纽约出租车大数据研究

傅豪 14307130013　　徐茂然 14300180099
大数据学院

摘要：纽约出租车数据中蕴含了上下客时间与经纬度、载客里程、出租车收费组成等丰富信息。借助 MapReduce 计算框架与 Spark 及其机器学习库 ML，本文侧重并解决了：（1）使用 K-Means 对乘客上车、下车地点进行聚类；（2）出租车收费分析；（3）司机驾驶行为分析；（4）基于上车地点与时间的车程及目的地预测。通过对上述问题的研究，我们试图挖掘纽约出租车大数据的潜在知识。

一、 项目背景

1. 纽约出租车概述

纽约市内运营的出租车以黄色为主，自 2013 年 8 月起，纽约出租车轿车委员会（TLC）向绿色出租车发放营业许可。绿色出租车的上客区域受到限制，在肯尼迪国际机场和拉瓜迪亚机场只能预约载客。此外，纽约的出租车还包括近年兴起的网约车，如优步（Uber）、来福车（Lyft）等。

2. 相关研究

TLC 在其官方网站公开发布出租车运营数据。前人的研究包括：绿色出租车的出现对黄色出租车载客量的影响；不同区域两种出租车的载客对比；乘客支付方式的变化；住宅区到机场的行车时间研究；天气与载客量的相关分析等。

二、 数据集

1. 数据集简介

本文的实验采用 2016 年黄色出租车的运营数据。数据集由近 1500 万条记录组成，每条记录代表一辆出租车的一次载客。记录具有 19 个属性，我们着重关注以下属性：上下车地点经纬度、上下车时间、旅程距离、总收费。收费总额由六个部分累加得到[1]。

2. 数据预处理

我们对经纬度不合理、上下车时间存在问题两类噪声数据进行了清洗。

经纬度　根据谷歌提供的数据，纽约的经纬度为（40.7，-74.0），而数据集中包含了严重偏离该点的上下客位置点，如大西洋、赤道等错误数据点。我们设定合理范围纬度 35.7~45.7，经度-79.0~-69.0，使用 Spark 对 RDD 的 filter 方法对两组经纬度域进行过滤，排除了错误地点的记录，为乘客上下车地点聚类清洗了离群点。

总时间　Python 的 pandas 工具包能够将字符型日期数据转化为日期时间对象（datetime）。乘客的下车时间与上车时间相减后，我们得到一次载客的时间花费。自然地，我们首先去除该差值小于 0 的数据记录。对剩余的数据，我们设定旅程时间上限为 720 分钟即半天，同样去除超过阈值的数据记录。

这两类噪声数据共计约 28 万条记录。

3. 数据可视化

本节通过图表展现数据集的一些初步统计结果。

1) 乘客上车点位置分布

借助 Carto 地理信息可视化工具，我们绘制了分时段的上车点分布热力图。

---

[1] 总收费 total_amount = 通行费 tolls_amount + 小费 tip_amount + MTA 税 mta_tax + 追加收费 imp_surcharge + 里程计费 fare_amount + 高峰时段/夜间额外收费 extra

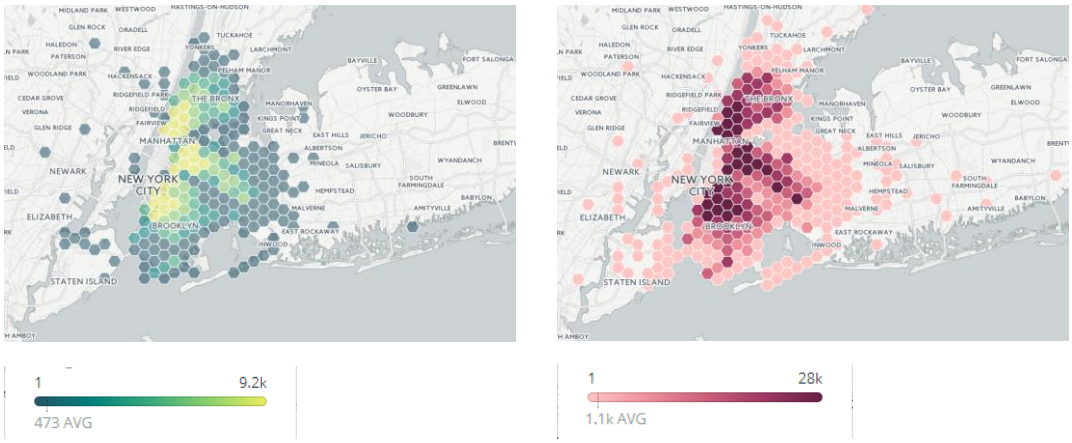

图 1：左：0 点到 12 点乘客上车点分布；右：12 点到 24 点乘客上车点分布

两个时间段上车点的疏密程度较为类似。从地理角度解读，市中心曼哈顿和布鲁克林地区无论何时均为上车点最密集的热区。右图在史坦顿岛（Staten Island，地图左下岛屿）较左图有所不同，推测可能与岛上有旅游景点有关。

2) 旅程时间

将下车时间与上车时间相减得到旅程时间。从频数直方图中可以看出，超过 70%的载客记录车程都在 20 分钟内。

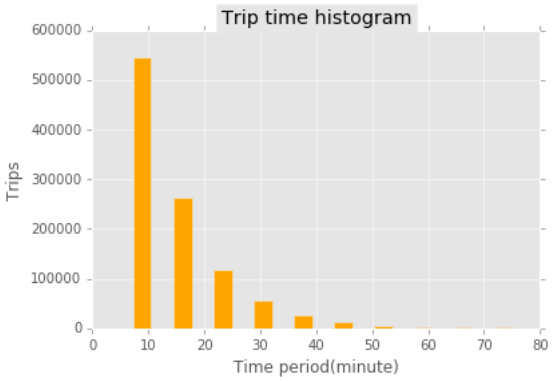

图 2：旅程用时直方图

3) 24 小时载客次数

按上车时间的不同时间段，我们统计一天 24 小时的载客次数。从图 3 不难看出每天 16 点之后纽约出租车的运营量逐步攀升并在 24 点前后达到最大值。

图 3：载客次数的小时分布直方图

三、 乘客上下车地点聚类

1. 经纬度 K-Means 聚类

以上车点为例，首先将经纬度数据载入 RDD，用 sample 对数据点做无重复的采样。调用 pyspark 的 ML 库聚类算法下的 K-Means 模型，以经度和纬度分别作为特征进行聚类。运行结果每一条数据均新增了 Label 属性，指示所分配簇的编号。

根据 Label 再次对每个簇中的点做无重复抽样，将这些样本通过可视化工具呈现，我们得到了如图 4 的效果。此例中 k=4。

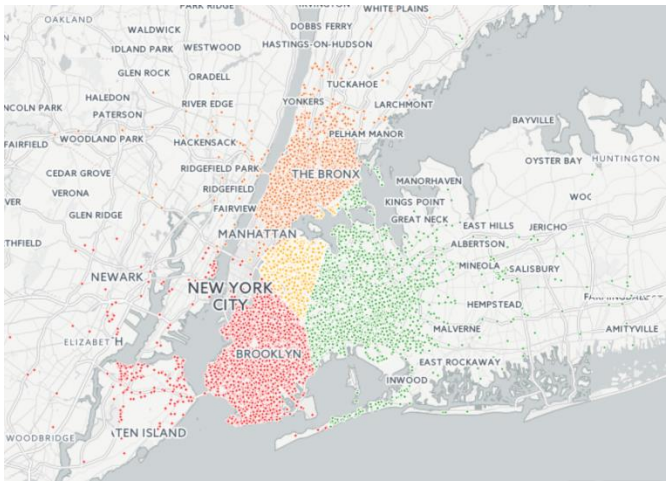

图 4：聚类结果可视化（k=4）

2. 聚类效果评价与 k 值选择

我们采用每个簇中样例个数的均匀程度来评价聚类效果。对不同的 k 分别进行 K-Means 并统计经纬度的分布情况。

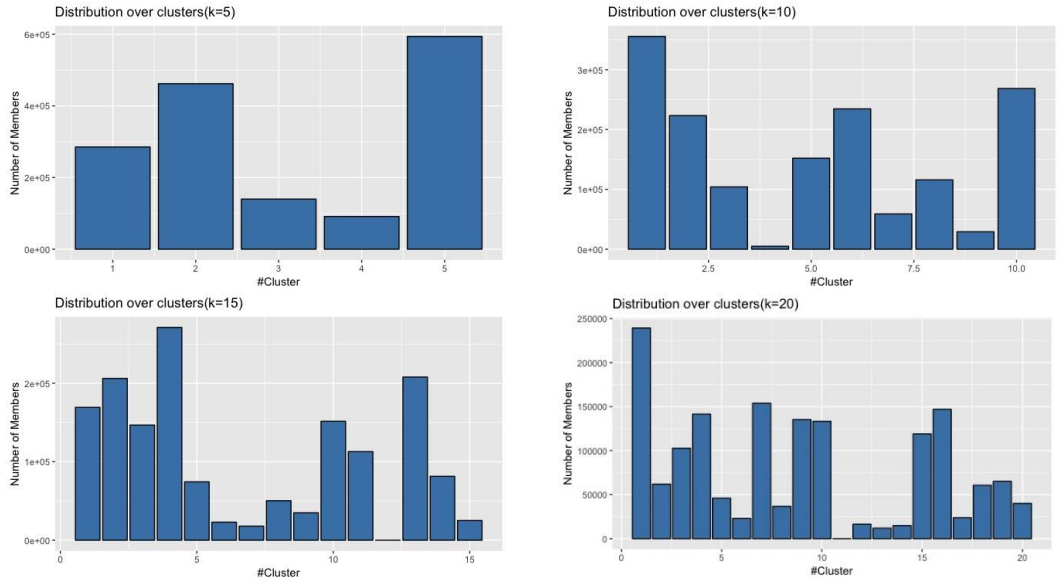

图 5：不同 k 值下各簇中的样例数目

在 k 等于 10、15 和 20 时，都存在一个簇内的样例数目明显少于其他。随着 k 的增大，样例的分布情况逐渐趋向不均衡。虽然 k 越大意味着数据点到聚类中心的总距离越小，即聚合程度越高，但需要指出的是，经纬度的值十分接近，更多的簇并不能在空间上更好地划分区域（图 6，k=10）。因此我们选取 k=5 作为 K-Means 算法的聚类中心数。

图 6：聚类结果可视化（k=10）

四、 出租车收费分析

1. 全年单日总收费

为了统计 2016 年每天所有营业出租车的总收费，我们采用 MapReduce 框架来解决这个计算问题。在 Map 阶段，我们从每一条记录的上车时间中抽取月份和日子，从而输出<(m, d), a>的键值对，其中 m、d、a 分别代表月、日以及收费额。中间过程的聚合确保了同一天的数据会进入同一个 Reduce 任务，这样在 Reduce 阶段我们就能对每一天的金额进行分别求和。

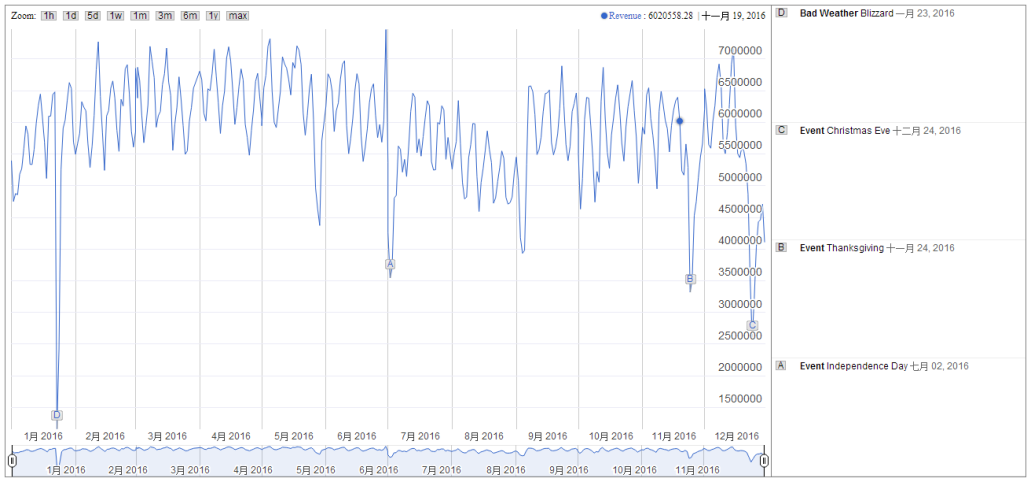

图 7：全年单日总收费（图表通过 Google Charts 制作）

在图中我们特别关注四个低谷点，如 A、B、C、D 所注。结合相关因素分析，我们得出：D 点当日纽约遭遇暴风雪，因此搭乘出租车的乘客数量骤降；而 A、B、C 分别是美国重大节日独立日、感恩节和圣诞节，故出租车营业收入下降也在情理之中。

2. 收费金额与地理位置关系研究

上车点与收费之间的关系是司机和乘客都感兴趣的问题。我们站在司机的视角，从“在哪里接客收费更高”出发对此展开研究。与上车点经纬度聚类类似，我们将总收费作为特征加入 K-Means 模型，试图引入收费对聚类的影响。另一方面，这里的聚类不同之处在于不再对 k 的取值进行比较选择，而是先取 k=20，在聚类结果中将总收费前 5 的簇提出。运用可视化工具把这些区域在地图上标注出来：

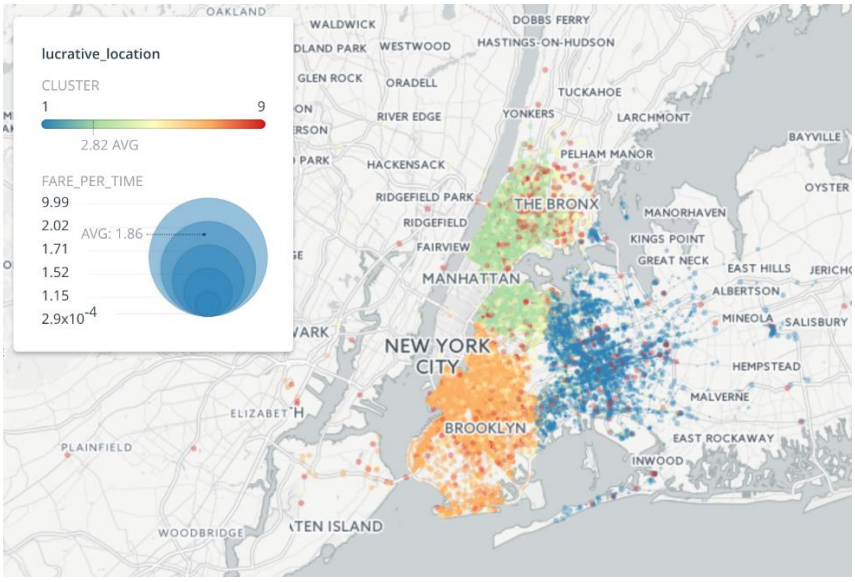

图 8：出租车高收费地点聚类[2]

图中，颜色区分不同簇，点的面积对应收费的多少（图中不十分清晰，可访问原地址）。除了较明显的三个区域外，还有无规律分布的点，意味着潜在高收费上车点的随机性。而成型的三个簇则覆盖了纽约市中心主要行政区，符合我们的直觉认识。

五、 司机驾驶行为分析

1. 发现“问题”司机

前文提到，出租车的收费由六个部分组成。然而在数据集中，有一部分记录的六项加和并不等于收费总额。对此，我们认为存在“问题”司机多收费（若加和小于总收费），而不是数据存在问题。

我们首先按条件筛选出潜在的“问题”司机，并提出问题：“问题”司机是否具有共同特征？由于数据集中可以用来显式刻画司机行为的属性较少，我们尝试用行车轨迹进行分析。对同一名司机（以 VendorID 区分）取一段连续时间内的上下客位置点，前后连接成为一个高维向量，即轨迹。统一向量长度后即可对两条轨迹计算欧氏距离等。通过空间轨迹聚类，我们查看是否有结果簇包含较多“问题”司机。实验中，聚类效果不十分显著。

2. 空载时间统计

我们把空载时间作为描述司机驾驶行为的另一个特征。空载时间由一段连续时间内一名司机的一次上客时间与前一次下客时间相减定义。我们按月份累加所有出租车空载时间，观察其月度变化。该数据可同出租车每日的营收额作比较：空载时间的峰值对应营收额的谷值，反之亦然。

[2] 详情可访问：https://moranxu23.carto.com/builder/f118cce4-5264-448e-bae2-96e31b7ad87f/embed

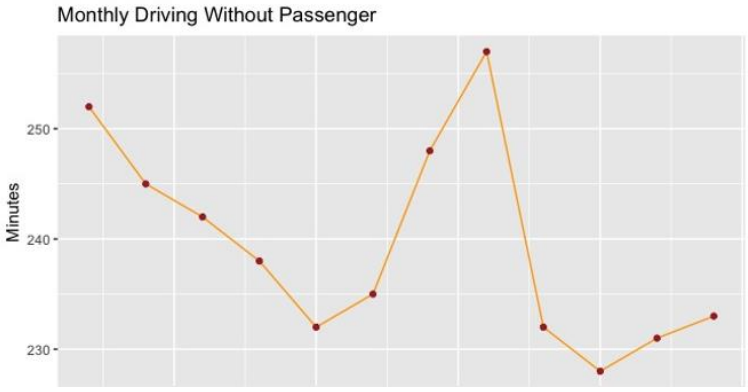

图 9：月度总空载时间

六、 车程与目的地预测

1. 旅程时间预测

旨在为乘客提供出行建议，我们试图解决以下问题：给定上车点与时间，可否对搭乘出租车前往某一目的地的花费时间进行预测？基于大量数据，我们把上车点经纬度、下车点经纬度、上车时间（小时，0~24）作为特征，旅程时间作为响应，训练回归模型。这样，输入一组新的起止点和上车时间后，模型便可给出旅程时间的预测。我们使用 pyspark 的 ML 库回归算法下的 Linear Regression、Decision Tree Regressor 等模型，在抽样的数据记录上划分训练集和测试集进行实验。在测试集上用平均绝对误差（MAE）计算准确率，线性回归模型为 5.1 分钟，决策树为 6.8 分钟。

2. 目的地预测

与旅程时间预测类似，我们研究某一特定上车点的记录最终目的地会去向何处。一个实例是华尔街精英下班的取出。将上车点经纬度在华尔街、上车时间 19 点到 20 点且是工作日的数据记录取出，首先在地图上可视化相应的下车位置。

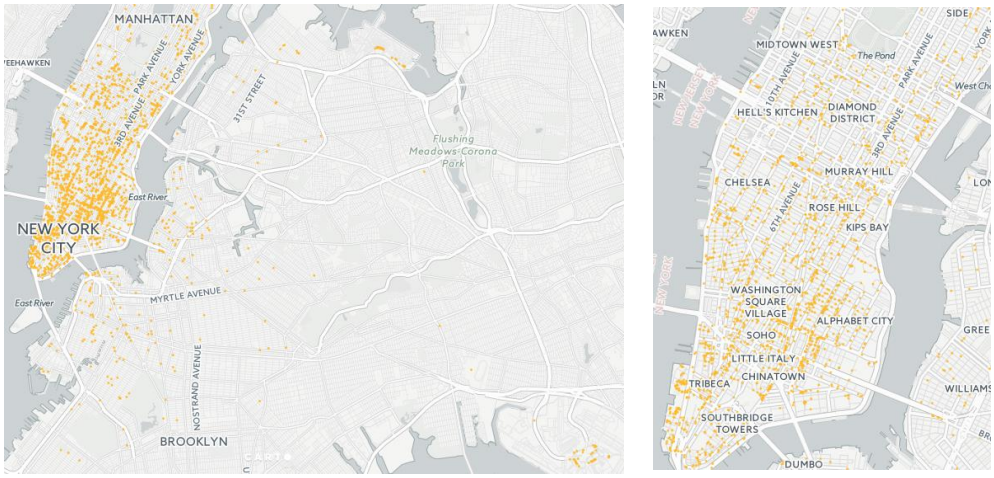

图 10：华尔街精英下班后去处（左：总览；右：局部放大）

从左图中可以看出两个主要的目的地区域。对左上区域放大后（右图）我们看到，这些目的地集中在住宅及餐饮区，包括字母城（Alphabet City）、小意大利（Little Italy）等。右下区域则为机场。此后，可对下车点经纬度再做聚类等分析，这里不加赘述。

七、 总结

本文在纽约黄色出租车数据集上做数据分析与研究，主要结果如下：

1. 运用 K-Means 对上车点经纬度聚类分析，选取了最优 k=5；
2. 运用 MapReduce 计算出租车全年单日收费总额，并绘制了 Google Charts 图表；
3. 聚类分析得到收费最高的上车点位置；
4. 追踪“问题”司机的行车轨迹；
5. 统计空载时间，结合其他因素对司机驾驶行为进行分析；
6. 运用回归模型预测旅程时间，最佳平均绝对误差为 5.1 分钟；
7. 基于出发点的目的地预测。

凭借 Spark、MapReduce 等工具，对单机难以处理的大规模数据进行有效的分析与机器学习等知识挖掘，得出了一系列有价值的信息。