Использование методов статистической верификации для поиска оптимальных кластеризаций

Пономарева Любовь Игоревна

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: lponomareva98@yandex.ru

Научный руководитель *— Сенько Олег Валентинович*

Целью методов кластерного анализа является разбиение выборок многомерных данных на группы объектов близких в смысле некоторой заданной меры сходства. Методы кластерного анализа могут использоваться как в качестве вспомогательных инструментов при решении задач прогнозирования или распознавания, так иметь самостоятельное значение, например, в задачах статистической обработки медицинских данных. В то же время оценка качества кластеризации становится предметом отдельного анализа, так как метрика сравнения не позволяет оценить полученную кластеризацию в смысле объективности существования кластерной структуры.

Мы использовали методы статистической верификации для поиска оптимальной кластерной структуры следующих данных:

- Показатели взаимосвязи 234 IgM и 285 IgG иммуноглобулинов при иммунных нарушениях в группе из 50 пациентов;
- Данные пациентов с шизофренией.

Способ оценивания статистической достоверности кластерной структуры заключался в сравнении качества кластеризации на реальной выборке с качеством кластеризации на искусственно сгенерированных выборках с тем же самыми числом объектов, признаков из фиксированного многомерного нормального распределения. Кластеризация принималась достоверной для данного числа выделенных кластеров, если значение показателя метрики на реальной выборке оказывалось больше значения 95%-ного квантиля метрики для искусственных данных. Для кластеризации наборов данных использовались алгоритмы иерархической кластеризации, DBScan, K-Means. Метрики для оценивания качества полученной кластерной структуры были заимствованы из [3], [1]. Мы сравнивали коэффициент силуэта для результатов кластеризации нашей выборки и случайной

выборки, который определяется следующим образом для отдельного объекта:

$$s = \frac{b - a}{max(a, b)},$$

где a - среднее расстояние от данного объекта до объектов из того же кластера, b - среднее расстояние от данного объекта до объектов из ближайшего кластера (отличного от того, в котором лежит сам объект). Также сравнивались значения индекса Данна:

$$D = \min_{i=1\ldots n_c; j=i+1\ldots n_c} \left\{ \frac{d\left(c_i, c_j\right)}{max_{k=1\ldots n_c}\left(diam\left(c_k\right)\right)} \right\}$$

где $d(c_i, c_j)$ - расстояние между кластерами c_i и c_j, и $diam(c_k)$ - диаметр кластера c_k. Диаметр кластера может быть найден как среднее расстояние между элементами кластера, между всеми элементами и центром кластера или как расстояние между самыми удаленными элементами. Чем больше значение индекса Данна, тем точнее результат кластеризации.

Указанный подход позволил выявить оптимальное число кластеров и найти коллективное решение кластеризации, которое является существенно более мощным по сравнению с решением, построенным по единственному критерию.

Литература

1. Кирилюк И. Л., Сенько О. В. Оценка качества кластеризации панельных данных с использованием методов Монте-Карло (на примере данных российской региональной экономики), 2016.
2. Сивоголовко Е. В. Методы оценки качества четкой кластеризации, 2011.
3. Сивоголовко Е. В. Оценка качества кластеризации в задачах интеллектуального анализа данных, 2014.
4. Halkidi M., Batistakis Y., Vazirgiannis M. On Clustering Validation Techniques, 2001.
5. Кузнецова А. В., Сенько О. В., Лобанов С. Анализ структуры данных по активности иммуноглобулинов, полученных с помощью гликановых микрочипов, 2019.
6. Сенько О. В. Перестановочный тест в методе оптимальных разбиений, Журнал вычислительной математики и математической физики, 2003.
7. Воронцов К В. Лекции по алгоритмам кластеризации и многомерного шкалирования, 2010.