

# ЗАСТОСУВАННЯ КЛАСТЕРИЗАЦІЇ В ЗАДАЧАХ НОРМАЛІЗАЦІЇ ЗОБРАЖЕНЬ

Пригодій А.І.

Науковий керівник – к.т.н., доц. Кобилін О.А.

Харківський національний університет радіоелектроніки  
(61166, Харків, пр. Науки, 14, каф. Інформатики, тел. (057) 702-14-19)  
e-mail: anna.pryhodii@nure.ua

The development of image processing techniques leads to the emergence of all new approaches to solving segmentation problems. For example, the problem of separation of adjoining objects in the image is still relevant. Among data mining tasks, clustering tasks occupy a special place. It should be noted that when solving clustering problems, the most popular algorithms are based on optimal partitioning of a set of data into clusters. Such algorithms are aimed at grouping data into clusters so that the objective function of the partitioning algorithm reaches an extremum (minimum). In this regard, it is possible to distinguish the C-means algorithm.

Кластеризація (або кластерний аналіз) – це задача розбиття множини об'єктів на групи, які називаються кластерами. У кожній групі повинні бути «схожі» об'єкти, а об'єкти різних груп повинні бути максимально різними. Основною особливістю кластеризації є те, що перелік груп чітко не заданий і формується у ході роботи алгоритму. Застосування кластерного аналізу зводиться до наступних етапів:

1. Відбір вибірки об'єктів для кластеризації.
2. Виявлення множини змінних, за якими будуть оцінюватися об'єкти у вибірці. При необхідності-нормалізація значень змінних.
3. Обчислення значень міри збіжності між об'єктами.
4. Застосування методу кластерного аналізу для створення груп схожих об'єктів (кластерів).
5. Отримання результатів аналізу.

Після отримання і аналізу результату можливе коректування метрики і методу кластеризації для отримання оптимального результату.

Для визначення «схожості» об'єктів потрібно створити вектор характеристик для кожного об'єкта – зазвичай, це набір числових значень. Після цього можна виконати нормалізацію для того, щоб всі компоненти давали однаковий внесок при розрахунку «відстані». В процесі нормалізації усі об'єкти приводяться до деякого діапазону, наприклад [-1,-1] або [0,1]. Вимірюється «відстань» між об'єктами-ступінь схожості:

1. Евклідова відстань:

$$\rho(x, x') = \sqrt{\sum_i^n (x_i - x'_i)^2}.$$

2. Квадрат евклідової відстані:

$$\rho(x, x') = \sum_i^n (x_i - x'_i)^2.$$

3. Відстань міських кварталів:

$$\rho(x, x') = \sum_i^n |x_i - x'_i|.$$

4. Відстань Чебишева:

$$\rho(x, x') = \max(|x_i - x'_i|).$$

5. Ступенева відстань:

$$\rho(x, x') = \sqrt[p]{\sum_i^n (x_i - x'_i)^p},$$

Вибір метрики повністю залежить від дослідника, оскільки результати кластеризації можуть суттєво відрізнятись при застосуванні різних метрик.

Найбільш популярним алгоритмом нечіткої кластеризації є алгоритм с-середніх (с-means). Він є модифікацією алгоритму k-середніх. Робота алгоритму:

1. Обрати початкове нечітке розбиття n об'єктів на k кластерів шляхом вибору матриці належності U розміром n x k.
2. Використовуючи матрицю U, - знайти значення критерію нечіткої похибки:

$$E^2(X, U) = \sum_{i=1}^N \sum_{k=1}^K U_{ik} \|x_i^{(k)} - c_k\|^2, \quad c_k = \sum_{i=1}^N U_{ik} x_i.$$

3. Перегрупувати об'єкти з метою зменшення значення критерію нечіткої похибки.
4. Повертатися до п.2 допоки зміни матриці U не стануть незначними.

Після того, як зображення розбиті на кластери відповідно до груп перетворень є можливість обирати відповідні нормалізатори для нормалізації цих зображень.

### Список використаних джерел:

1. Воронцов К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования //М.: МГУ. – 2007.
2. Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review //ACM computing surveys (CSUR). – 1999. – Т. 31. – №. 3. – С. 264-323.
3. Котов А., Красильников Н. Кластеризация данных //Котов А., Красильников. – 2006. – №. 2006.
4. Мандель И. Д. Кластерный анализ //М.: Финансы и статистика. – 1988. – Т. 176.