СРАВНЕНИЕ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ В ЗАДАЧЕ ОБРАБОТКИ GPS ДАННЫХ

Иванин Н. С., Романов А. А.
Факультет компьютерных систем и сетей, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: {nikivnik, antekromanov}@gmail.com

В статье рассматривается возможность применения различных алгоритмов кластеризации для обработки GPS траекторий. Основной задачей при обработке GPS траекторий является выделение таких кластеров, которые относятся к одному дорожному сегменту. Для решения этой задачи рассматриваются алгоритмы single linkage clustering, k-means, DBSCAN. Для каждого алгоритма производится анализ его преимуществ и недостатков, а также проверяется эффективность на реальных GPS траекториях.

Введение

В настоящее время наблюдается значительное увеличение количества транспортных средств, оснащенных GPS датчиками, которые позволяют определять их местоположение и записывать пройденные ими траектории. Зачастую существует необходимость извлечения информации из коллекции записанных траекторий, например, дороги по которой двигалось транспортное средство или статистики скорости.

Для извлечения информации из нескольких GPS траекторий необходимо объединить эти траектории в группы. Основной проблемой при объединении траекторий является неточность измерений GPS датчика. Согласно [1] точность GPS датчиков в современных смартфонах составляет 4,9 метра. Для решения проблемы погрешности измерений GPS датчика возможно применить различные алгоритмы кластеризации.

I. Алгоритм Single Linkage Clustering

Авторы работы [2] используют алгоритм single linkage clustering(SLC), описанный в [3]. Этот метод строит иерархию кластеров в зависимости от расстояния между траекториями. На первом шаге рассчитываются расстояния между траекториями, причем каждая из траекторий образует свой собственный кластер. Затем кластеры, имеющие наименьшее расстояние объединяются в один кластер. Эта процедура повторяется до тех пор, пока не выполнятся заданные условия, такие как количество кластеров либо расстояние между элементами кластера.

К достоинствам алгоритма SLC можно отнести отсутствие необходимости задания числа кластеров, а также возможность образования кластеров не эллиптической формы. К недостаткам можно отнести чувствительность к данным с большой погрешностью измерений и выбросам, сложности при разделении кластеров большого размера, сложности при выделении кластеров разного размера.

II. Алгоритм k-means

В работе [4] предлагается использовать алгоритм k-means [5] с ограничениями. Целью этого алгоритма является нахождение k кластеров, расстояния между точками в которых минимально. В ходе своей работы этот алгоритм минимизирует сумму квадратов расстояний между точками и центром кластера.

Авторы [4] вводят дополнительные ограничения, накладываемые на точки:

- сильная связь. Точки обязаны находится в одном кластере;
- пустая связь. Точки не должны находится в одном кластере.

К достоинствам алгоритма k-means можно отнести довольно высокую скорость работы. Однако у алгоритма есть несколько недостатков. Первый из них – это необходимость задания точного числа кластеров. Также на первом этапе своей работы алгоритм выбирает центры кластеров случайным образом, что приводит к выделению разных кластеров при последовательных запусках алгоритма.

III. Алгоритм DBSCAN

В работе [6] для кластеризации GPS данных предлагается использовать алгоритм DBSCAN [7]. Этот алгоритм, в отличие от k-means не предполагает, что у кластера есть заранее известная форма. С точки зрения алгоритма DBSCAN кластер – это соединенные регионы, имеющие высокую плотность точек.

У алгоритма DBSCAN есть несколько важных преимуществ. Во-первых, для этого алгоритма не требуется начальное число кластеров. Также точки, имеющие значительные погрешности измерений и находящиеся в стороне от остальных точек, принимаются как шумовые и не добавляются в кластеры. Алгоритм DBSCAN показывает довольно хорошие результаты при поиске кластеров произвольной формы. Основным недостатком алгоритма является снижение

качества выделения кластеров из областей точек с изменяющейся плотностью.

IV. Проверка алгоритмов кластеризации на реальных данных

Для эксперимента были получены 20 GPS траекторий, записанных велосипедистами в г. Минске и проходящими через одинаковые дорожные участки. В ходе эксперимента использовались реализации алгоритмов SLC, k-means, DBSCAN из библиотеки scikit-learn[8].

При кластеризации важным является выбор метрики расстояния. Наиболее точные результаты для географических координат дает использование расстояния Евклида и Хаверсайн. По сравнению с расстоянием Хаверсайн расстояние Евклида на небольших расстояниях между точками дает маленькую погрешность, например на расстоянии 100 км между точками погрешность составляет 2 метра.

Для алгоритма SLC были выбраны следующие параметры: количество кластеров было выбрано равным 12, в качестве метрики расстояния между точками использовалось расстояние Хаверсайн, для объединения кластеров использовался критерий single. Результаты работы алгоритма SLC представлены на рисунке 1.

Рис. 1 – Результаты работы алгоритма SLC

Для алгоритма k-means были заданы следующие параметры: количество кластеров было принято равным 12, для выбора начальных кластеров использовался алгоритм k-means++, максимальное число итераций было установлено равным 300, для формирования кластеров использовался комбинированный алгоритм. Результаты кластеризации представлены на рисунке 2.

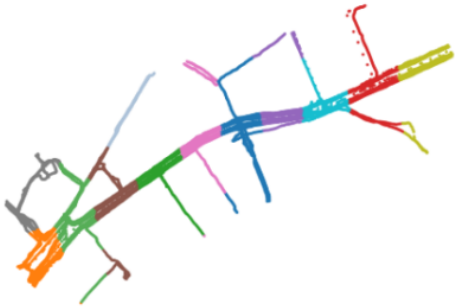

Рис. 2 – Работа алгоритма k-means

Для алгоритма DBSCAN были заданы параметры eps, min_samples, metric, algorithm. Параметр eps был выбран равным четырем средним расстояниям между точками и равнялся 7. Значение параметра min_samples был принято равным 5, что позволило улучшить кластеризацию точек, относящихся к линейным участкам дорог. Параметр algorithm определяет алгоритм поиска ближайшего соседа. Выбранный алгоритм BallTree при своей работе использует бинарное дерево, представляющие собой иерархию кластеров с точками. Точки, находящиеся в листьях дерева считаются близкими. Результаты работы DBSCAN представлены на рисунке 3.

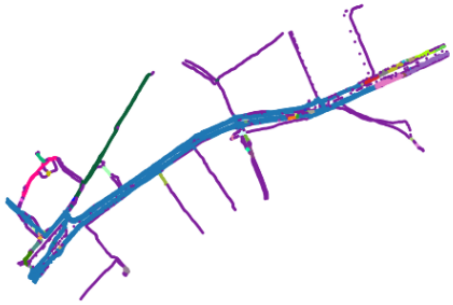

Рис. 3 – Работа алгоритма DBSCAN

Как видно из рисунков наилучшие результаты показал алгоритм DBSCAN. Он позволил выделить точки, относящиеся к одному дорожному сегменту(участок дороги без поворотов), что позволяет использовать этот алгоритм в задачах сопоставления карт и разбиения траектории на отдельные участки. Алгоритм k-means может быть использован в задаче разделения дорожной сети на равные участки.

Список литературы

1. GPS Accuracy [Электронный ресурс] / Режим доступа: https://www.gps.gov/systems/gps – Дата доступа: 09.10.2019.
2. Crowdatlas: self updating maps for cloud and personal use / Y. Wang [et al.] // Proc. of 11th inter. conf. mobile systems, applications and services. – 2013. – P. 27–40.
3. Sibson, R. Slink: An optimally efficient algorithm for the single-link cluster method / R. Sibson // The Computer Journal. – 1973. – Vol. 6, № 1. – P. 30–34.
4. Constrained k-means clustering with background knowledge / K. Wagsta [et al.] // Proc. of the 18 International Conference on Machine Learning(ICML 2001). – 2001. – P. 577-–584.
5. MacQueen, J. Some methods for classification and analysis of multivariate observations / J. MacQueen // Proc. of the 5 Symposium on Math, Statistics, and Probability. – 1967. – Vol. 1. – P. 281-–296.
6. Boeing, G. Clustering to Reduce Spatial Data Set Size / G. Boeing // SSRN Electronic Journal. 10.2139/ssrn.3145515. – 2018. – 7 p.
7. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / M. Ester [et al.] // Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining. – 1996. – P. 226–231.
8. scikit-learn: Machine Learning in Python [Электронный ресурс] / Режим доступа: https://scikit-learn.org – Дата доступа: 09.10.2019.