轨迹相似度的聚类较简单的空间聚类更为复杂,常用的方法包括:

一、常用的特征提取

1.过滤噪声或数据补全

1）滤波

2）速度加速度

3）线性插值（常出现在保持数据维度等长、缺失值补全等场景）

2.特征选择

1）粗糙集

2）频繁项集

二、聚类方式

1.特征提取后直接进行聚类

1）特征提取 => 聚类

2.通过一些对特征提取进一步的转换和延伸得到聚类

1）特征提取 => 粗糙集 + 聚类

2）特征提取 => 灰色 + 聚类

3）特征提取 => HMM + 聚类

三、将相似度判断规则加入进来再进行聚类

1）特征提取 => 重新编写距离计算函数 => 利用新的聚类计算函数聚类

2）特征提取 => 通过一些相似度规则构造距离相似度矩阵 => 谱聚类、PIC或AP聚类

**注：这个方法还可以解决一些特征不定长的问题**

四、不同聚类算法对轨迹聚类表现的评价

——可以参考《聚类算法观点》

1. K-均值聚类算法

1）K-means是最经典的聚类算法，其思想简单且效率较高，计算复杂度仅为O(n)

2）该算法对“噪声”和孤立点数据敏感，聚类中心会因此而出现偏离。

3）要求球形分布

2. K-medoids较好地解决K-均值算法中的“噪声”敏感问题，不过由于其采用新的质点计算规则，也使得算法的时间复杂度为O(n2) 。要求球形分布。

3.DBScan

1）克服了对分布的要求

2）对阈值设定比较敏感

4.Mean-Shift

1）克服了对分布的要求

2）运算量较大

5. Density-Peak

1）克服了对分布的要求

2）算法复杂度不高

3）方法不要求特征定长

五、聚类后进行异常检测

1.常用模式

1）聚类 => 检测到中心点的距离 => 标记为异常

2）时间分箱（用于时序数据） + 聚类 => 检测到中心点的距离 => 标记为准异常 => 将准异常投票决定最异常的若干个

2.监控异常类别

1）聚类 => 人工识别异常类别 => 监控异常类别

**注：应用场景如监控车站人群中的行李丢失人群这一大类（如果我们给聚类出的某类这样定义的话）**