数据挖掘大作业报告

开课学院： 电子信息与电气工程学院 **2020 年 4 月 20 日**

学院	电院	班级	B1903294	姓名	盛子豪	学号	119032910114
作业名称		聚类算法的软件实现				指导教授	何星

1 数据集介绍及预处理

1.1 数据集介绍

本次作业所使用的数据集是下一代模拟（NGSIM）中的 Lankershim Boulevard 数据集。2005 年 6 月 16 日，NGSIM 项目的研究人员在加利福尼亚州洛杉矶环球城附近的 Lankershim 大道上收集了详细的车辆轨迹数据。研究区域由三至四条南北走向的车道组成，覆盖三个信号交叉口，长度约为 500 米（1600 英尺）。这些数据是将视频中的车辆轨迹数据识别出来，分辨率为十分之一秒。

车辆轨迹文档中一共有 25 个属性，分别为 VehicleID, FrameID, TotalFrames, GlobalTime, LocalX, LocalY, GlobalX, GlobalY, vLength, vWidth, vClass, vVel, vAcc, LaneID, Oone, DZone, IntID, SectionID, Direction, Movement, Preceding, Following, SpaceHeadway, TimeHeadway, Location。

1.2 数据预处理

此次作业使用的属性为 VehicleID, GlobalTime 和 LocalY。VehicleID 表示车辆的序号。GlobalTime 表示自 1970 年 1 月 1 日起已用时间，单位是毫秒。为了方便表示，我把它转换为以实验开始的时间为起点，单位为秒的数据。LocalY 表示车辆前中心沿 Lankershim 大道中线的纵向坐标，以英尺为单位，起点位于研究区南部边界。

图1为选取了三辆车前 500 英尺的轨迹的原始数据，可以看到存在一些噪声点，这些噪声点是在识别轨迹的过程中出现了一些时间或者空间上的偏差所导致的，可以通过编程查找一小段区间内显著不同的数据点很容易地清洗掉。然后由于噪声的产生而导致缺失的数据点可以通过均值填充。经过数据清洗和填充预处理后的轨迹展示在图2中。

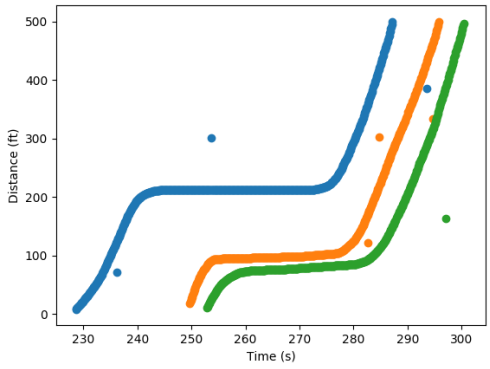

图 1: 原始轨迹

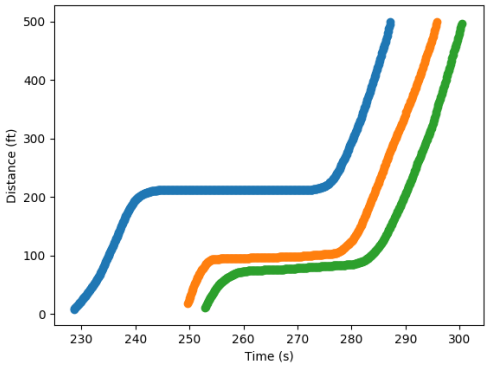

图 2: 预处理后的轨迹

2 算法描述

2.1 DBSCAN

DBSCAN 是 Density-Based Spatial Clustering of Applications with Noise 的缩写，是一种简单有效的基于密度的聚类算法。该算法通过不断生长足够高密度区域来进行聚类; 它能从含有噪声的空间数据库中发现任意形状的聚类。这种算法假定类别可以通过样本分布的紧密程度来决定。同一类别的样本，它们是比较紧密的，也就是说，对于属于一个类别的样本，在这个样本的不远处很大可能有同一类别的样本。

应用 DBSCAN 算法时，我们需要估计数据集中特定点的密度，特定点的密度是通过计算该点在指定半径下数据点个数 (包括特定点)，这种计算得到的某个点的密度也被称为局部密度。计算数据集中每个点的密度时，我们需要把每个点归为以下三类:

1. 如果点的局部密度大于某个阈值，称这个点为核心点。
2. 如果点的局部密度小于某个阈值，但是它落在核心点的邻域内，称这个点为边界点。
3. 如果点不属于核心点也不属于边界点，称点为噪声点。

算法流程如下:

- 如果所有点已经处理，停止
- 对于以前没有处理的特定点，检查它是否是核心点
- 如果不是核心点

  将其标记为噪声点
- 如果是核心点，将其标记并
  - 使用这一点形成一个新的聚类 C_{new}，并包括集群内的邻域内或边界上的所有点
  - 将所有这些在邻域内的点插入队列中
  - 当队列不为空
    - 从队列中删除一个点
    - 如果这个点不是核心点，则将其标记为边界点
    - 如果这个点是核心点，则标记它并检查其邻居中以前没有分配给类的每个点
    - 对于每一未分配的相邻点，将该点分配给当前类 C_{new}，将该点插入队列中

2.2 谱聚类

谱聚类是一种基于图论的聚类方法，通过对样本数据的拉普拉斯矩阵的特征向量进行聚类，从而达到对样本数据聚类的母的。谱聚类可以理解为将高维空间的数据映射到低维，然后在低维空间用其它聚类算法（如 KMeans）进行聚类。。对于给定包含 N 个样本的数据集 $\{x_1, x_2, \cdots, x_N\}$，谱聚类的目标是将样本分配到 k 个类中。算法流程如下:

- 构建样本间的邻接矩阵
- 通过邻接矩阵，利用 kNN 方法获得相似矩阵
- 通过相似度矩阵获得度矩阵
- 通过度矩阵和相似度矩阵获得拉普拉斯矩阵

- 求解拉普拉斯矩阵的特征向量
- 根据特征向量对特征向量构成的矩阵进行行聚类，实现对数据集的 k 聚类。

谱聚类最后一步可使用 k-means 算法。算法流程如下：

- 选择一个含有随机选择样本的 k 个簇的初始划分，计算这些簇的质心。
- 根据欧氏距离把剩余的每个样本分配到距离它最近的簇质心的一个划分。
- 计算被分配到每个簇的样本的均值向量，作为新的簇的质心。
- 重复第二和第三个步骤，直到 k 个簇的质心不再发生变化或者准则函数收敛。

3 结果及参数分析

3.1 结果

图3展示了使用 DBSCAN 算法进行聚类后的结果，可以看出此算法把车辆行驶的轨迹聚为一类，把车辆停止的轨迹聚为一类。DBSCAN 算法一共有两个参数：半径 r 和最小局部密度阈值 d。图4展示了使用谱聚类算法后的结果，可以看出谱聚类把每个车辆的轨迹聚为一类。谱聚类有两个参数：簇个数 n 和 kNN 的参数 k。

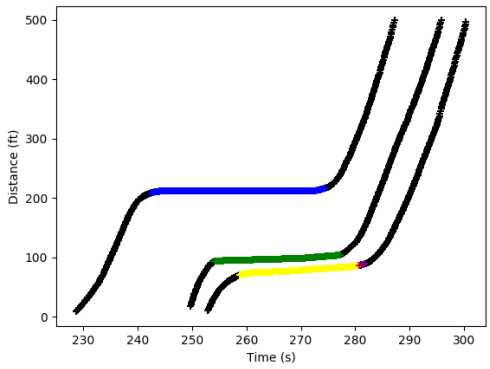

图 3: DBSCAN: r=0.3, d=2

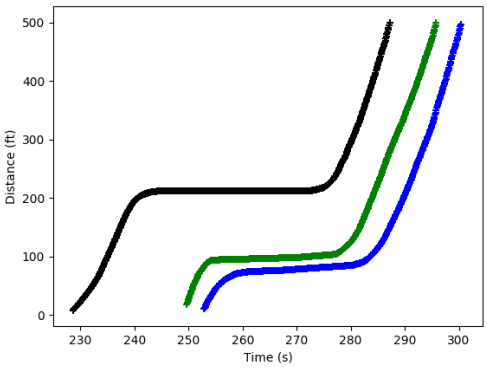

图 4: 谱聚类：n=3, k=5

3.2 参数分析

当改变 DBSCAN 的参数时，得到下图的结果。当半径过大时，会把相离较近的轨迹聚为一类，相反，半径过小时，即使是一辆车的轨迹也会被聚为好几个类。当最小局部密度阈值设的较大时，会导致很少有数据点会超过它，从而都被识别为边界点或噪声点。当最小局部密度阈值设的较小时，会产生较多的核心点，使聚出的类变多。

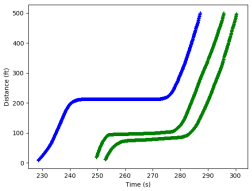

图 5: r=5, d=2

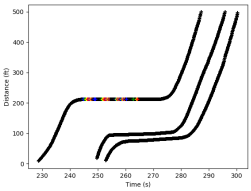

图 6: r=0.1, d=2

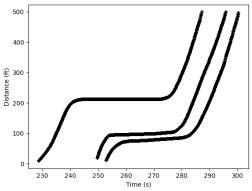

图 7: r=0.3, d=10

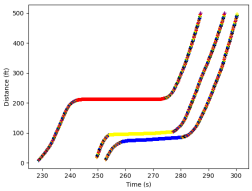

图 8: r=0.3, d=1

当改变谱聚类算法的参数时，得到下图的结果。当参数簇的个数变化，聚出来的类的个数也相应变化。当 kNN 的参数 k 变化时，会影响相似矩阵的生成，从而导致聚类的结果发生变化。

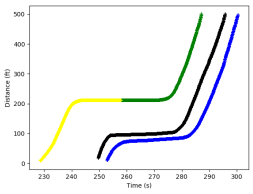

图 9: n=4, k=5

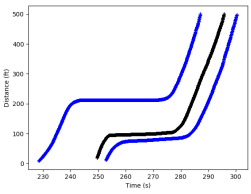

图 10: n=2, k=5

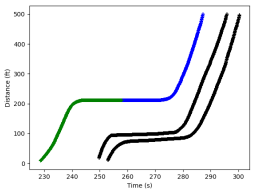

图 11: n=3, k=10

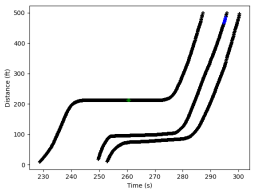

图 12: n=3, k=1

4 心得体会

通过此次大作业，我更加深入地理解了聚类算法的设计流程，并且通过编程得到进一步巩固。在最终得到比较好的聚类结果的过程之中，我感受到了调参的重要性。在完成了两种聚类算法的程序后，我与自己最近的研究方向相结合，最终实现了对车辆轨迹的聚类。一种算法可以把车辆行驶和停止的轨迹当作不同的簇，一种算法可以把每个车辆的轨迹作为簇识别出来。这次大作业让我开始用数据挖掘的角度审视我的研究方向，也对自己的研究有了一些新的启发。