# 一、客户聚类

客户是本文的研究主体，客户行为反应了市场对于医药产品的需求和偏好。本文数据集较大，有客户样本161,086个，因此聚类分析时需要考虑计算复杂度。

## 基本原理

### K 均值聚类

【公式和介绍就不写了，这里我只写一下复杂度分析，后同】

对于K均值聚类，其时间复杂度大约为，其中n是样本数量，k是簇数量，d是特征维度，iter是迭代次数。其空间复杂度主要取决于存储数据集和聚类中心的空间占用，大约为，其中n是样本数量，d是特征的维度，k是簇的数量。

### 流形学习

流形学习是用于非线性降维和数据可视化的方法。其假设数据分布在低维流形（manifold）而随机分布在非高维空间，其特性适用于描述数据的非线性结构。常用的流形学习算法包括t-SNE（t-distributed Stochastic Neighbor Embedding）、Isomap（Isometric Mapping）、LLE（Locally Linear Embedding）等。

【可能需要扩充写一下，个人感觉不需要写太多具体原理，因为由于计算复杂度太高，本数据集无法应用】。

流形学习的时间复杂度取决于具体算法，流形学习的时间复杂度一般比K均值聚类高，通常大于，其中n是样本数量，d是特征的维度。由于其需要计算样本之间的距离矩阵或者邻接矩阵，并进行特征值分解等操作，因此复杂度较高。流形学习的空间复杂度取决于具体的聚类算法，但通常情况下，流形学习的空间复杂度与样本数量和特征维度有关，可能比K均值聚类更高。

### 轮廓系数

【参考网上资料】

轮廓系数由于计算复杂度太高，也没有用，改用后面两种指标。

### Calinski-Harabasz 得分

【参考网上资料】

### Davies-Bouldin 得分

【参考网上资料】

## 实验结果

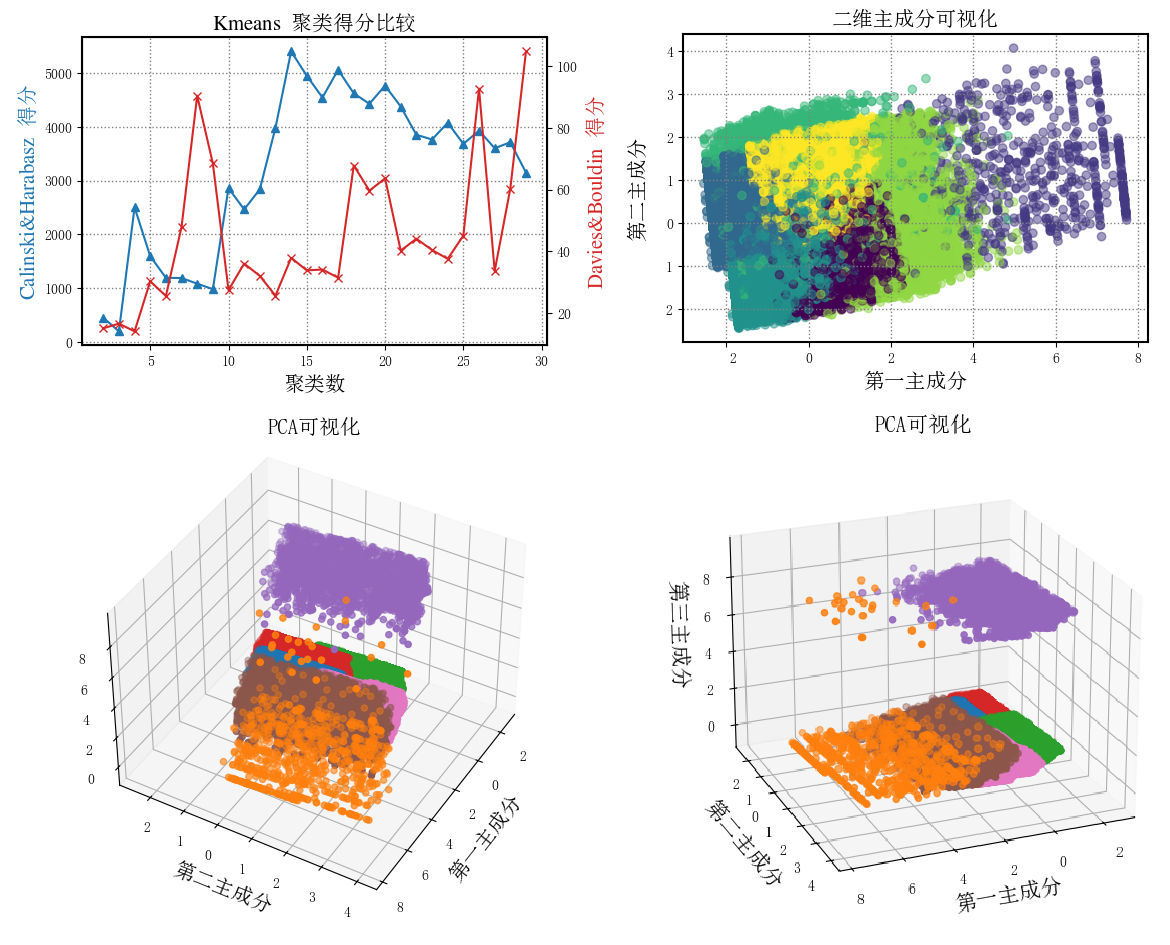


图 1 聚类结果与可视化

图 1左图中横坐标是聚类簇个数，左右两边颜色对应两个评判指标分数两个分数在不同聚类数下评判结果具有显著差异，现实业务中也不允许将客户类型划分太大，因此采用聚类簇为7，通过PCA降维得到图 1右图。我们进一步将特征降到三维并且可视化绘制得到图 2，其中不同颜色表示不同聚类类型。

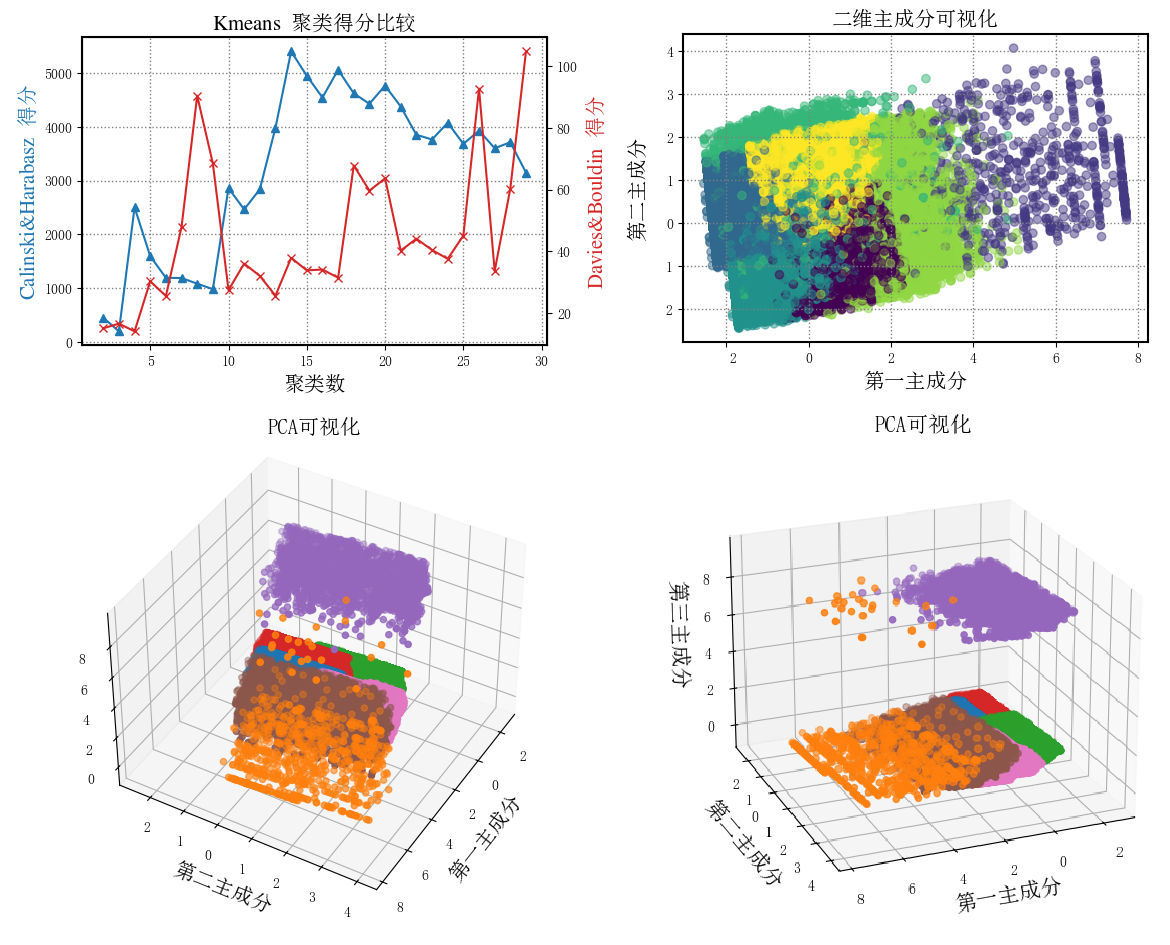


图 2 三维特征可视化

观察发现该数据形态接近流形，但受限于计算复杂度，很难采用流形学习方法聚类。