Струков В.М., к.т.н,
зав. кафедри інформаційних технологій
Харківський національний університет внутрішніх справ

КРИТЕРІЙ І АЛГОРИТМ ВИБОРУ ПЕРВИННИХ ЦЕНТРІВ КЛАСТЕРІВ В АЛГОРИТМІ КЛАСТЕРИЗАЦІЇ K-MEANS

Останнім часом незворотною є тенденція стрімкого зростання потоків інформації у віртуальному просторі і необхідністю її накопичення та обробки. Обсяги накопиченої інформації досягають десятки і сотні мільйонів записів (десятки терабайтів) і з кожною миттю продовжують зростати. За таких обставин все більш актуальним стає застосування методів аналітичної обробки великих масивів даних Data Mining, Web Mining [1]. В основі цих наукоємних технологій лежать методи класифікації-кластеризації. Найбільш поширеними є алгоритми під загальною назвою K-means. Головний недолік цих алгоритмів – заздалегідь задана кількість кластерів, на які необхідно розбити вихідний набір об'єктів. Крім того, в алгоритмах цього класу на першій стадії при визначенні первинних центрів шуканих кластерів, як правило, випадковим чином обираються К об'єктів. У випадах надвеликих обсягів оброблюваних даних такий підхід неприпустимий, оскільки з великою вірогідністю передбачає «погане» початкове рішення здачі визначення первинних кластерів. А в алгоритмах K-means початкое рішення має визначальне значення. Автору невідомі ефективні алгоритми визначення первинних центрів кластерів.

Припустимо, що кластери добре виражені, тобто є угрупування даних, які досить чітко відокремлені одне від одного. Крім того, дані мають надвелику розмірність і вимірюються у числовій шкалі.

Пропонується підзадачу визначення первинних центрів кластерів вирішувати як задачу оптимізації деякого критерію, що забезпечує максимальне розсіювання первинних центрів кластерів. Обгрунтуванням вибору такого критерію є припущення про те, що кластери добре виражені (виділені і відокремлені), отже, їх реальні центри розташовані в такий спосіб, що вони доставляють екстремум функції розсіювання деяких параметрів первинних центрів кластерів. В якості такого параметру пропонується використовувати показник середньоквадратичного відхилення від середньої відстані між первинними центрами.

Нехай $X=\{x_{ij}\} \in R$ - матриця «об'єкт-властивість», де x_{ij} – значення j-ї властивості i-го об'єкта a, i=1,2, …, m; j=1,2, …, n, при цьому кожний об'єкт може бути описаний вектором-рядком властивостей $x_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \in R$. Вважатимемо, що значення всіх n ознак задані у звичайній числовій шкалі, хоча, як показано в [2], до такого випадку можна звести і ситуацію, коли значення ознак об'єктів задані в інших шкалах.

K – кількість кластерів у задачі;

$X_c=\{x_i^c\}$, i=1, 2, …, K – множина шуканих центрів кластерів.

Нехай також у просторі об'єктів задана відстань між ними у звичайній евклідовій метриці:

$$d_{il} = \sqrt{\sum_{j=1}^{n} (x_{ij} - x_{lj})^2}. \qquad (1)$$

Тоді через d_{ij}^c позначимо відстань між центрами кластерів x_i^c і x_j^c. Відповідно, середня відстань від об'єкта x_i^c до усіх інших центрів (об'єктів множини X_c) може бути записана у вигляді:

$$d_i^c = \frac{1}{K-1} \sum_{x_j^c \in X^c, x_j^c \neq x_i^c} d_{ij}^c. \qquad (2)$$

Таким чином, цільова функція у підзадачі пошуку первинних центрів К кластерів буде мати наступний вигляд:

$$f^c = \sum_{j=1}^{K} (d_j^c - \frac{1}{K} \sum_{i=1}^{K} d_i^c)^2 \to \min. \qquad (3)$$

В основу запропонованого алгоритму покладені наступні принципи:

- перший центр обирається як один з двох найбільш близьких об'єктів;
- другий центр обирається як найбільш віддалений від першого центру;
- всі наступні первинні центри обираються за критерієм (3).

Експериментальне дослідження програмної реалізації запропонованого підходу на тестових прикладах підтвердило його ефективність.

Список використаних джерел

1. Aggarwal C.C., Reddy C.K. Data clustering. Algorithms and Applications. Cham: Springer Ltd. Publ., Switzerland, 2015.- 734p.

2. Бодянский Е.В., Струков В.М., Узлов Д.Ю. Задача оценки близости многомерных объектов анализа данных. Управляющие системы и машины. – 2016.- № 6. – С. 67-72.