出租车热点区域划分判别分析

杨文静，ZY1321222

(北京航空航天大学 软件学院 100191)

**摘要：**北京市出租车在空载时会趋向附近的热点区域，如何划分区域是其中的一个问题。本文采用判别分析分析其中热点区域划分的合理性。本文首先采用聚类算法K-means将出租车载客事件发生的经纬度聚类为10簇，然后对其聚类结果进行判别分析。实验结果表明，采用K-means方法聚类产生的划分效果良好。

**关键字：**判别分析, K-means, 载客事件

# 引言

现有的车载网络通信中，路由协议往往侧重于个别指标，缺乏对多项指标的综合考虑，往往在个别指标上性能优越，但无法优化多项指标，网络整体性能难以获得较大的提升。现有的车载网络中，考虑车辆的行为特征有助于路由协议的设计，提高路由效率。如何找到热点区域并划分区域进行路由活动是其中研究的一个难点。

北京市出租车在空载时会趋向附近的热点区域，如何划分区域是其中的一个问题。本文采用判别分析分析其中热点区域划分的合理性。本文首先采用聚类算法K-means[1]将出租车载客事件发生的经纬度聚类为10簇，然后对其聚类结果进行判别分析[2]。

本文接下来主要从以下三节来论述，第二章对载客事件的地点属性进行聚类分析。第三章对聚类分析的结果进行判别分析。最后总结全文。

# 载客事件地点聚类分析

聚类分析是根据研究对象的特征对研究对象进行分类的多元统计分析技术的总称，它直接比较各事物之间的性质，将性质相近的归为一类，将性质差别较大的归入不同的类。本文采用的是K-均值聚类分析，它又称快速聚类法，它是非系统聚类法中最常用的聚类法，它的优点是占有内存少、计算量小、处理速度快，特别适合大样本的聚类分析[1]。

K-means算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。符合我们区域划分要求。其算法过程如下：

1）从N组数据中随机选取K组数据作为质心。

2）对剩余的每个文档测量其到每个质心的距离，并把它归到最近的质心的类。

3）重新计算已经得到的各个类的质心。

4）迭代2～3步直至新的质心与原质心相等或小于指定阈值，算法结束。

为了利于分析，对K值得选取暂时选定为10，采用欧几里德距离作为分簇标准。

## K-Means配置

K-Means设置簇数目为10，设置最大迭代数为100，输出簇的中心值以及每个数据所在的簇。其配置界面如图 1所示。

|  |  |
| --- | --- |
|  |  |
|  |

图 1 K-Means配置

## 聚类结果及其分析

聚类分析效果如图 2所示。聚类的初始值由系统随机指定。初始的分簇中心和最终的分簇中心如表 1、表 2所示。

表 1 初始中心值

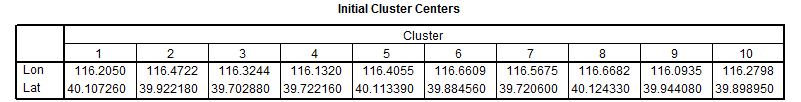


表 2最终中心值

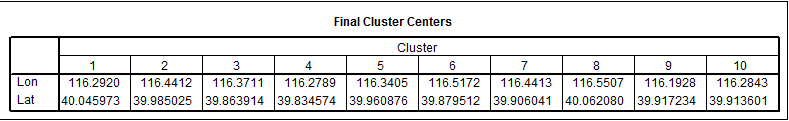


表 3 簇内节点数目

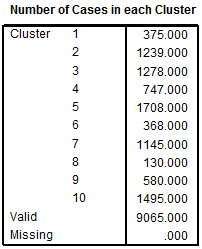




图 2 聚类效果图

由以上分簇结果可见，划分区域为不相互交叉的，簇内具有较好的聚合程度。每组数据距分簇中心的距离的数量级为0.01。. 聚类后扩展数据如图 3所示。QCL\_1代表簇的case编号，QCL\_2代表距离本数据所在簇的中心的距离，单位为度。

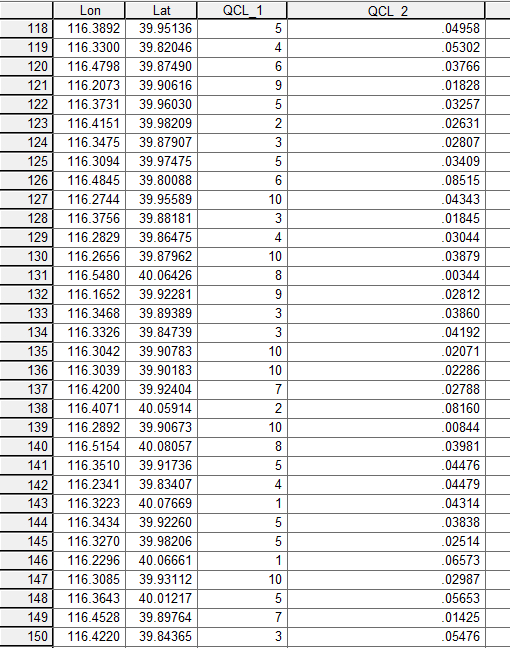


图 3 聚类后扩展数据示例

# 载客事件区域划分判别分析

## 判别分析参数设置

具体的判别分析参数配置如图 4所示。

|  |  |
| --- | --- |
|  |  |
|  |
|  | |

图 4 判别分析参数配置

在分组变量“Grouping Variable”栏中选中QCL\_1,即簇的case值。该值应为离散值。并设置范围为0到10. 在分簇过程中为1到10. 多设置一个0标签是为了发现新的分簇。自变量的位置的经纬度。采用“Enter independents together”选项，认为所有的自变量都能对观测量提供丰富的信息。描述的统计量为均值，输出Fisher判别函数系数和未标准化的函数系数。最后设置先验概率，采用的协方差举证以及选择生成和输出的分类结果。

## 判别分析结果和分析

通过以上配置，得到的分析结果如下。

图 5为具体的判别方程全图。



图 5 判别方程全图

表 4给出了数据分组后的各组均值、标准差。以及未加权的观察数目和已经加权的观察数目。

表 4 数据分簇统计量

****

表 5 典型特征函数的特征值

表 6 lambda值得卡方转化和卡方检验

****

表 7 Fisher判别函数系数表

****

表 7 表示Fisher函数的系数表。由此可知每组的判别函数,例如第一组的判别函数为：F1 = 143414.3Lon+63544.298Lat-9611318.

表 8 判别回带统计表

****

表8表明，对于第一类，使用判别回带分类，共有数据375条，正确分类的371条，其中应分为第二类的2条，第五类的两条，准确率达到98.9. 类似的其他分类均具有大于96%的准确率，最高准确率可达100%，即没有错分。由此可以说明判别函数可信。

# 结论

本文采用判别分析分析其中K-Means方法划分热点区域的合理性。本文首先采用聚类算法K-means将出租车载客事件发生的经纬度聚类为10簇，然后对其聚类结果进行判别分析。实验结果表明，采用K-means方法聚类产生的划分效果良好。采用K-Means方法对区域进行划分具有较高的划分准确性。

**参考文献**

[1]Dan Pelleg,Andrew W. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters.[M]，2000: 727-734

[2]孙海燕,周梦,李卫国,冯伟. 应用数理统计[M]. 北京:北京航空航天大学数学与系统科学学院,2012年9月.

# 附录：分组的判别函数图像





















# 附录：其他判别分析结果

****

****