий підхід
5) ієрархічний підхід

Один з найпоширеніших алгоритмів кластеризації - метод кластеризації методом k-середніх.

Метод k-середніх - це метод кластерного аналізу, мета якого є поділ спостережень (з простору) на k кластерів, при цьому кожне спостереження відноситься до того кластеру, до центру (центроїду) якого воно найближче.

В якості запобіжного близькості використовується Евклідова відстань:

$$\rho(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^{n}(x_p - y_p)^2}, \text{ где } x, y \in R^n$$

Отже, розглянемо ряд спостережень $(x^{(1)}, x^{(2)}, \ldots, x^{(m)})$, $x^{(j)} \in R^n$.

Метод k-середніх розділяє m спостережень на k груп (або кластерів) (k ≤ m) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$, щоб мінімізувати сумарне квадратичне відхилення точок кластерів від центроїдів цих кластерів:

$$\min\left[\sum_{i=1}^{k}\sum_{x^{(j)} \in S_i}\left\|x^{(j)} - \mu_i\right\|^2\right], \text{ где } x^{(j)} \in R^n, \mu_i \in R^n$$

μ_i - центроїд для кластера S_i.

Отже, якщо міра близькості до центроїда визначена, то розбиття об'єктів на кластери зводиться до визначення центроїдів цих кластерів. Число кластерів k задається дослідником заздалегідь.

Розглянемо початковий набір k середніх (центроїдів) $\mu_1, \ldots, \mu_k$ в кластерах $S_1, S_2, \ldots, S_k$.

На першому етапі центроїди кластерів вибираються випадково або за певним правилом (наприклад, вибрати центроїди, максимізує початкові відстані між кластерами).

Потім центр ваги кожного i-го кластера переобчислюють за таким правилом:

$$\mu_j = \frac{1}{s_j}\sum_{x^{(j)} \in S_j} x^{(j)}$$

Таким чином, алгоритм k-середніх полягає в переобчислення на кожному кроці центроїда для кожного кластера, отриманого на попередньому кроці.

Отже, ще раз підкреслимо деякі особливості методу k-середніх:

Як метрики використовується Евклідова відстань

Число кластерів заздалегідь не відомо і вибирається дослідником заздалегідь

Якість кластеризації залежить від початкового розбиття.

*Науковий керівник – Чумаченко Д.І., к.т.н., доцент, доцент кафедри математичного моделювання та штучного інтелекту ХАІ.