УДК 004

ОБЗОР И СРАВНЕНИЕ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ДЛЯ ЗАДАЧ АНАЛИЗА НАВИГАЦИОННЫХ ДАННЫХ

Пономарева А.В., Мейта Р.В.
Научный руководитель: Шамин А.А., доцент, к.т.н.

Национальный Исследовательский Томский политехнический университет,
634050, Россия, г. Томск, пр. Ленина, 30
E-mail: avp35@tpu.ru

This paper presents a study of the possibility of application of cluster analysis methods to the data sets from navigation receivers. The navigation data from the moving objects have a number of features, it makes the application to them of some class of clustering algorithms impossible. Such features include, in particular, the prevalence of clusters of complex shape different from circle.

Ключевые слова: *навигация, кластерный анализ, GPS, k-means, c-means, DBSCAN, Data Mining.*

Key words: *navigation, cluster analysis, GPS, k-means, c-means, DBSCAN, Data Mining.*

Введение

Методы Data Mining сегодня широко применяются для анализа больших объемов данных с целью выявления в них скрытой информации и логических связей. Одним из типов задач, решаемых методами Data Mining, являются задачи кластерного анализа. Кластерный анализ успешно применяется для исследования навигационных данных от спутниковых систем GPS/ГЛОНАСС. С точки зрения задачи кластеризации, наибольший интерес представляют значения координаты, времени, скорости и направления движения отдельной навигационной точки. Основываясь на вышеприведенной информации об исследуемых данных, приведем сравнение наиболее часто применяемых алгоритмов кластеризации.

Алгоритм k-средних

Алгоритм k-средних основан на работах Стюарта Ллойда [1] и Гуго Штейнгауза [2]. Входными данными для алгоритма являются количество кластеров и координаты центроидов (количество центроидов соответствует количеству кластеров). На каждой последующей итерации алгоритма центры масс кластеров вычисляются снова. Итеративный процесс продолжается, пока центры масс не перестанут менять свои координаты или количество итераций не достигнет установленного предела.

В контексте кластеризации навигационных данных данный алгоритм является не самым подходящим вариантом по следующим причинам:

- не всегда представляется возможным верно выделить количество кластеров, так же, как и начальные центры кластеров;
- навигационные данные содержат шум и выбросы, что оказывает большое влияние на правильность выделения кластеров данным способом;
- алгоритм плохо выделяет кластеры специфической формы, а навигационные треки являются примером подобных кластеров.

Нечеткий алгоритм c-средних

Данный алгоритм является модификацией алгоритма k-средних, он основан на вычислении вероятности отнесения элемента к тому или иному кластеру [3]. Как и в алгоритме

k-средних, на первом этапе задается количество кластеров и центры масс кластеров. Однако в данном алгоритме также задается мера схожести и матрица весов принадлежности рассматриваемого элемента к каждому из кластеров. На каждой итерации с учетом меры схожести и матрицы весов происходит пересчет центров масс кластеров и весов.

Преимуществом данного алгоритма является возможность определения степени принадлежности элемента к тому или иному кластеру. Данное свойство помогает преодолеть трудности, возникающие при применении алгоритма k-средних. Однако по-прежнему остается сложность применения алгоритма в рассматриваемой предметной области по тем же причинам, что и в методе k-средних.

Алгоритм DBSCAN

Алгоритм позволяет определить принадлежность элемента кластеру на основе плотностного распределения элементов исходного множества [4]. Начальными данными алгоритма являются расстояние ε, которое определяет радиус соседства точки, и minPts – минимальное количество точек, находящихся в соседстве с рассматриваемой точкой. На каждой итерации рассматривается точка множества. Если в ее ε-окрестности количество точек больше или равно minPts, то точка добавляется в кластер. В противном случае точка относится к шуму. Таким образом, после обхода всех точек формируются кластеры и массив точек, отнесенных к шуму.

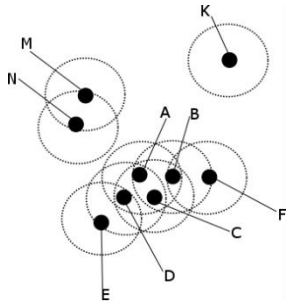

Рис. 11. Выделение кластеров методом DBSCAN

Алгоритм довольно успешно справляется с задачей кластеризации данных. При этом удается выделить шум и осуществить разбиение множества на кластеры специфической формы, что удовлетворяет требованиям предметной области. Однако данный алгоритм также имеет недостаток: принадлежность граничных точек (точек, которые одинаково достижимы из точек двух и более кластеров) к какому-либо кластеру определяется порядком обработки точек.

Выводы

На основе проведенного исследования алгоритмов кластеризации можно сделать вывод о том, что в условиях специфики навигационных данных наиболее подходящим для их анализа является DBSCAN и его производные. Основной особенностью навигационных данных является тот факт, что точки в пространстве сгруппированы в вытянутые области, плохо поддающиеся кластеризации центроидными методами. В это же время подход, реализованный в алгоритме DBSCAN, позволяет находить кластеры произвольной формы.

Список литературы

1. Lloyd S.P. Least squares quantization in pcm // IEEE Transactions on Information Theory, vol. 28 (2), pp. 129–136, March 1982.
2. Steinhaus H. On the division of material bodies into parts ("Sur la division des corps materiels en parties") // Bull. Acad. Polon. Sci., C1. III – vol IV, pp. 801–804, 1956.
3. Jantzen J. Neurofuzzy Modelling // Electronic publishing.
4. Ester M., Kriegel H.-P., Sander J., Xu X., Simoudis E., Han J., Fayyad U.M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). – 1996. – pp. 226–231.

УДК 004

РАЗМЕЩЕНИЕ СТАНЦИЙ НА ТОПОЛОГИЧЕСКОМ ПОЛЕ

Пузанов А.Л.
Научный руководитель: Погребной А.В.

Национальный Исследовательский Томский политехнический университет,
634050, Россия, г. Томск, пр. Ленина, 30
E-mail: tema.puzanov94@gmail.com

В данной статье рассматривается задача, определяющая места расположения станций на топологическом поле и соответствующую конфигурацию связей станций с терминальными точками в технической системе. На основе результатов анализа была сформирована совокупность конкретных видов станций, которые одновременно удовлетворяют требованиям по своевременному выполнению программной нагрузки и подключению терминальных точек. При решении данной задачи будем исходить из того, что на топологическом поле расположены точки. Соответственно, станции имеют векторы подключения, способные подключить эти точки. В нашем случае принимается, что все точки одного типа, а станции могут различаться только количеством подключаемых точек. При этом станции выступают в роли некоторых центров, способных подключить (обслужить) ограниченное число точек и выполнять функции сбора данных, управления, передачи сообщений и другие. Для таких приложений на первое место выдвигается задача разбиения множества объектов (терминальных точек) на заданное число подмножеств, так чтобы суммарное расстояние между объектами в подмножествах было минимальным. В данной статье будет изложен метод разбиения множеств на подмножества и возможности их применения при решении рассматриваемой задачи.

Данный подход решения задачи размещения станций основан на разрезании графа TG (топологический граф) на подграфы с минимальной суммой связей вершин в подграфе и максимальной суммой связей между вершинами из разных подграфов. Вместе с тем при разработке алгоритма оказалось востребованным применение задачи подключения точек.

Вначале мы имеем некоторое количество точек.

Далее решаем транспортную задачу (ТЗ) или задачу распределения точек по станциям. Размерность матрицы ТЗ определяется по числу станций и числу точек, в нашем случае 5*25, элементы матрицы – расстояния между станциями и точками, одна точка может быть подключена лишь к одной станции. Вначале мы находим полюса, к которым будем подключать остальные точки, для этого берем точки максимально удаленные друг от друга, далее алгоритм минимизирует суммарные расстояния от станций до подключаемых к ним точек. Под-