驾驶员行为数据的聚类分析

一、 数据预处理

该数据记录了 63 为驾驶员在平均经历的 5 次高风险情况中，最危险的一次情况的各参数。数据集规模较小，各数据的分布情况如图 2。由于每组数据都代表一个驾驶员的行为习惯，因此也保留一些离群数据。

二、 特征选择

该问题选取的主要特征有：本车雷达的与前车雷达的纵向相对距离（$range\_x$）、纵向加速度（$long\_accel$）及驾驶员遇到前车刹车后的平均反应时间（$reaction\_time$）。其在三维图中的分布情况如下：（绘图网址 https://cnsknowall.com/）

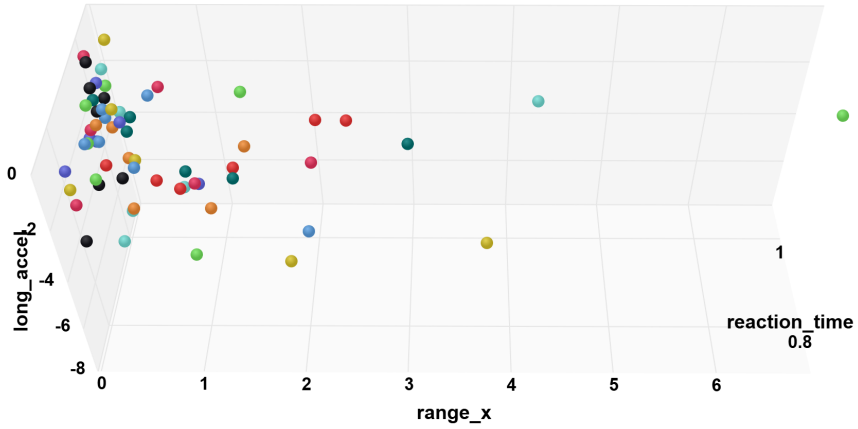

图 **1**　样本点三维分布图

假设本车雷达的与前车雷达的纵向相对距离即可用于表示纵向上相邻两车的车间距。从该图可知：数据点大部分集中在较低的 $range\_x$ 和较小 $long\_accel$ 区域，在 $reaction\_time$ 方向分布相对均匀。可知大多数驾驶员习惯保持较小的车间距，并且对风险情况表现更淡定。

三、 相似性度量

3.1 描述性统计分析

对三类特征数据进行描述性统计分析，如下表所示：

表 **1**　驾驶员行为数据描述统计表

	range_x(m)	**long_accel(m/s^2)**	**reaction_time(s)**
均值	0.870603	-1.743255	0.862326
标准差	1.177007	1.212131	0.080565
最大值	6.848001	-0.00600	1.079167
最小值	0.032000	-6.873600	0.708215
上四分位数	1.040000	-0.879000	0.907218
中位数	0.416000	-1.461000	0.872814
下四分位数	0.176000	-2.400000	0.793736

分别对三类特征数据进行可视化如下：（绘图网址 https://cnsknowall.com/）

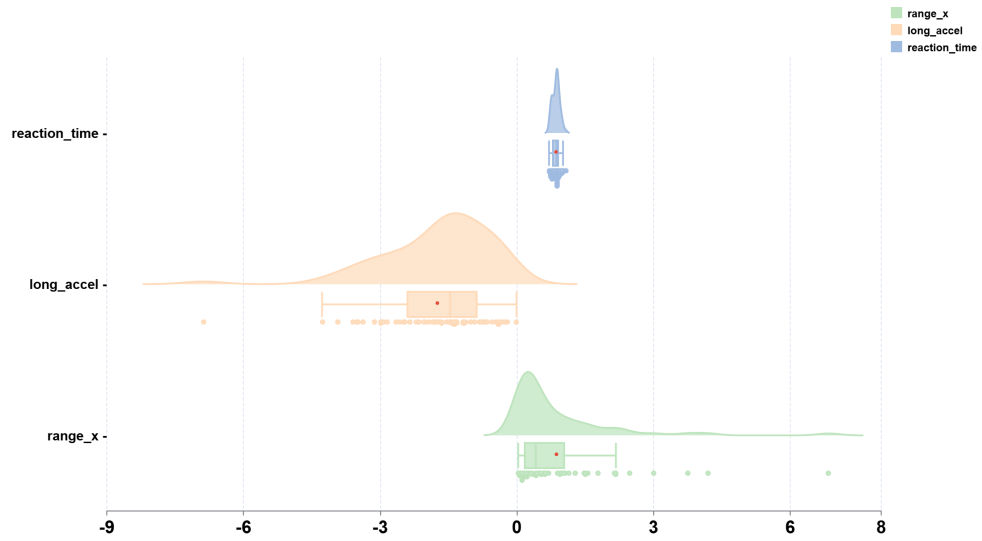

图 **2**　驾驶员避险行为数据分布情况

图表显示了三类特征的统计特征和分布情况。首先，从 $range\_x$ 来看，其均值为 0.87 米，最小

值仅为 0.032 米，而最大值达到 6.84 米，分布具有较大的右偏性。再结合图中的分布图，可以看到大部分数据集中在较小范围内，但也存在一些“异常点”，表明少数驾驶员保持了较大的车距。该类数据的中位数为 0.416 米，低于均值，进一步说明了数据呈现右偏。上述结果表明驾驶员整体较倾向于与前车保持较小车距，但少数驾驶员可能极为谨慎而保持较大距离。

其次，对于 $long\_accel$ 来说，其均值为 $-1.74m/s^2$，且数据分布较为宽广；最小值为 $-6.87m/s^2$，说明是急刹车的情况，而最大值为 $-0.006m/s^2$，表明非常轻微的刹车。从图中的数据分布情况看，该变量呈现出单峰分布，且大多数值集中在 -3 到 $0m/s^2$ 之间。数据的分布反映出多数驾驶员在紧急情况下采取中等程度的刹车，极少数情况下会出现非常急剧的刹车反应。

最后，对于 $reaction\_time$ 而言，其的均值为 $0.86s$，分布相对集中；中位数（$0.87s$）接近于均值，而标准差仅为 $0.08s$。图表显示 $reaction\_time$ 的分布较为对称，且无明显异常点。这表明驾驶员的反应时间在中较为一致，反应灵敏度整体较高。

综合来看，这三组数据的分析表明：

- 仅存在少数非常谨慎的驾驶员。但大多数驾驶员保持较短车距，这可能增加遇见紧急情况的行驶风险。
- 刹车加速度的分布显示驾驶员在应激情况下的反应强度差异较大，少数驾驶员倾向于急刹车。
- 反应时间整体表现较好，说明驾驶员在面对突发事件时的响应能力较为稳定。

从安全性的角度，建议关注那些 $range\_x$ 小且 $long\_accel$ 绝对值大的驾驶员，因为这类驾驶员的驾驶风格可能更具风险。

3.2 相似相异性分析

整体而言，数据可以分为数值数据，标称数据，二元数据等。每一类数据都有各自的相似相异性度量方法。由于该问题下的三类特征数据全部属于数值型数据，因此可以用距离来度量相似相异性。一般采用闵可夫斯基距离，其公式如下：

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h} \tag{1}$$

其中 $x_{i1}, x_{i2}, \cdots x_{ip}$，$x_{j1}, x_{j2}, \cdots x_{jp}$ 分别为两个对象的 p 维的数据，通常情况下采用 $L2$ 范数，其公式如下：

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots (x_{ip} - x_{jp})^2} \tag{2}$$

本文用 L2 范数来衡量驾驶员之间的相似相异性，对 63 位驾驶员的紧急避险数据进行 L2 范数计算，得到下图的 L2 范数矩阵：（代码见 $3.1.py$）

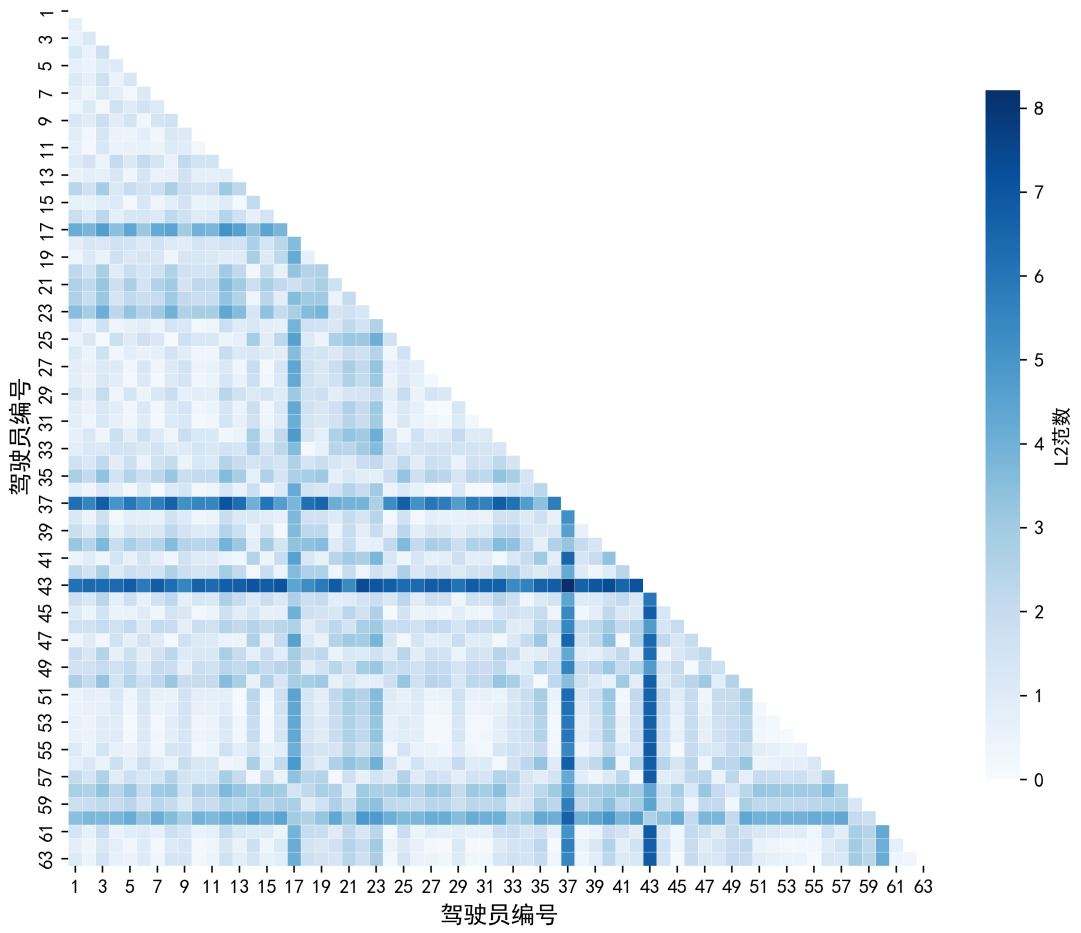

图 **3**　驾驶员紧急避险数据 **L2** 范数

图中的每个颜色代表了两个驾驶员在三个特征（前后车间距、纵向加速度、反应时间）上的相似相异性。L2 范数越大，颜色越深，表示这两个驾驶员在这三个特征上的相异性越大，意味着他们的驾驶风格、应急反应等方面的行为差异也更明显。

从图中可以观察到，一些驾驶员之间的 L2 范数较小，说明这些驾驶员在紧急避险条件下表现得比较相似，拥有相似的驾驶习惯或应急反应策略。而相反，那些 L2 范数较大的驾驶员，他们的驾驶行为相异性则较为显著，在面对紧急情况时展现出截然不同的反应模式。尤其注意到，37 和 43 号驾驶员与其他所有驾驶员有近乎完全相异的驾驶习惯和紧急避险行为。

四、 聚类分析及结果评价

对本题采用 AGNES 层次聚类，其基本步骤如下：

1. 每个样本都被认为是一个独立的簇。
2. 计算所有簇之间的距离，使用欧氏距离。
3. 选择距离最小的两个簇进行合并，形成一个新的簇。

4. 合并后，采用中心链接更新簇之间的距离。

5. 重复步骤 2、3、4。

设置不同的簇类数目，并计算每一个簇类数目下聚类的轮廓系数，得到如下折线图：（本节所有图代码见 4.1.*py*）

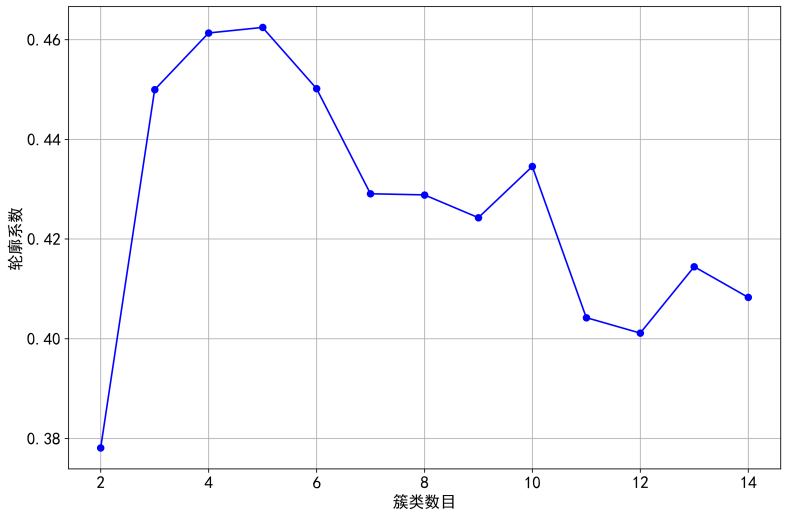

图 **4**　不同簇类数的轮廓系数

轮廓系数本质上衡量的是每个样本点到其簇内样本的距离与其最近簇结构之间距离的比值，轮廓系数系数越大，代表聚类越好。由上图可知：簇类数目在 2 到 5 时，轮廓系数呈上升趋势。在 5 之后，总体上呈下降趋势。为了便于解释，本文选择 4 个簇，对数据进行层次聚类，得到聚类效果图如下：

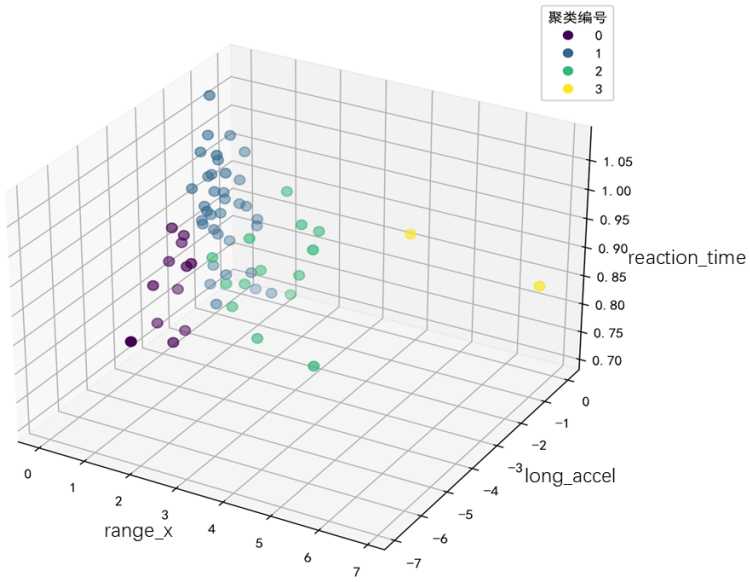

图 **5**　聚类效果图

对每一类驾驶员的解释如下：

- 类别 0（激进型）：这一组驾驶员习惯保持较小的车距，反应时间较短，刹车力度（应激反应）更为强硬。表明他们在面对突发情况时表现出更为激进的反应特征，靠近前车且刹车迅速。这类驾驶员最有可能是“激进型”驾驶风格的典型代表。他们虽然在应对突发情况时具备较快的反应能力，但其较大的应激反映和低车距可能增加事故风险，需要特别关注安全教育。
- 类别 1（老练型）：这一组驾驶员保持车距较近，刹车力度（应激反应）中等偏低，反应时间较为适中。这三个特征反映了他们在驾驶时更倾向于较为紧凑的跟车距离，更注重行驶效率；在紧急情况下即使刹车反应时间有长有短，但都可以采取较为温和的刹车强度。是典型的“老练型”驾驶风格。
- 类别 2（均衡型）：这一组驾驶员驾驶时车距保持适中，刹车力度（应激反应）较强，反应时间也较快。这组驾驶员表现出较为均衡的驾驶风格，既保持了合理的谨慎性，又在紧急情况下敏捷地做出反应。是典型的“均衡型”驾驶风格，既不激进也不保守，适合大多数正常驾驶场景。
- 类别 3（保守型）：这两位驾驶员习惯保持很大的车距，刹车力度（应激反应）偏弱，反应时间较快。说明他们在行车安全性上表现出高度的谨慎，属于典型的“保守型”风格，可能是新手。他们倾向于尽可能保持安全距离，应激反应也比较温和。