Е. А. Троцкая, Н. Б. Осипенко
(ГГУ им. Ф. Скорины, Гомель)

КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ МЕТОДОВ *k*-СРЕДНИХ И *к*-БЛИЖАЙШИХ СОСЕДЕЙ

В наше время кластеризация является эффективным способом разделения больших объемов данных на группы для того, чтобы можно было наглядно оценить различия между ними. Кластерный анализ позволяет открыть в данных ранее неизвестные закономерности, которые практически невозможно исследовать другими способами, а также представить их в удобной для пользователя форме.

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. 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.

Наиболее распространен среди неиерархических методов алгоритм k-средних, также называемый быстрым кластерным анализом. Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, – наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Основной идеей является то, что метод k-средних относит каждое обучающее наблюдение к одному из k кластеров (где k – число радиальных элементов) таким образом, чтобы каждый кластер был представлен центроидом соответствующих наблюдений, а каждое наблюдение отстояло бы от центроида своего кластера меньше, чем от центроидов всех других кластеров. Затем координаты центроидов копируются в радиальные элементы. Цель здесь состоит в том, чтобы найти набор центров, наилучшим образом представляющий распределение обучающих наблюдений. Основной проблемой возникающей при использовании алгоритма k-средних является проблема изначального выбора количества кластеров. К сожалению, нет общих теоретических решений, чтобы найти оптимальное количество кластеров для любого заданного набора данных. Поэтому для упрощения освоения метода была написана программа на C++ в среде Borland C++, реализующая пошаговое выполнение метода для случая, когда число кластеров не более 5, а признаков – не более 3

На рисунке 1 приведена интерпретация пошаговой работы алгоритма *k*-средних.

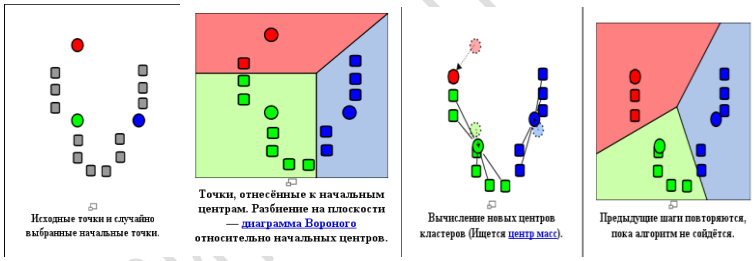

Рисунок 1 – Пример работы алгоритма *k*-средних

В методе k-ближайших соседей расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Это правило должно, в известном смысле, нанизывать объекты вместе для формирования кластеров, и результирующие кластеры имеют тенденцию быть представленными длинными «цепочками». Алгоритм способен выделить среди всех наблюдений k известных объектов (k-ближайших соседей), похожих на новый неизвестный ранее объект, и на основе этого принять решение о его принадлежности к выделенным ранее классам. Важной задачей данного алгоритма является подбор коэффициента k – количество записей, которые будут считаться похожими.