УДК 004.4

ИСПОЛЬЗОВАНИЕ МЕТОДА КЛАСТЕРИЗАЦИИ K-СРЕДНИХ ДЛЯ ОПТИМИЗАЦИИ ОТОБРАЖЕНИЯ ПРОСТРАНСТВЕННЫХ ДАННЫХ

Приходько А.С., Хмелевой С.В.
Донецкий национальный технический университет
кафедра автоматизированных систем управления
E-mail: prikhodko-anna@mail.ru

Рассмотрены существующие проблемы и вопросы алгоритма k-средних. Выбраны наилучшие методы их решения. Определены основные параметры и метрики, необходимые для эффективной кластеризации.

Общая постановка проблемы

В наше время, наиболее распространенной геоинформационной системой является GoogleMaps. Но и в данной системе существуют свои проблемы с отображением большого количества пространственных данных. Отображение геоинформационных данных может занять большое количество времени – даже для высокоскоростного Интернета такие операции могут стать серьезным испытанием, не говоря уже о скорости подключения у среднестатистического пользователя. Одним из решений данной проблемы является кластеризация пространственных данных.

На мой взгляд, наиболее подходящий метод кластеризации, позволяющий оптимизировать отображение пространственных данных, – метод *k*-средних. Данный метод позволит: сократить количество отображаемых пространственных объектов без потери данных, разбить на кластеры выборку пространственных данных единожды, обеспечить четкую кластеризацию.

Также существуют и проблемы в реализации метода *k*-средних. На данный момент не существует такого метода кластеризации, который был бы универсальным. Для каждой отдельной задачи необходимо четко указать требования и цели для того, чтобы правильно определить основные параметры и метрики, необходимые для эффективной кластеризации. Так, например, выбор способа определения начальных центров кластеров в дальнейшем может оказать большое влияние на результат кластеризации.

Таким образом, основные проблемы в реализации метода кластеризации *k*-средних заключаются в определении оптимальных параметров и начальных данных метода в соответствии с поставленной задачей.

Кластеризация, основные понятия и цели

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных групп должны быть отличны друг от друга. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.[4]

Кластерный анализ выполняет следующие основные задачи:

– Разработка типологии или классификации.
– Исследование полезных концептуальных схем группирования объектов.
– Порождение гипотез на основе исследования данных.
– Проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Независимо от предмета изучения, применение кластерного анализа предполагает следующие этапы:

– Отбор выборки для кластеризации.
– Определение множества переменных, по которым будут оцениваться объекты в выборке.
– Вычисление значений той или иной меры сходства между объектами.
– Применение метода кластерного анализа для создания групп сходных объектов.
– Проверка достоверности результатов кластерного решения.

Кластерный анализ предъявляет следующие требования к данным:

– показатели не должны коррелировать между собой.
– показатели должны быть безразмерными.
– распределение показателей должно быть близко к нормальному распределению.
– показатели должны отвечать требованию «устойчивости», под которой понимается отсутствие влияния на их значения случайных факторов.
– выборка должна быть однородна, не содержать «выбросов».

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Цели кластеризации:

– Понимание данных, путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»).
– Сжатие данных. Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера.
– Обнаружение новизны (англ. Noveltydetection). Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

В первом случае число кластеров стараются сделать поменьше. Во втором случае важнее обеспечить высокую степень сходства объектов внутри каждого кластера, а кластеров может быть сколько угодно. В третьем случае наибольший интерес представляют отдельные объекты, не вписывающиеся ни в один из кластеров [1].

Алгоритм кластеризации *k*-средних

Исходя из требований к системе и выбранных предпочтений, можно определиться с конкретным алгоритмом кластеризации.

Задачу кластеризации можно рассматривать как построение оптимального разбиения объектов на группы. При этом оптимальность может быть определена как требование минимизации среднеквадратической ошибки разбиения.

Алгоритмы квадратичной ошибки относятся к типу плоских алгоритмов. Наиболее распространен в этой категории алгоритм *k*-средних, также называемый быстрым кластерным анализом. Полное описание алгоритма можно найти в работе Хартигана и Вонга (HartiganandWong, 1978). В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров.

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. Выбор числа *k* может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции [2:3].

Общая идея алгоритма: заданное фиксированное число *k* кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга. Работа алгоритма делится на несколько этапов:

1. Случайно выбрать *k* точек, являющихся начальными «центрами масс» кластеров.
2. Отнести каждый объект к кластеру с ближайшим «центром масс».
3. Пересчитать «центры масс» кластеров согласно их текущему составу.
4. Если критерий остановки алгоритма не удовлетворен, вернуться к п. 2 [6].

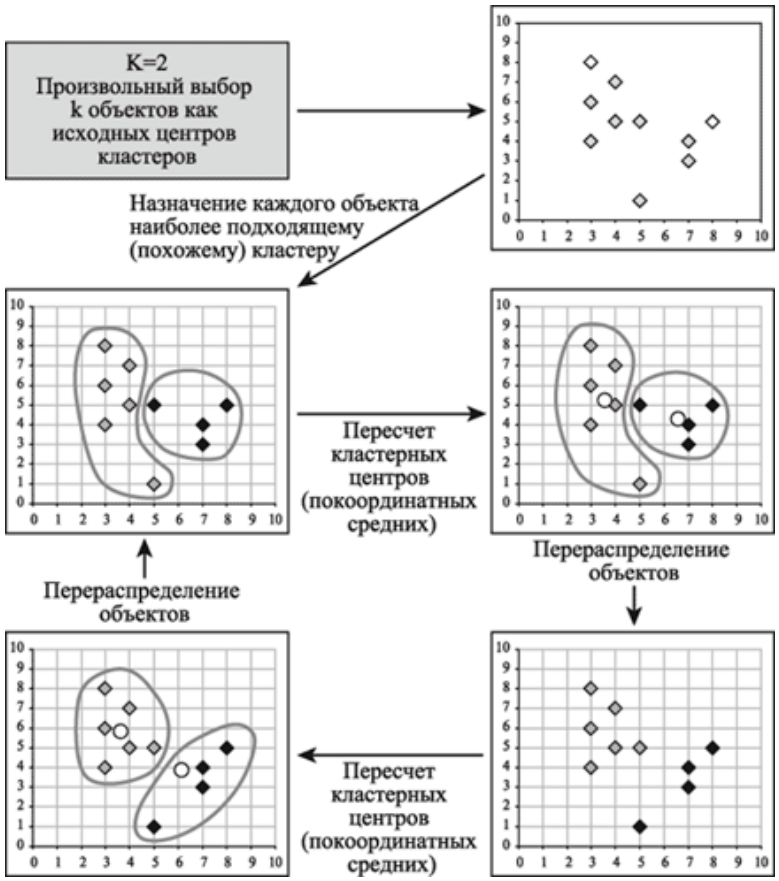

Рисунок 1. Алгоритм метода *k*-средних

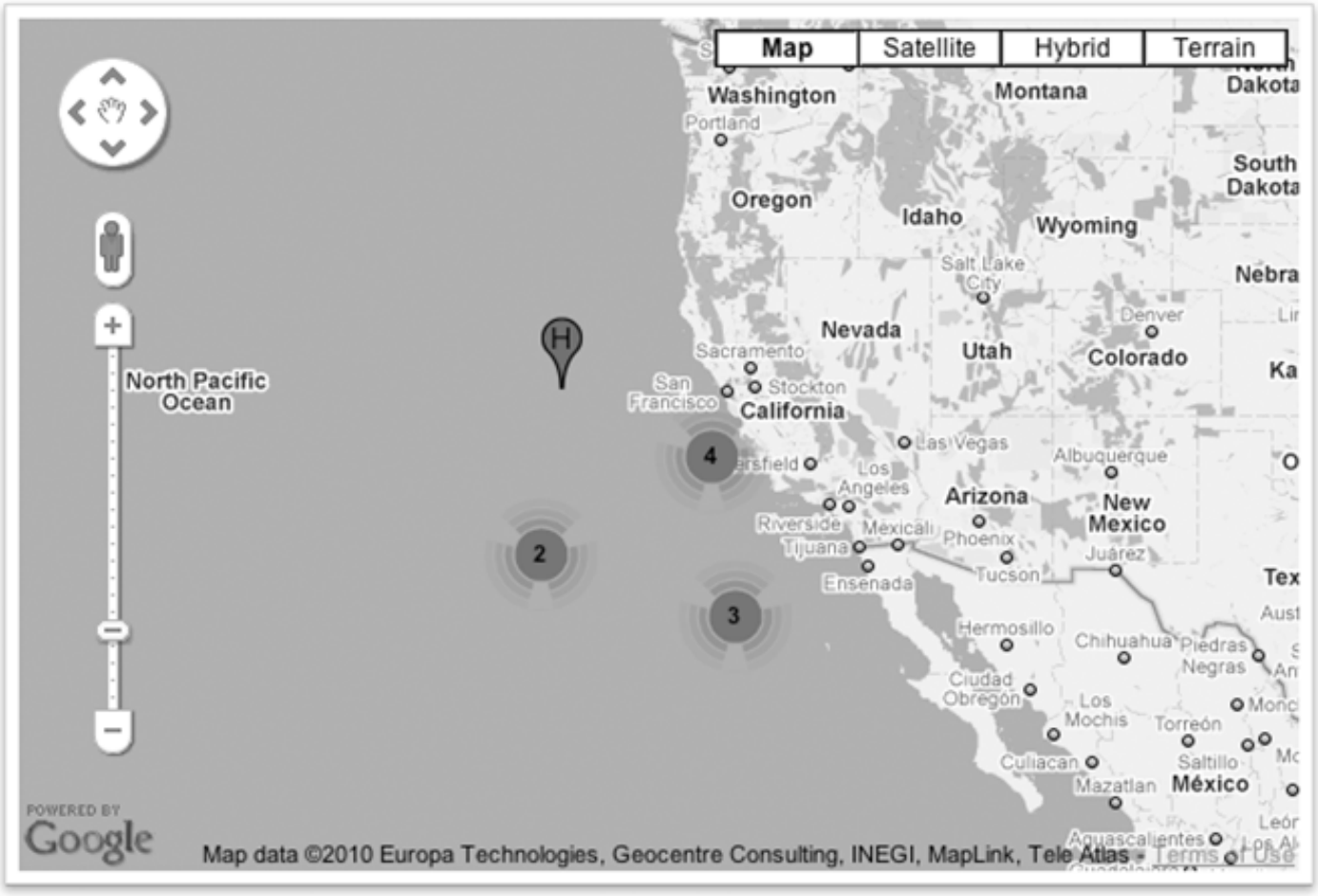

Рисунок 2. Предполагаемый результат кластеризации

Достоинства алгоритма k-средних:

1. простота использования;
2. быстрота использования;
3. понятность и прозрачность алгоритма.

Недостатки алгоритма k-средних:

1. алгоритм чувствителен к выбросам, которые могут искажать среднее;
2. необходимо заранее знать количество кластеров;
3. алгоритм очень чувствителен к выбору начальных центров кластеров.[7]

Метрики, параметры и начальные данные метода k-средних для задачи кластеризации пространственных данных

Критерий останова

В качестве критерия остановки работы алгоритма обычно выбирают минимальное изменение среднеквадратической ошибки.

$$e^2(X, L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left\| x_i^{(j)} - c_j, \right. \quad (1)$$

где c_j — «центр масс» кластера j (точка со средними значениями характеристик для данного кластера).

Так же возможно останавливать работу алгоритма, если на шаге 2 не было объектов, переместившихся из кластера в кластер [9].

Но наиболее эффективна комбинация этих двух критериев останова. Также необходимо ограничить количество итераций алгоритма.

Расстояние

Для каждой пары объектов измеряется «расстояние» между ними — степень похожести.

Наиболее распространенная функция расстояния – Евклидово расстояние. Представляет собой геометрическое расстояние в многомерном пространстве [8]:

$$\rho(x, x') = \sqrt{\sum_i^n (x_i - x_i')^2}. \quad (2)$$

Выбор метрики полностью лежит на исследователе, поскольку результаты кластеризации могут существенно отличаться при использовании разных мер.

Для данного случая наиболее подходящее «расстояние» – Евклидово, так как данная мера идеально подходит для расчета географических расстояний. Остальные «расстояния» менее подходящие, так как они решают специфические задачи, которые только усложнят кластеризацию пространственных данных.

Количество кластеров

Количество кластеров должно зависеть от начальной выборки пространственных данных. На данный момент нет четкого решения для определения оптимального количества кластеров. Эту проблему можно преодолеть проведением иерархического анализа со случайно отобранной выборкой наблюдений и, таким образом, определить оптимальное количество кластеров. Либо использовать метод определения количества кластеров, который основывается на нахождении кластеров, распределенных по некоему закону.

Начальные центры кластеров

Классический вариант подразумевает случайный выбор кластеров, что очень часто является источником погрешности. Как вариант решения, необходимо проводить исследования объекта для более точного определения центров начальных кластеров. В данном случае на начальном этапе центры кластеров также можно взять из предварительно проведённого иерархического анализа [5].

Выводы

По данным сравнительного анализа основных алгоритмов кластеризации можно сделать вывод, что наиболее подходящим алгоритм для разбиения на кластеры пространственных данных является алгоритм квадратичной ошибки. А именно, метод k-средних.

Данный метод наиболее эффективно минимизирует количество отображаемых данных без потери информации. Недостаток метода – необходимость указывать количество кластеров, никак не повлияет на результаты работы.

Остальные алгоритмы только усложнят работу, увеличат время обработки данных и нагрузку на трафик.

Основные проблемы алгоритма решены посредством детального анализа поставленных задач. Определить конкретные способы решения проблем можно только

экспериментальным путем.

После получений результатов кластерного анализа методом k-средних следует проверить правильность кластеризации (то есть оценить, насколько кластеры отличаются друг от друга). Для этого рассчитываются средние значения для каждого кластера. При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех измерений или хотя бы большей их части. Таким образом, можно найти оптимальные параметры и начальные данные для алгоритма.

Перечень источников

[1] Воронцов К.В. Алгоритмы кластеризации и многомерного шкалирования. Курс лекций. МГУ, 2007.

[2] Jain A., Murty M., Flynn P. DataClustering: A Review. // ACM ComputingSurveys. 1999. Vol. 31, no. 3.

[3] Котов А., Красильников Н. Кластеризация данных. 2006.

[4] Мандель И.Д. Кластерный анализ. — М.: Финансы и Статистика, 1988.

[5] Прикладная статистика: классификация и снижение размерности. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин — М.: Финансы и статистика, 1989.

[6] Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных — www.machinelearning.ru/

[7] Чубукова И.А. Курс лекций «DataMining», Интернет-университет информационных технологий —www.intuit.ru/department/database/

[8] Интернет энциклопедия – http://ru.wikipedia.org/wiki/Кластерный_анализ

[9] Журавлев Ю.И., Рязанов В.В., Сенько О.В. «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. ISBN 5-7036-0108-8.