УДК 681.518

ОЦЕНКА ЭФФЕКТИВНОСТИ РЕЗУЛЬТАТОВ КЛАСТЕРИЗАЦИИ ДАННЫХ

С. А. Петров, ассистент; С. И. Фесенко, студент,
Сумский государственный университет, Украина
sergpet@gmail.com

При управлении и прогнозировании различных экономических, социальных и других явлений часто возникает проблема, связанная с многомерностью их описания. Наиболее действенным инструментом исследования таких процессов являются методы многомерного анализа, к числу которых относится и кластерный анализ.

Основным назначением кластерного анализа является разбиение множества объектов на группы (кластеры) по определённому признаку, то есть выявление в множестве соответствующей структуры.

Для решения такой задачи используют два подхода: эвристический, который, исходя из интуитивных соображений, реализует схему разделения объектов на группы, и экстремальный, использующий для этого критерий оптимальности. Самое трудное в задаче классификации это определение меры однородности объектов. За меру однородности объектов принимают расстояние Махалонобиса или его частные случаи: эвклидово расстояние, взвешенное евклидово расстояние и Хемингово расстояние. Также важным является и определение расстояния между кластерами. В задачах с небольшим количеством объектов, где анализ структуры более важен и существует проблема определения количества кластеров, используются иерархические методы, такие, как: метод ближнего соседа (single linkage), метод дальнего соседа (complete linkage), метод средней связи (pair group average), центроидный метод (метод медианной связи) [4]. Если же число кластеров заранее задано или его можно априорно определить, то для классификации чаще всего используют параллельные кластер - процедуры, в которых реализуется идея оптимизации разбиения в соответствии с некоторым функционалом качества. К таким методам относят метод k - средних. Он является наиболее распространённым среди неиерархических методов, благодаря своей простоте, быстроте, понятности и прозрачности алгоритма.

После проведения кластеризации любым из методов важным моментом является вычисление эффективности кластеризации. Это можно сделать, вычислив вероятность ошибки [2] или используя критерий функциональной эффективности [3]. Вероятность ошибки R может быть определена, например, по формуле [2]:

$$R = 0.5\left\{1-\exp\left(-\frac{m\ln(\frac{D}{d})}{\frac{D^m}{d^m}-1}\right)+\exp\left(\frac{m\ln(\frac{D}{d})}{1-\frac{d^m}{D^m}}\right)\right\}, \qquad (2)$$

где m – мерность пространства параметров; d -наиболее вероятное значение ближайшего расстояния; D -наиболее вероятное значение межкластерного расстояния. Вероятность ошибки R можно принять за критерий качества кластеризации: чем меньше вероятность ошибки, тем выше качество кластеризации. Эффективность также может быть оценена с помощью критерия функциональной эффективности по Шеннону [3]. Эта формула имеет вид.

$$E = 1 + \frac{1}{2}\left(\frac{\alpha}{\alpha + D_2} \log_2 \frac{\alpha}{\alpha + D_2} + \frac{D_1}{D_1 + \beta} \log_2 \frac{D_1}{D_1 + \beta} + \right.$$

$$\left. + \frac{\beta}{D_1 + \beta} \log_2 \frac{\beta}{D_1 + \beta} + \frac{D_2}{\alpha + D_2} \log_2 \frac{D_2}{\alpha + D_2} \right), \qquad (3)$$

где D_1, D_2, α, β - значения точностных характеристик.

Значение R является достаточно надёжным индикатором для процедуры исключения неинформативных признаков, что поможет сделать массивы социально-экономической информации более компактными и наглядными для дальнейшей обработки и принятия решений, не накладывая никаких ограничений на вид рассматриваемых объектов, то есть позволяет рассматривать исходные данные практически произвольной природы.

1. Коваль П. Н. Использование кластеризации при анализе данных / П. Н. Коваль // УСиМ.– 2010. – №6. – С. 32-34.

2. Алехин Е. И. Многомерные статистические методы / Е. И. Алехин.– Орел: Издательский центр ГОУ ВПО ОГУ, 2007. – 37с.

3. Довбиш А. С. Основи проектування інтелектуальних систем: навчальний посібник / А. С. Довбиш.– Суми: Вид-во СумДУ, 2009. – 171 с.

4. Jain A. K., Murty M. N., Flynn P. J. Dataclustering: a review / A. K.Jain, M. N. Murty, P. J. Flynn // ACM Computeing Surveys(CSUR).–1999. – Volume 31 Issue 3. – 69 p.