Задание 3
Содержание

Описание

Третье практическое задание заключается в реализации одного из алгоритмов кластеризации.
В качестве исходных данных используются данные о пространственной трехмерной карте дорог на севере Дании. Карта представлена в виде набора точек в пространстве с указанием их GPS координат и высоты над уровнем моря, а так же идентификатора ребра дорожной сети, к которому принадлежит точка (идентификатор пути из OpenStreetMap). Данный набор данных используется для тестирования алгоритмов оценки расхода топлива и выбросов CO_2 при движении по разным маршрутам. Обзор и сравнение таких алгоритмов можно найти в статье
http://cs.au.dk/~mkaul/papers/p269-guo.pdf.
Описанные там алгоритмы основываются на анализе параметров передвижения автомобилей (скорость, время простоя и прочие).
В задании же предлагается провести кластеризацию участков дорог, основываясь на гипотезе о том, что на ровных прямых участках расход топлива, а следовательно и выбросы продуктов сгорания, меньше, чем на извилистых участках со значительными перепадами высот (данная гипотеза подтверждается моделями оценки расхода топлива, представленными в статье).

Описание формата данных

Данные о трехмерной карте дорог представлены в виде набора точек в пространстве с указанием их GPS координат и высоты над уровнем моря. Формат представления – текст. Каждая строка – точка карты. Точка задается четырьмя значениями, разделенными запятой

- OpenStreetMap ID пути, которому принадлежит точка
- Долгота (восточная) в десятичных градусах
- Широта (северная) в десятичных градусах
- Высота (над уровнем моря) в метрах

Формализация задачи

Практическое задание состоит из трех частей

- Подготовка данных
- Кластеризация участков
- Визуализация
- Оценка полученной кластеризации

Подготовка данных

«Участок» - это часть дороги. В исходных данных каждый участок представлен набором входящих в него точек. Первый этап задания заключается в том, чтобы из облака точек получить набор участков, которые можно кластеризовать. Таким образом необходимо выделить из набора данных участки дорожной сети, которые затем будут кластеризованы. Стоит учесть, что по каждому участку можно проехать в двух направлениях, при этом, если есть уклон, то при езде «в горку» будет больший расход топлива, чем при движении «под горку». Поэтому некоторые участки, для которых перепад высот в конечной и начальной точках существенен, могут входить в конечный набор дважды: по разу для каждого направления езды.
Стоит так же обратить внимание на то, что в исходных данных точки задаются GPS координатами дополненными высотой, то по таким трехмерным координатам не совсем корректно считать

расстояние. Опять же, так как цель задания – кластеризация, то можно осуществлять перевод GPS координат в метрические следующим упрощенным образом:

- Переместить начало отсчета в произвольную точку выборки (присвоить ей координаты (0, 0); у остальных точек уменьшить на соответствующие значения координаты)
- Перевести координаты из градусов в метры простым умножением: широты на 111 000 (протяженность пути при перемещении на 1^{o} широты), а долготы на 60 000 (протяженность пути при перемещении на 1^{o} долготы при широте в 57^{o}).

Кластеризация участков дороги

Цель кластеризации – разбить все множество участков дороги на группы по степени их экологичности. Оценка экологичности должна основываться на гипотезе «чем ровнее путь, тем меньше расход топлива, чем больше поворотов и перепадов высот на участке пути, тем расход топлива больше». Данная гипотеза равносильна утверждению, что на ровных участках водитель может поддерживать равномерную скорость, а на извилистых должен постоянно тормозить/разгоняться, что ведет к перепадам скорости, что, как показано в приведенном выше исследовании, приводит к повышению расхода топлива.
Предлагается разработать вектор признаков, описывающий степень извилистости дороги: количество поворотов, их крутизна (величина, обратная радиусу) и протяженность, уклон дороги, количество подъемов/спусков и т.д. Вектор признаков должен содержать не менее 4х компонент.
Рассчитать для участков из набора данных их признаки и провести кластеризацию. Число кластеров предлагается подобрать самостоятельно (но не менее 3х: наиболее экономичные участки, средние и плохие с точки зрения экологичности)

Формальная постановка задачи

- На множестве участков дороги необходимо определить вектор признаков
- Необходимо выбрать и реализовать произвольный алгоритм кластеризации исходных данных на пространстве векторов-признаков.

Визуализация

После разбиения участков на кластеры необходимо визуализировать карту дорожной сети с учетом полученной кластеризации

Оценка кластеризации

Предлагается провести оценку кластеризации, используя одну или несколько из моделей расхода топлива, приведенных в статье http://cs.au.dk/~mkaul/papers/p269-guo.pdf: предлагается применить модель для оценки расхода топлива на участках и проверить разницу между полученными расходами: для участков одного кластера разница должна быть не велика, тогда как для участков разных кластеров наоборот, существенной.

Так как модели из статьи оперируют со скоростью и ускорением, предлагается построить модель, дающую некоторое приближение отсутствующих данных по участку. Например, восстанавливать скорость можно следующим образом: берем Vb (км/ч) за базовую скорость на прямом ровном участке (например, 70 км/ч). Дальше в каждой точке по паре соседних (если они не на прямой, если на прямой, то скорость не меняется) оцениваем [горизонтальный] радиус изгиба дороги. Если радиус в данной точке меньше, чем радиус в предыдущей, то значит траектория входит в поворит и скорость нужно сбавлять (сюда же можно включить моделирование ускорения): например, на delta.V = k * deltaR ^ alpha, где k и alpha - некоторые коэффициенты, значения которых можно подобрать , deltaR -

разница в радиусах. Если же радиус в точке больше радиуса в предыдущей, то дорога выходит из поворота и скорость можно набирать (на ту же deltaV). Так же можно воспользоваться таблицами вроде такой: http://www.rlib.net/racing/safe-turns/safe-turns2.htm и прикинуть сопоставить каждому повороту определенную скорость прохождения, в зависимости от радиуса. Аналогично при выходе из поворота, скорость, наоборот, набирается.

Критерии оценки

Задание оценивается по бинарной шкале: зачтено/не зачтено.
Ниже приведены подзадачи, решаемые в рамках выполнения практического задания, и получаемые за них баллы. Для зачета задания необходимо набрать 7.

- *Предобработка данных (0)*
  Обязательная часть задания.
- *Кластеризация участков (4 – …)*
  Обязательная часть задания.
  - Построение вектора признаков и реализация одного алгоритма кластеризации + отчет о применяемых метриках и методах кластеризации (4)
  - Реализация каждого последующего алгоритма кластеризации и разных метрик с описанием различий в результатах кластеризации (3)
- *Визуализация (0)*
  Обязательная часть задания. Без визуализации результатов кластеризации задание не принимается.
- *Оценка кластеризации (3)*
  Опциональная часть задания. Позволяет набрать дополнительные баллы. Включает в себя
  - разработку метода восстановления скорости и ускорения по участку дороги
  - оценку расхода топлива по одной или нескольким моделям
  - сравнение оценки для участков одного и разных кластеров

Дополнительная информация

Исходные данные так же могут быть взяты здесь:
http://archive.ics.uci.edu/ml/datasets/3D+Road+Network+%28North+Jutland%2C+Denmark%29#