数据分析与数据挖掘 Homework 2

周子龙 1851201

April 5, 2021

摘要

在第一问的糖尿病人的数据中，并运用了自定义的特征向量相似度来作为距离计算的方式，在对数据初步了解后，剔除了所有的关键特征确缺失的数据，并且通过领域知识提出了一些无关特征，最后使用了高斯混合模型以及 K-均值聚类的方法对该数据进行聚类分析，得到了一些可解释的内容。在第二问中，我首先通过对所有用户观看数据按照用户机顶盒卡号进行了几个维度的统计，并通过该统计数据进行 k-均值聚类，得到了一些可解释的结果。

1 problem1

1.1 预处理

第一问是关于 UCI 的糖尿病人数据进行分析，得到数据后，我查询了一些糖尿病领域知识，并试图对原始数据进行初步的特征筛选。

1.1.1 剔除重复记录

之后我对数据进行了描述性统计分析，发现有部分病人（patient_nbr 相同）具有多条记录，以 patient_nbr 为 1152 的病人为例，每一条记录代表了该病人一次的住院纪录，不同字段纪录了该病人的一些生理特征以及住院数据。出于简单考虑，我对所有相同病人的多次记录按照 encounter_id 排序后取第一个，其他全部剔除。

1.1.2 特征数值化

由于在本数据集中,较多字段是以“Yes”以及“No”这样的方式出现,为了方便数值处理,我对全部的这类字段建立了一个字段映射表，比如将 “Yes” 处理为 1，“No” 处理为 0。需要指出的是：“glyburide-metformin”、” glipizide-metformin “、” glimepiride-pioglitazone “、” metformin-rosiglitazone “、” metformin-pioglitazone “，这几个字段在整个数据集内全部取的是 “No”，对分类结果没有任何影响，故可以直接删去。

1.1.3 归一化处理

本数据集中存在部分描述性特征以及数值性特征，他们所表示的特征信息不同，数值间尺度不一，为了方便后续处理，在这里我对所有数据进行了 minmax 归一化处理，将所有特征映射到 0-1 之间的

一个数值。由于我在第一次处理的时候没有进行归一化，导致了所有数据距离几乎由某一个尺度较大的特征影响，致使整个数据几乎不可分。

1.1.4 自定义距离

在上文中提到的归一化处理仅仅适用于可以量化的距离相比较，对于描述性特征进行归一化处理是没有意义的。因为这一映射完全是认为界定的，假设特征 A 由三个可能的描述构成：a，b，c；认为地将其分别映射为 1，2，3 后进行归一化处理，a，b，c 之间的“差异性”并不能够按照简单的欧式距离来定义。在缺乏领域知识的情况下，不妨直接按照如下定义来计算特征 A 之间不同取值的距离：

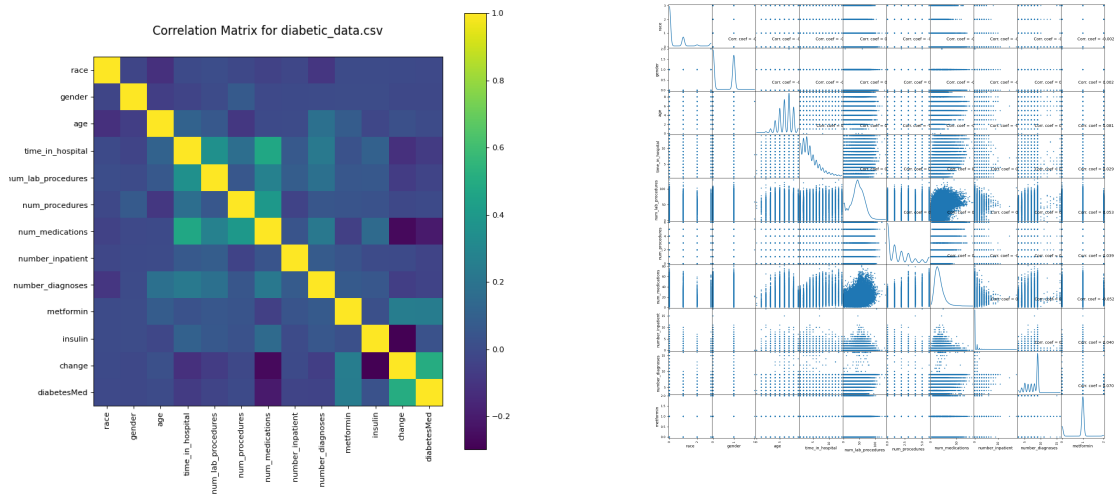
$$dis(x_1, x_2) = \begin{cases} 1 & \text{二者完全相同} \\ 0 & \text{否则} \end{cases} \tag{1}$$

1.1.5 上采样

由于部分数据过于离散，如果不加处理直接运用将会得到一个特别稀疏的特征矩阵，为了避免这一点，我对其中一些较为稀疏的特征进行了一些随机组合的卷积操作，将其中 25 维稀疏特征卷积得到了一个 5 维的上采样结果，添加到余下的特征中进行分类。

1.1.6 描述性统计

进行完如上操作后，我对最后的特征进行了协方差分析，得到了如下的协方差热力图以及不同特征之间的两两散点图。



(a) 协方差热力图　　(b) 协方差热力图

图 1: 对原始数据进行描述性统计得到的结果

从图中可以看出，数据集中的部分描述性统计量如：住院次数、手术次数、等具有一定的相关性，而其他特征几乎不具有相关性。

1.2 确定聚类数量

1.2.1 通过可视化方法确认

由于高维数据不利于可视化，为了得到更为直观的效果，我对原始数据进行了主成分分析，得到如下结果：

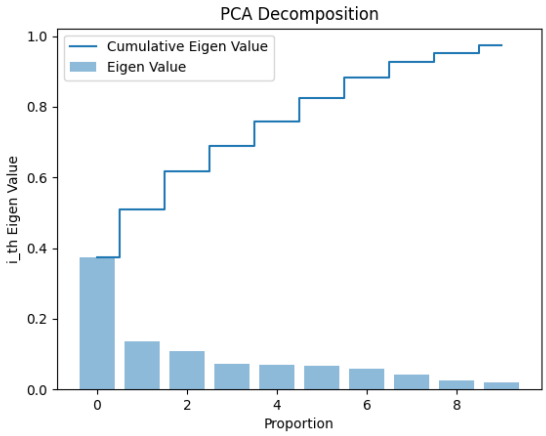


图 2: 成分碎石图

通过提取 2 和 3 个主成分，数据的方差解释性分别为 50.9% 和 61.7%。对原始数据绘制散点图后得到如下结果：由上图，我们可以很明显的看出在二维空间中 3 个类别较为合适；在三维空间中似乎

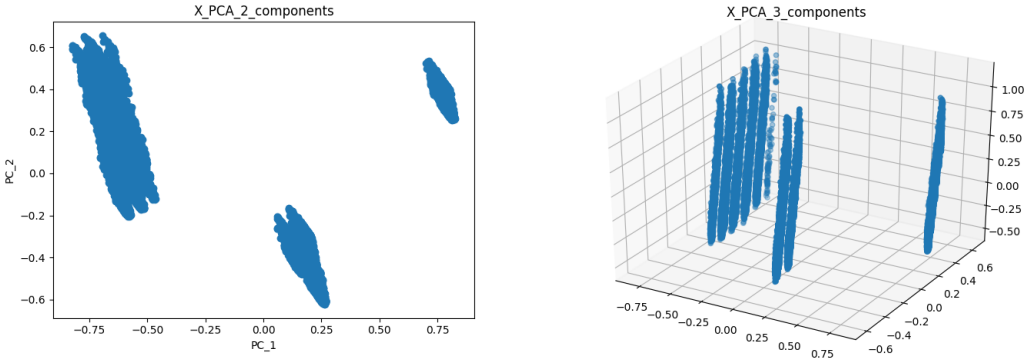
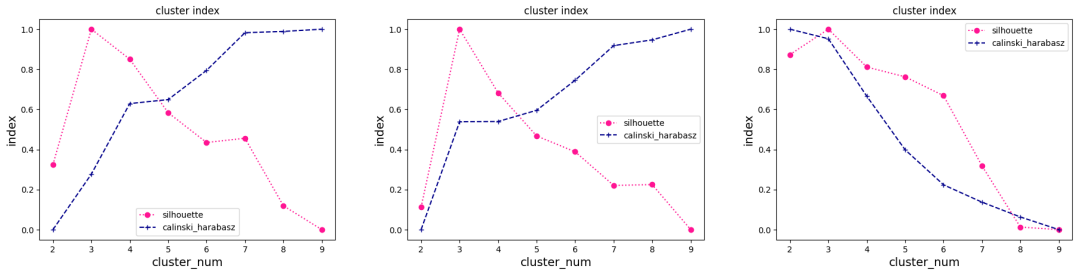


图 3: 提取成分得到的散点图

有很多较为狭长的高斯分布，在考虑聚类算法是可以优先考虑应用高斯混合模型进行逼近。

1.2.2 通过一般评判标准确定

较为一般的聚类评判标准有许多，在这里我使用了 silhouette 以及 calinski harabasz 标准，其他的标准也多是基于类间、内方差进行判断，在趋势上和后者差不多。而 silhouette 是对每一个样本点进行单独计算，故而效率较为低下，但是结果更为可信。通过对原始数据、选取两个特征的数据、选取三个特征的数据进行 k-均值聚类并使用如上的两个标准进行判断，得到了如下结果：



(a) 2-组分特征聚类得分　(b) 3-组分特征聚类得分　(c) 原始特征聚类得分

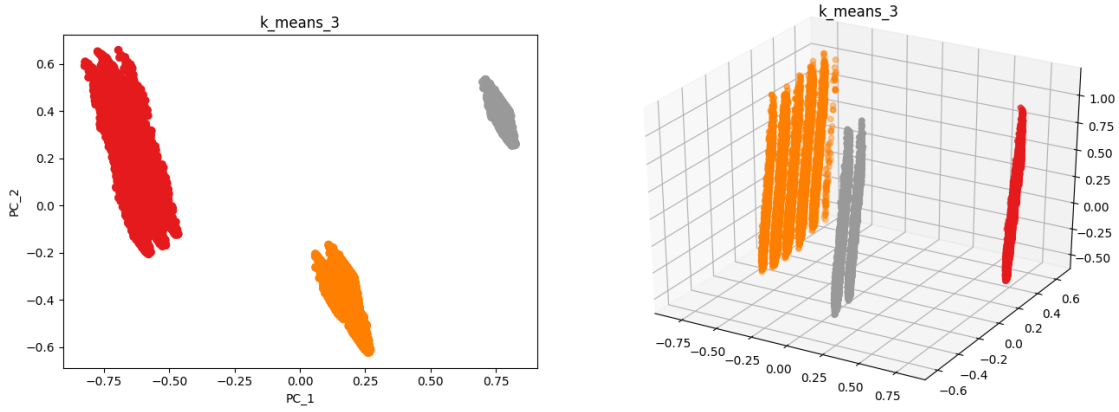
图 4: 聚类得分

通过上图，可以很明显的看出，silhouette 在各个组分中 3 个聚类的情况下均达到了极大值，故而应当选取 3 作为聚类数进行聚类。但是这里的结果对于 calinski harabasz 标准来说有些奇怪，在原始数据中该得分随着组分的增加逐渐下降，于提取部分组分得到的结果趋势不一致，可能需要进一步分析。其次，calinski harabasz 标准无法准确判断应当选择何种组分，若选择第一个平台拐点处，则对于 2 特征组分应当选取 4 个类别，3 特征组分应当选取 3 个类别，原始特征无法准确分析。

1.3 聚类结果与分析

1.3.1 K-均值聚类

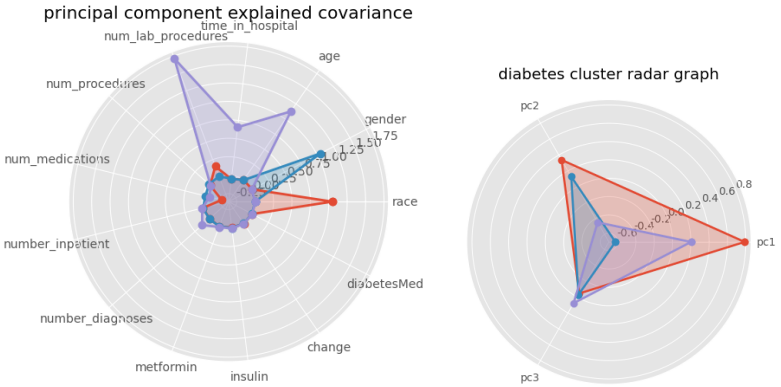
对各组分数据进行 3-均值聚类得到的结果如下：



(a) 3-均值聚类在 2 个主成分下的可视化结果　(b) 3-均值聚类在 3 个主成分下的可视化结果

图 5: 聚类结果

其个组分组成以及对不同成分下的聚类结果的聚类中心结果如下：



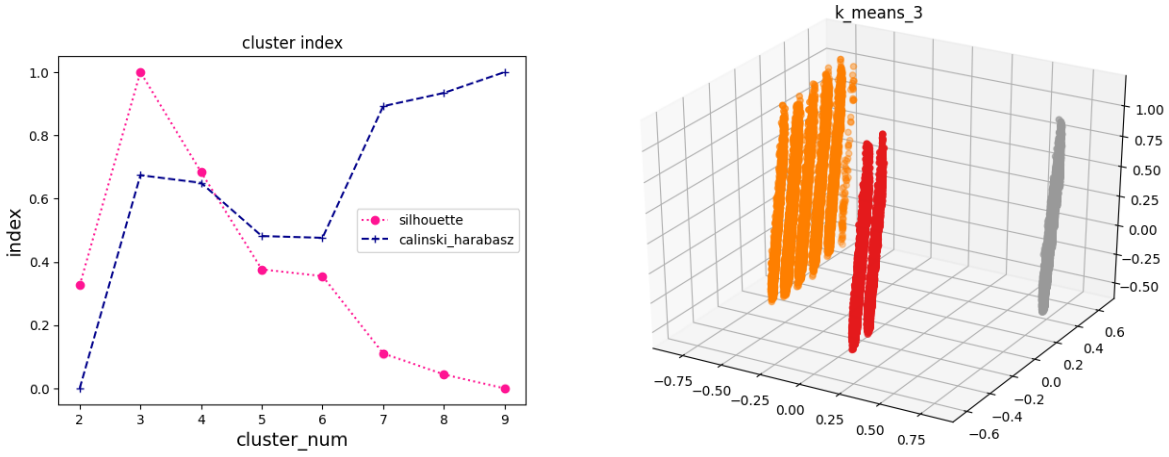
(a) 3 组分方差解释性　　(b) 进行 3-均值聚类得到的雷达图

图 6: 主成分解释

其中的三个组分的方差解释性分别为：82.95%，0.13.66%，1.27%，累计方差解释为：97.88%；可以很明显的看到，在聚类中，第一组分的方差主要贡献为：年龄、手术次数、住院时间，以及较少部分的诊断次数；第二组分的主要贡献为性别；第三组分为人种；这可以理解为第一组分主要为患者的医疗数据，这在不同患者之间产生了较大的差异，我们的手成分也是按照这个分布的。在第一类中，我们可以认为某一性别以及某一人中的患者更多。

1.3.2 高斯混合模型聚类

对于高斯混合模型，其聚类得分以及最高得分结果如下：



(a) GMM 聚类数得分　　(b) 3-GMM 聚类结果

图 7: GMM 模型得到的聚类可视化结果

分析不再赘叙述

1.3.3 进一步处理

由于缺少领域知识，我无法对患者的医疗住院数据进行聚类，这里仅仅是按照了这些组分进行的聚类分析。在获得相关领域知识后，可以对数据更好的进行预标注，可以尝试单独利用患者的医疗检测数据数据进行组分分析，也可尝试将医疗检测数据与描述性数据进行结合，应当可以得到更好的可解释的聚类结果。

2 problem2

2.1 数据预处理

对第二问的用户数据进行聚类前，我首先对用户收藏记录（下称收藏记录），用户常看直播频道记录样例（下称直播记录）进行了大概查看，发现其数据量巨大，收藏记录大概是百万级别的数据量，直播大概是十万级的数据量。

2.1.1 缺失值与异常值处理

这两类数据中有部分明显缺失值以及明显异常值，如在直播记录中，SID 为 602 的频道名称有：欢笑剧场、幸福剧场；在收藏记录中，CODE 为 OTHE1000000000892860 分别有两种不同的频道内容：你是我的答案、名侦探柯南 16；进一步分析，我认为在仅有的数据中完全无法区分这种异常情况，以及在若以 CODE 为主键进行筛选聚类数据过大，导致特征急剧膨胀，因此此类异常不影响我的聚类结果。

2.1.2 原始数据筛选

对于一些语意上完全重复的字段，我直接删去了其中一个，如 CODE、MD5、频道名称完全为一类特征的不同表现形式，在实际处理中，我仅仅保留了 CODE 作为该类特征的表现形式。进一步观察，我发现 FOLDERCODE，指的是节目 ID，而 CODE 可能是用户收藏记录的唯一主码，如在已知数据集中，节目名为：“亮剑”的记录有 255 条，其 FOLDER_ID 均为 4d007de12f2a241bd7000155，但每条记录的 CODE 各异。

2.2 特征提取

对于收藏记录,我对原始数据按照 STBID 进行聚合,对聚合后的数据按照 SHOW_TYPE 进行统计,在可见数据集中,SHOW_TYPE 字段仅有四种可能的取值,每当某一用户数据出现一类 SHOW_TYPE 的记录，我就对该用户的特称字段的相应内容进行加一操作，此外，由于某一个 CHANNEL_ID 以及 FOLDER_ID 字段取值可能行过大，在缺少相关领域知识的前提下我选择不对这些字段进行采样。

对于直播记录，我也采取的是以 STBID 为主键进行聚集的策略，在进行初步的统计分析后，我发现所有可能的电视节目 ID 有 283 种，大部分用户的记录为 20 条左右，同样地，这样也面临着维度过高，特征过于稀疏的问题。一个可能比较好的做法是将这些电视节目按照一定的条目归类，对所有同类的电视节目的观看记录进行叠加，但是由于缺少相关领域知识以及时间，我没有采取这种特征提取的方法。

出于简单考虑，我对原始数据进行了简单的统计意义上的采样，我一共提取了某一用户的如下统计特征：收看的频道数目 (num)，收看的的累计观看次数 (sum)、收看频道的标准差 (std)、平均收看次数 (mean)、以及达到 80% 的收看次数的频道数目 (mostN)；通过这几个特征进行聚类，我试图对用户从行为上进行聚类分析。

当然，如果条件允许，完全可以融入一些相关的领域知识，以便对用户的喜好等进行分析。

对提取到的特征进行描述性统计得到如下结果：

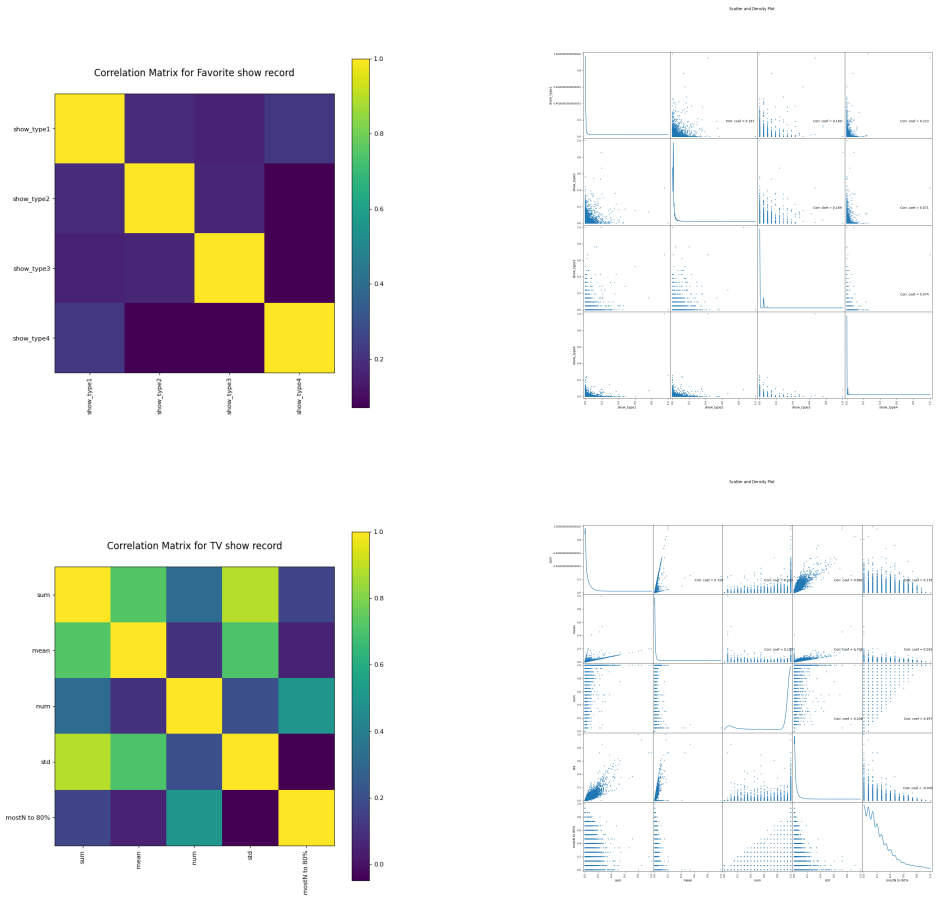
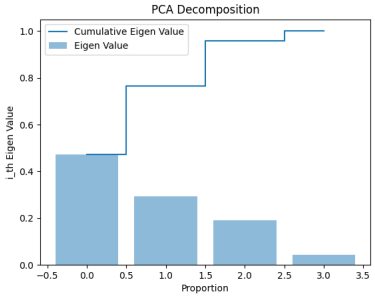
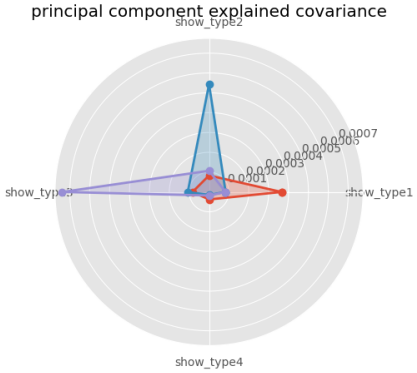


图 8: 上为用户收藏数据的相关结果，下图为直播记录的相关结果

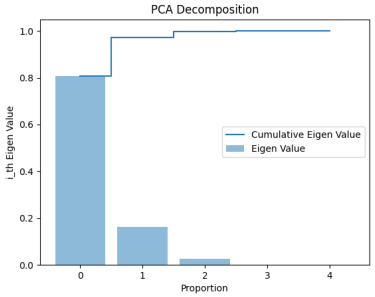
其中，对于这两者的数据的主成分分解结果如下：



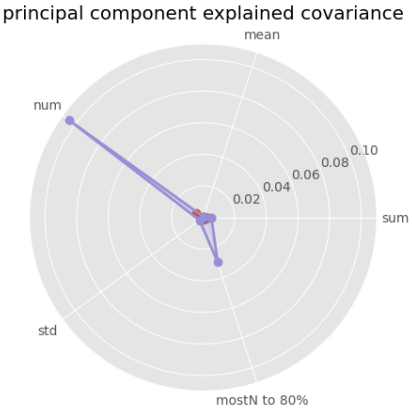
(a) 收藏数据主成分碎石图



(b) 收藏数据 3-组分方差解释



(c) 直播数据主成分碎石图

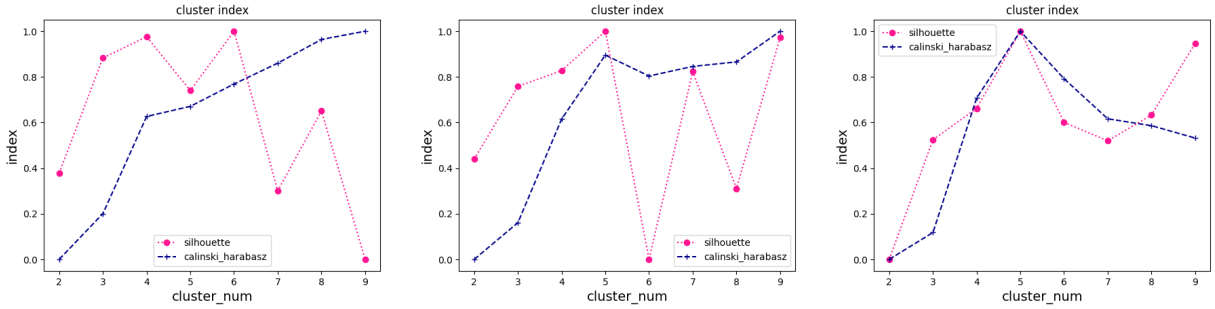


(d) 直播数据 3-组分方差解释

2.3 聚类与结果分析

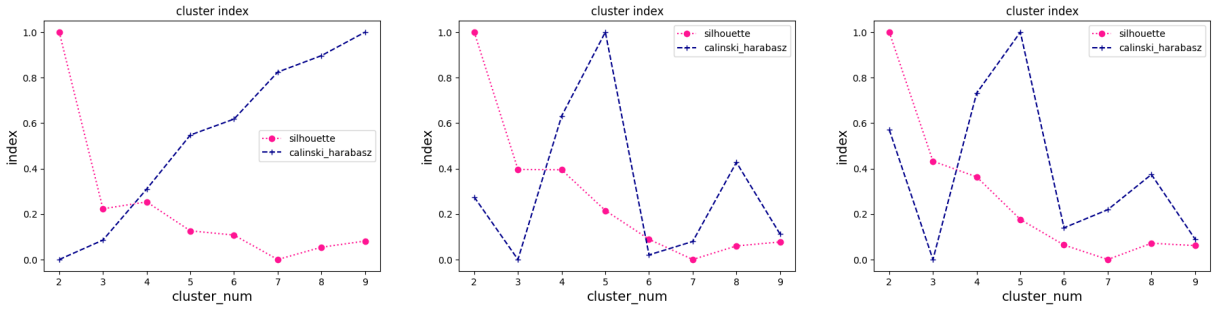
2.3.1 K-均值聚类

对如上数据进行 K-均值聚类后进行结果判断，得到下表： 由图，对于收藏数据，选择 5 个组分比



(e) 2-组分聚类得分　(f) 3-组分聚类得分　(g) 原始数据聚类得分

图 9: 收藏数据聚类结果得分图



(a) 2-组分聚类得分　(b) 3-组分聚类得分　(c) 原始数据聚类得分

图 10: 直播数据聚类结果得分图

较合适，对于直播数据，选择 2 或 5 个组分比较合适；同样地，在这里也出现了不同维度下聚类得分的差异性问题。

2.3.2 结果分析

这里直接给出二者的聚类中心数据:

	show type 1	show type 2	show type 3	show type 4
cls1	31.1	82.5	554.9	6.4
cls2	35.6	103.4	0.0	16.0
cls3	117.5	1098.3	134.1	31.9
cls4	736.6	1113.6	3474.6	172.1
cls5	1393.9	264.2	158.7	52.8

表 1: 收藏记录聚类中心，数据进行了一些处理

由于缺乏领域知识，这里不便于分析。

	sum	mean	num	std	mostN 80%
cls1	46.0	9.4	994.3	18.4	526.9
cls2	39.3	8.2	969.1	30.2	166.9
cls3	5.8	4.9	264.0	9.2	46.9

表 2: 直播记录聚类中心，数据进行了一些处理

这里可以粗略地将用户分为三类，第一类用户收看的频道数目较多，种类大，次数多，并且并没有特殊的偏好，可能是多年龄段群体家庭；第二类收看总数略少，但是总体量也比较大，有特殊的偏好，喜欢收看特定类型的频道，可能为单一年龄段家庭；第三类观看电视数目明显较少，并没有特殊偏好，并且基本上集中在少部分频道上，可能对应老年群体。

最后是一点点对作业的小建议：真的觉得作业有点多。而且不是太明确，我个人感觉大量的时间花费在读数据，处理数据，出图，写报告上，在做聚类上做的反而不是特别多。如果可以的话希望以后作业能够更加明确一些些？比较有针对一些，可以及时训练一下学到的算法之类的结果，最后的大作业可以稍微开放一些，做一些探索性的尝试。