Универсальные критерии кластеризации и вопросы устойчивости

Рязанов В. В., Арсеев А. С., Коточигов К. Л.
rvv@ccas.ru
Москва, Вычислительный центр РАН

Задачи кластеризации (или автоматической классификации, таксономии, самообучения, обучения без учителя, группировки) образуют важный раздел интеллектуального анализа данных. Существует несколько постановок кластерного анализа, но основная состоит в поиске разбиения выборок объектов (при заданных признаковых пространствах или матрицах близостей объектов) на классы эквивалентности (кластеры), причем эквивалентность объектов кластеров определяется каждым алгоритмом по-своему. Принципы, согласно которым объекты объединяются в один кластер, являются обычно «внутренним делом» конкретного алгоритма кластеризации. Пользователь, зная данные принципы, может в определенных пределах интерпретировать результаты каждого конкретного метода [1–4].

В отличие от задач распознавания, где существуют единые стандартные критерии оценки алгоритмов (оценка вероятности ошибки, эмпирический риск, и другие), в настоящее время не существует универсальных общепризнанных критериев качества решения задачи кластеризации. Соответственно, при отсутствии внешней суперцели, решения различных алгоритмов кластеризации сложно оценивать и сравнивать. Действительно, есть методы кластеризации, где ищется экстремум некоторого функционала качества разбиения (например, дисперсионный и родственные ему критерии, определитель матрицы внутригруппового разброса, и другие). Метод k-внутригрупповых средних находит группировки, где каждый объект находится ближе к среднему своей группировки, чем к среднему любой другой. Данные группировки и объявляются кластерами. Кластеры в методах иерархической группировки вычисляется согласно последовательному локальному объединению более мелких группировок в более крупные. В алгоритме ФОРЕЛЬ кластером объявляется группировка объектов, принадлежащая некоторому шару фиксированного радиуса, обладающего свойством: центр шара совпадает со средним принадлежащих ему объектов.

В настоящем докладе в основу общего универсального подхода предлагается положить идею устойчивости решений относительно малых изменений выборки. Здесь возможны различные критерии оценки качества кластеризаций. Назовем m-выборкой произвольную выборку из m объектов, а $(m-1)$-выборкой — выборку, полученную из m-выборки удалением

некоторого объекта. Определяется близость кластеризаций m-выборки и $(m-1)$-выборки на заданное число кластеров. По близостям кластеризаций m-выборки и всевозможных $(m-1)$-выборок оценивается качество кластеризации исходной m-выборки. Возможны и другие варианты определения качества кластеризаций.

Данные определения не связаны с сутью конкретного метода кластеризации, а отражают степень изменения кластеризаций относительно вариации выборок. Для некоторых алгоритмов кластерного анализа получены эффективные методы вычисления качества кластеризаций. Приводятся результаты анализа оценки качества кластеризаций на модельных и реальных задачах.

Литература

[1] *Айвазян С. А. и др.* Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989.

[2] *Дуда Р., Харт П.* Распознавание образов и анализ сцен. М.: Мир, 1976. — 511 с.

[3] *Загоруйко Н. Г.* Методы распознавания и их применение. М.: Сов. радио, 1972. — 206 с.

[4] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Изд-во Института математики, 1999.