... поиск оптимального числа кластеров

В большинстве случаев при использовании кластеризации исследователь сталкивается с дилеммой выбора количества выделяемых в результате кластеров (разбиений) (Everitt, 1979; Sneath & Sokal, 1973). Не относящиеся к иерархическим методы обычно принимают на вход заранее заданное число кластеров, которое, таким образом, должно быть известно до начала самой процедуры анализа. В то же время иерархические алгоритмы позволяют получать серию последовательных разбиений, начиная с числа кластеров, равному числу наблюдений n, и до единственного кластера, включающего весь датасет полностью. К настоящему времени разработаны различные подходы к определения оптимального числа кластеров (Dubes & Jain, 1979; Milligan, 1981c; Perruchet, 1983). В случае иерархических методов кластеризации наиболее простым способом является визуальная оценка дендрограммы - т.е. графа, в котором наблюдения (observation) объединены в группировки-кластеры на основе сходства. При этом визуально выявляют сравнительно плотные ветви, соответствующие объективно присутствующим кластерам в данных. В соответствии с этим метод характеризуется субъективнсотью, что необходимо учитывать при его применении (Aldenderfer and Blashfield, 1984). Коэффициент аггломерации (agglomeration coefficient), т.е. численное значение, при котором различные наблюдения(?, cases) объединены в кластер, является основой двух других сходных методик. Первая включает построение кривой зависимости данного коэффициента (ось ординат) от количества кластеров (ось абсцисс). Выраженные плато на таком графике предполагают, что при этом происходит объединение очень различных кластеров. В соответствии с этим надлежащее количество кластеров устанавливают по перегибу на графике ("правило локтя", "elbow"). (Milligan, Cooper, 1985)

Предложено разделение множества методик определения оптимального числа кластеров на две группы ( A. D. Gordon. Classification. Chapman & Hall/CRC, Boca Raton, FL, 2 edition, 1999.): глобальные и локальные. В случае глобальных методов оптимальность некоторого количества кластеров, Например, оценивается при помощи некоторого критерия, причем оптимальное число кластеров получают путем сравнения значений такого критерия для ряда последовательных значений числа кластеров. Недостатком таких методов является необходимость прежде всего, присутствует ли в данных более 1 кластера, идентичного самому набору данных. При этом подобной проблемы не возникает, если есть следующие из сторонней информации основания предполагать наличие таких группировок. Локальные методы призваны тестировать гипотезу о том, что пара кластеров должны быть объединены, и применимы только к иерархически вложенным разбиениям. При этом следует избегать однозначного толкования в связи с тем, что при этом используются многочисленные тесты (Gordon, 1999). Таким образом, количество кластеров в большинстве случаев является неизвестным параметром, который должен быть либо задан пользователем на основании имеющейся информации о наборе данных и соответствующий области знаний, либо определен некоторым дополнительным методом.

В случае использованного датасета кластеризация получена для метрики, которая была полученная напрямую как набор попарных расстояний в дорожной сети. При этом не рассматривались исходные значения, соответствующие отдельным наблюдениям (деревням). Такой тип данных ограничивает круг методик оценки оптимального числа кластеров. В частности, среди 30 доступных для расчета с помощью библиотеки среды R NbClust (Charrad M., Ghazzali N., Boiteau V., Niknafs A. (2014). "NbClust: An R Package for Determining

the Relevant Number of Clusters in a Data Set.", "Journal of Statistical Software, 61(6), 1-36.",

"URL <http://www.jstatsoft.org/v61/i06/>".) индексов при наличии матрицы расстояний, но не исходных данных применимы только 5. При использовании данной библиотеки были получены графики вида “значение критерия по оси y - число кластеров” при изменении последнего от 2 до 50. Экстремумы (для разных индексов это могут быть либо минимумы, либо максимумы) или перегибы указывают число кластеров, которые согласно данному алгоритму следует считать оптимальным. Среди 5 рассмотренных такие характерные элементы имели 3 графика , в каждом случае соответствующие 18, 33 и 42 кластерам (Рис. ...), причем для индексов frey и cindex наиболее выраженные экстремумы соответсвуют 18 кластерам. Примечательно, что данное число соответствует числу административных районов в рассмотренной поселенческой сети. Оптимальное число кластеров согласно индексу dunn (максимальное значение критерия) выделяет в качестве оптимального числа группировок 42-50, что следует признать скорее неинформативным. Нарастание графика, в частности, предполагает, что глобального максимума не было достигнуто.