УДК 004.79

ФРАТАВЧАН Т.М., АНТОНЮК С.В., ФРАТАВЧАН В.Г.
Чернівецький національний університет ім. Ю.Федьковича,
(Україна)

ПОБУДОВА КЛАСТЕРИЗАТОРА НА ОСНОВІ ФОРМ ЕРМІТА ДЛЯ КЛАСТЕРІВ ЗІ СКЛАДНОЮ КОНФІГУРАЦІЄЮ

Пропонується алгоритм для кластеризації множин у випадках неопуклих, багатомодальних областей та областей із взаємними перетинами. Для кластеризації використовуються інтерполяційні форми Ерміта.

Кластерний аналіз (кластеризацію даних) розуміють як задачу розбиття деякої вибірки об'єктів на підмножини (кластери) таким чином, щоб екземпляри одного кластеру були схожі між собою за деяким набором критеріїв, а екземпляри різних кластерів відрізнялися істотно. Якщо об'єкти задаються векторами характеристик, мірою схожості можна вважати декартову, чебишевську або манхетенську відстань.

Для процесу кластеризації розроблено багато методів та алгоритмів, які є ефективними, якщо області кластерів мають компактну локальну форму. У випадках, коли області локалізації кластерів мають складну нерегулярну форму (багатомодальну, спіралевидну), задача кластеризації суттєво ускладнюється.

В роботі пропонується алгоритм кластеризації для випадків, коли об'єкти задаються векторами числових дійсних характеристик і кількість кластерів наперед відома. Самі кластери можуть мати нетривіальну топологічну конфігурацію (кільцеподібну, спіралеподібну, кометоподібну форму).

Як відомо, процедура кластеризації визначається як функція, яка кожному екземпляру x вибірки X ставить у відповідність номер класу:

$$f: x \to c, \quad x \in X, c \in C,$$

де x – багатовимірний вектор характеристик об'єкта;

c – номер класу;

X – загальна вибірка;

C – множина номерів класів.

Для визначення функції кластеризації кожному кластеру у просторі характеристик ставиться у відповідність апроксимуюча параметрична крива третього порядку – багатовимірна інтерполяційна форма Ерміта. Для визначення інтерполяційної форми поточного кластера потрібно вказати чотири вектори – координати початкової та кінцевої точок, а також початковий і кінцевий напрямок:

$$P(t) = (2t^3 - 3t^2 + 1)p_0 + (t^3 - 2t^2 + t)m_0 + (-2t^3 + 3t^2)p_1 + (t^3 - t^2)m_1,$$
$$t \in [0,1].$$

Тут $P(t)$ – багатовимірна точка у просторі характеристик;

p_0, p_1 – початкова та кінцева точки кривої Ерміта;

m_0, m_1 – початковий та кінцевий напрямок кривої Ерміта.

Оцінка належності об'єкта до класу визначається відстанню від вектора характеристик об'єкта до інтерполяційної кривої Ерміта цього класу.

Кластеризація здійснюється як варіант машинного навчання. Розглядаються два варіанти навчання – навчання з учителем та самонавчання. При навчанні для кожного класу визначаються

параметри кривих Ерміта, тобто для кожного кластеру вибираються початкова та кінцева точка, початковий та кінцевий напрямок.

У випадку навчання з учителем з кожного кластеру вибирається репрезентативна навчальна послідовність об'єктів (навчальна вибірка). Крива Ерміта вибирається таким чином, щоб сумарна відстань від векторів навчальної послідовності до кривої була мінімальною.

У випадку самонавчання, для *k* кластерів одночасно вибираються *k* кривих Ерміта. Об'єкти загальної вибірки розподіляються по кластерах так, щоб сумарна відстань від векторів характеристик до інтерполяційних кривих класів була мінімальною.

Визначення коефіцієнтів у кривих Ерміта для кожного класу, як для навчання з учителем, так і при самонавчанні, здійснюється як оптимізаційний процес. Цільовою функцією в оптимізаційній задачі є сумарна відстань від точок вибірки до кривих Ерміта. Параметрами оптимізації виступають початкові та кінцеві точки і початкові та кінцеві напрямки кривих Ерміта:

$$D(P_0, P_1, M_0, M_1) \to \min,$$

$$P_0, P_1, M_0, M_1 \subset \Omega \subset R^n,$$

де $D(P_0, P_1, M_0, M_1)$ – сумарна відстань від точок вибірки до кривих Ерміта кожного з кластерів;

P_0, P_1, M_0, M_1 – множини початкових та кінцевих точок і множини початкових та кінцевих напрямків у формах Ерміта;

Ω – область локалізації кластерів та параметрів у просторі характеристик;

n – розмірність простору характеристик.

Очевидно, що при навчанні з учителем оптимізаційна задача є простішою за аналогічну задачу при самонавчанні, оскільки при навчанні з учителем можна визначити форму Ерміта для кожного кластеру окремо. Цільова функція при такому підході має $4*n$ параметри, а вибірка обмежена навчальною послідовністю одного класу. При самонавчанні кількість параметрів збільшується до $4*n*k$ (n – розмір простору, k – кількість кластерів), а вибірка містить всі екземпляри вхідної множини. Область Ω обмежується n-сфероїдом, що містить загальну вибірку.

Для розв'язування оптимізаційної задачі пропонується використання генетичного алгоритму. Генетичний алгоритм можна застосувати як для навчання з учителем, так і для самонавчання. Для великих розмірів простору характеристик, з урахуванням формату вхідних даних, можна застосувати модифікацію генетичного алгоритму зі спрощеними операціями кросовера і мутації [1, 2], який для даної задачі має покращені показники швидкодії.

ПЕРЕЛІК ЛІТЕРАТУРИ

1. Valerii FRATAVCHAN, Tonia FRATAVCHAN. One Pattern Recognition Method for Complex Geometric Clusters Configuration. Proceedings of the 14th International Conference on Development and Application Systems, DAS 2018. (24-26, May 2018, Suceava – Romania), pp.200-203. URL: http://www.dasconference.ro/dvd2018/data/papers/D51-paper.pdf

2. Фратавчан В.Г., Фратавчан Т.М., Лукашів Т.О., Літвінчук Ю.А. Методи та системи штучного інтелекту: навчальний посібник. Чернівці: ЧНУ, 2023, – 115 с.