М. В. Приставко, Н. Б. Осипенко
(ГГУ им. Ф. Скорины, Гомель)

МЕТОД КЛАСТЕРНОГО АНАЛИЗА *K*-СРЕДНИХ

Кластерный анализ – один из методов многомерного анализа, предназначенный для группировки (кластеризации) совокупности элементов, которые характеризуются многими факторами, и получения однородных групп (кластеров). В широко распространенной универсальной статистической системе Statistica фирмы StatSoft, Inc., созданной в начале 90-х годов для среды Windows, содержится широкий набор процедур кластерного анализа, включая иерархическое объединение, двухвходовое объединение, метод *k*-средних; алгоритмы оптимизированы для анализа очень больших проектов, например, методом *k*-средних можно анализировать 400000 наблюдений с 10 переменными. Система Statistica позволяет проводить исчерпывающий, всесторонний анализ данных, представлять результаты анализа в виде таблиц и графиков, автоматически создавать отчеты о проделанной работе. Она состоит из отдельных модулей (факторный анализ, канонический анализ, дискриминантный анализ, кластерный анализ и т.д.), каждый из которых является полноценным Windows-приложением.

Модуль кластерного анализа в пакете Statistica реализован при помощи трех методов: иерархическое объединение, двухвходовое объединение, метод *k*-средних. Наиболее распространенным является метод *k*-средних, также называемый быстрым кластерным анализом. Его распространенность объясняется простотой и быстротой использования, понятностью и прозрачностью алгоритма. Однако в системе Statistica метод *k*-средних реализован таким образом, что неопытному пользователю трудно понять, как работает алгоритм, нет наглядной интерпретации метода. Поэтому для упрощения освоения метода была написана программа на C++ в среде Borland C++ Builder, реализующая пошаговое выполнение метода для случая, когда число кластеров не более 5, а признаков – не более 3.

Рассмотрим алгоритм *k*-средних в пошаговой реализации. На первом шаге происходит первоначальное распределение объектов по кластерам. Выбирается *k* точек – центров кластеров. Выбор начальных центроидов может осуществляться следующим образом: выбор *k*-наблюдений для максимизации начального расстояния; случайный выбор *k*-наблюдений; выбор первых *k*-наблюдений. В результате каждый объект назначен определенному кластеру. На втором шаге итеративно определяются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются. Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий: кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации; максимальное число итераций фиксировано. Выбор числа кластеров является сложным вопросом. Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т.д., сравнивая полученные результаты.

На рисунке 1 отображен пример работы алгоритма *k*-средних для *k*, равного двум.

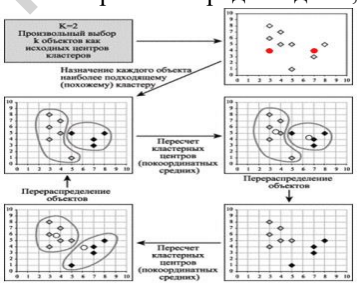

Рисунок 1 – Пример работы алгоритма *k*-средних (*k*=2)