УДК 519.237.8

Упрощенный показатель силуэта кластерной структуры

В.В. Журавлева, А.А. Куракина
АГУ, г. Барнаул

Методы кластеризации предназначены для разбиения совокупности объектов на однородные группы (кластеры или классы). Кластер можно охарактеризовать как группу объектов, имеющих общие свойства, среди которых выделяют внутреннюю однородность объектов в кластере (или компактность) и изолированность объектов разных кластеров (отделимость) [1–2].

Показатель «Силуэт» каждого кластера определяется следующим образом [3]: пусть объект x_j принадлежит кластеру c_p. Обозначим среднее расстояние от этого объекта до других объектов из того же кластера c_p через a_{pj}. Теперь обозначим среднее расстояние от x_j до объектов из другого кластера c_q ($q \neq p$) через d_{qj}. Зададим

$$b_{p_j} = \min d_{q_j} \tag{1}$$

как меру несхожести выбранного объекта с ближайшим кластером. Таким образом, «силуэт» каждого отдельного объекта определяется по формуле

$$S_{x_j} = (b_{p_j} - a_{p_j}) / \max\left(a_{p_j}, b_{p_j}\right). \tag{2}$$

Значения показателя силуэта ограничены отрезком [−1; 1]. Очевидно, что высокое значение показателя S_{x_j} характеризует собой «лучшую» принадлежность объекта x_j к кластеру p.

Силуэтом кластера называется средняя величина показателя силуэта всех объектов кластера. Оценка для всей кластерной структуры достигается усреднением показателя по всем объектам [3]:

$$SWC = \frac{1}{N}\sum_{j=1}^{N} S_{x_j}. \tag{3}$$

Лучшее разбиение характеризуется наибольшим значением показателя *SWC*, что достигается в том случае, когда расстояния внутри кластеров a_{pj} малы, а расстояния между элементами соседних кластеров b_{p_j} велики.

Определение меры несхожести для каждого объекта с ближайшим кластером по формуле (1) на большом массиве данных требует проведения полного перебора пар объектов. В некоторых работах предлага-

ется сравнивать расстояния от каждого объекта до центров других кластеров и вычислять упрощенный индекс силуэта. Такой подход дает адекватный результат лишь для структуры с выпуклыми хорошо разделенными кластерами.

Предлагается при сравнении кластерных структур для больших массивов данных следующий подход: при оценке силуэта кластеров учитывать не все объекты, а только эталонные представители (для большей точности таковых эталонов должно быть достаточно много). Итак, пусть имеется N сложных кластеров C_k, каждый из которых состоит из n_k мини-кластеров $c_{k1}, \dots, c_{kn_k} \in C_k$. Пусть мини-кластер c_{kn_j} описан центром z_{kn_j} и количеством входящих объектов m_{kn_j}. В работах [4-5] описаны методы кластеризации, которые строят кластеры сложной формы на основе мини-кластеров.

Будем определять силуэты s_{kn_j} по формуле, аналогичной (2), где в качестве объектов при вычислении расстояний берутся центры z_{kn_j} мини-кластеров с учетом их «веса» m_{kn_j}.

Тогда формула для силуэта кластера C_k примет вид:

$$S_k = \frac{1}{N_k}\sum_{j=1}^{k} s_{kn_j} m_{kn_j}. \quad (4)$$

где $N_k = \sum_{j=1}^{k} m_{kn_j}$ – количество объектов в выбранном кластере.

Силуэт кластерной структуры будет получен как усреднение силуэтов кластеров

$$S = \frac{1}{N}\sum_{k=1}^{n_k} S_k N_k. \quad (5)$$

В качестве замечаний к вышеизложенному следует указать, что при вычислении показателя силуэта в качестве функции расстояния целесообразно брать ту же метрику, которая была использована при построении кластерной структуры.

Библиографический список

1. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999. – 270 с.

2. Миркин Б. Г. Методы кластер-анализа для поддержки принятия решений: обзор. – М. Изд. дом Национального университета «Высшая школа экономики», 2011. – 88 с.

3. Сивоголовко Е.В. Оценка качества кластеризации в задачах интеллектуального анализа данных: Дис. … канд. физ.-мат. наук. – СПб. – 2014. – 92 с.