2023 年第二届“钉钉杯”大学生
大数据挑战赛论文

题　　目：　智能手机用户监测数据分析

摘要

经过数据探索，发现原始数据存在一定的错误、冗余及缺失。数据清洗后，尝试从原始数据（数值、类别、时间）中提炼统计特征，以便于问题一和问题二的解决。

本节首先选取合适量化指标预处理，然后分别采用原型（K-Means++）、密度（DBSCAN）、层次（AGNES）三种聚类算法对用户进行聚类，遵循“肘部法则”选择合理的聚类数量 K 值；最后，根据聚类结果对不同类别的用户画像，分析不同群体用户的特征。

cumcmthesis 是为全国大学生数学建模竞赛编写的 LaTeX 模板, 旨在让大家专注于论文的内容写作, 而不用花费过多精力在格式的定制和调整上. 本手册是相应的参考, 其中提供了一些环境和命令可以让模板的使用更为方便. 同时需要注意, 使用者需要有一定的 LaTeX 的使用经验, 至少要会使用常用宏包的一些功能, 比如参考文献，数学公式，图片使用，列表环境等等. 例子文件参看 `example.tex`.

2020 年建模比赛格式变化说明

今年的格式变化主要就是三个地方，如下：

1. 论文第一页为承诺书，**内容进行了调整**。
2. 编号页格式进行了格式调整。
3. 这是 19 年调整了，这里延续说明下。论文正文（**不要目录**，尽量控制在 20 页以内）；正文之后是论文附录（页数不限）。

`https://www.latexstudio.net` 陆续推出了更优质的资源，欢迎学习。

欢迎大家到 QQ 群里沟通交流：91940767/478023327/640633524。我们也开通了问答区交流 LaTeX 技术：`https://ask.latexstudio.net`，欢迎大家前来交流，有问题就来这里，与大神零距离。

关注我们的微信公众号：

关键词：TEX　图片　表格　公式

目录

一、 问题重述

1.1 问题背景

智能手机已成为现代社会人们生活不可或缺的一部分，其普及和发展给人们带来了巨大的生活便利和娱乐享受。近年中国智能手机市场品牌竞争进一步加剧，中国超越美国成为全球第一大智能手机市场。随着智能手机市场快速增长，智能手机用户群体愈发多样，智能手机软件满目琳琅，研究智能手机用户的行为模式和使用偏好对于理解用户需求、预测用户行为和优化产品与服务具有重要意义。通过对智能手机用户监测数据的分析，可以为智能手机制造商、软件开发者、广告商和营销人员等提供有益的信息及有价值的洞察，指导他们制定战略和决策，更好地贴合用户需求并提供更佳的用户体验。

1.2 问题要求

问题一 针对问题一，赛题要求 (1) 根据用户常用所属的 20 类 APP 的数据对用户聚类，(2) 对不同类别的用户画像，分析不同群体用户的特征。

二、 数据探索、预处理与特征提取

2.1 符号说明与数据概览

原始数据集包含 app 类别辅助表格（A.app_class.csv）与 21 天监测数据（B^*.dayxx.txt），来源、符号、意义及数据类型如下表 1 所示。

表 1　数据集原始特征

来源	符号	意义	类型
$A\&B^*$	$appid$	用户的 id，唯一标识一名用户	类别变量
A	app_class	应用的 id，唯一标识一个 APP	类别变量
B^*	app_type	APP 类型：系统自带、用户安装	类别变量
B^*	$start\_day$	使用起始天，取值 1-30	数值变量
B^*	$start\_time$	使用起始时间	时间变量
B^*	end_day	使用结束天	数值变量
B^*	end_time	使用结束时间	时间变量
B^*	$duration$	使用时长（秒）	数值变量
B^*	up_flow	上行流量	数值变量
B^*	$down\_flow$	下行流量	数值变量

将 B^* 与 A 进行“左连接”得到 B（同时舍弃重复值），app_class 为 a～t 的代表 A 中 20 个常用类别（B 含 3931 种）；NaN 则代表所属类别未知的不常用 APP（B 含 32506 种）。B 中，在类别已知的常用 20 类 APP 中，t 类数量最多（1406），r 类最少（41）。

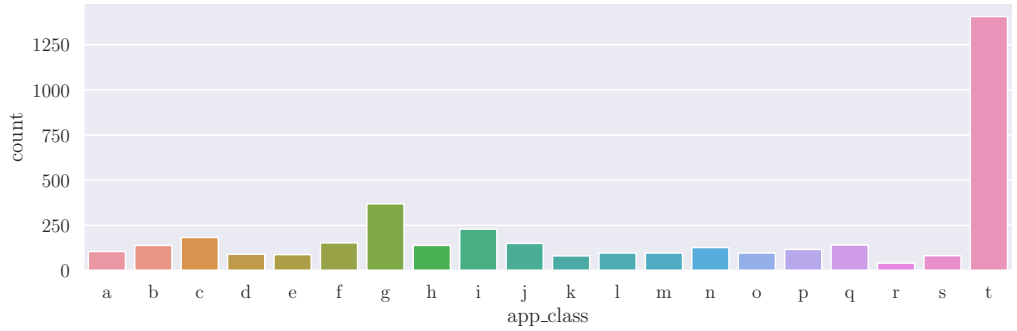


图 1　B 中各类 APP 计数图

2.2 数据探索之类别变量

2.2.1 单一变量

表 2　类别变量统计描述（以 day01 为例）

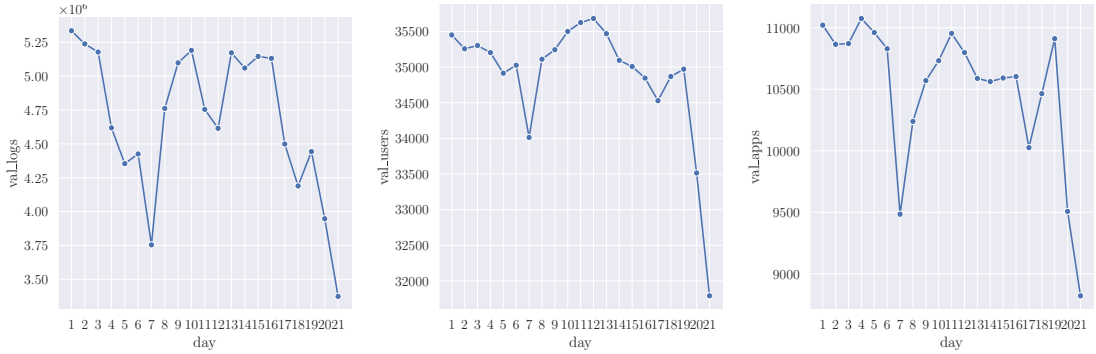
	uid	*appid*	*app_type*	*app_class*
count	5335803	5335803	5335803	5335803
unique	35451	11021	4	21
top	A9E4AAC5B8E05D2A4E35E0D4F2994F37	3309	usr	NaN
freq	2629	924309	2987468	2432606

据表 6，*app_type* 只有两类【系统预装、用户安装】，存在异常。通过数据探索，发现表格存在 ['sys', 'usr', ' 用户', ' 预装'] 四种取值，故将中文全部替换成英文。

app_class 有 21 类，这是因为在"左连接"操作时，将 NaN 也作为一种 APP 类型，这是由于此处数据缺失本身就代表一种资讯（小众 APP），并非随机发生或人为故意。如果将 NaN 也视作一种 *app_class*，那么数据 B 不存在缺失值。

2.2.2 天数变化

另外，从 21 天的类型变量数据可以发现，每日活跃用户、APP、日志条数在各天都有所差异，如图 2 所示。



(a) 日志条数　　**(b) 活跃用户数**　　**(c) 活跃 APP 数**

图 2　day01～day21 类别变量的变化折线图

粗略观察图 2b，活跃用户数在 7 和 21 存在明显波谷，似乎和"星期"有某种关联；对照图 2a、图 2b、图 2c三表分析，似乎 APP 活跃情况（种类、请求）与天数有所关联，甚至可以猜测某些小众 APP 被某些特定用户群体所使用甚至是青睐。

2.2.3 变量关联

APP 自身包含 *appid*、*app_class* 以及 *app_type* 属性，因此可以抽取这三列建立 C。

表 3　APP 统计描述

	appid	*app_type*	*app_class*
count	37276	37276	37276
unique	36437	2	21
top	19582	usr	NaN
freq	2	34153	32958

据表 3，数据探索发现有 839 个 app 多重类型，即对于某些用户而言，该软件为系统预装，对另一些而言，则为自行安装。这似乎表明，不同的用户在软件下载与安装层面，有可相互区分的行为特征。另外，21 类 APP 的数量情况如图 3 所示。

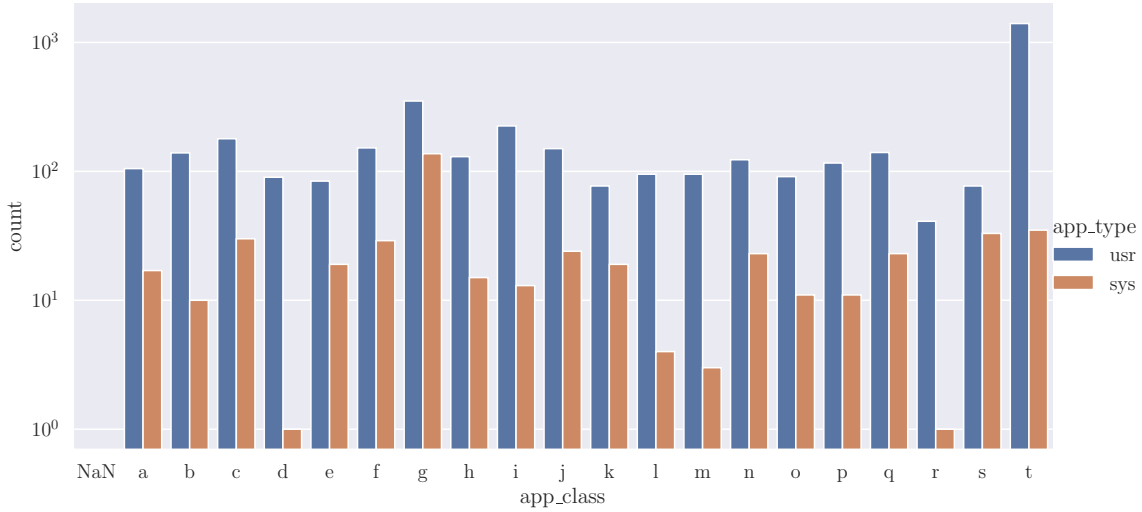


图 3　21 类 APP 计数图

据图 3，未分类的 APP 并不是小数目（有一部分为系统预装），t 类在 APP 多样性表现上依然出众。此外，不同种类的 APP 就类型（预装、用户）而言相差较为悬殊，例如：g 类，系统预装相对较多；其余类别，用户预装较为普遍，尤其是 d 和 r 类。

值得区分的是，在新建 APP 统计数据 C 中，计数图反映了各类 APP 的多样程度（市场垄断程度）。而在监测数据 B 中，某个（类）APP 请求日志（行数）计数图则反映了该用户在使用各类 APP 时的活跃（点击）行为，这将在下一小节进行探索。

2.3 数据探索之数值变量

2.3.1 单一变量

表 4　数值变量统计描述（以 day01 为例）

	start_day	*end_day*	*duration*	*up_flow*	*down_flow*
count	5335803	5335803	5335803	5335803	5335803
mean	0.975107	1	2151.604772	607572.168995	158163.759549
std	16.843899	0	1455335.155631	11015502.975274	6538529.614936
min	-16524	1	1	0	0
25%	1	1	3	0	0
50%	1	1	10	0	0
75%	1	1	36	1278	1063
max	1	1	1427769883	3639473769	3292713011

数据具体说明指出：“*start_day*：使用起始天，取值 1-30（注：第一天数据的头两行的使用起始天取值为 0，说明是在这一天的前一天开始使用的）。”然而，表 4 显示其最小值为 -16542，因此可以判断 *start_day* 存在异常值；而这会直接导致 *duration*、*up_flow*、*down_flow* 偏差过大。因此，须要对这两列进行数据清洗，删除异常值。

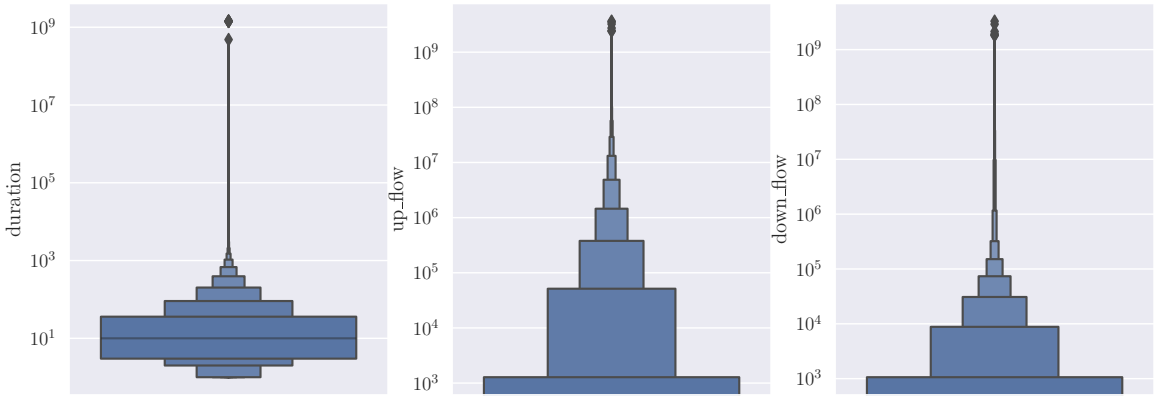


图 4　使用时长、上下行流量增强箱形图（以 day01 为例）

据统计，在第 1～21 天监测数据中，99.98% 的记录使用时长不超过 9158.02。针对异常案例进行分析，例如，$uid = 64B3E40461C56847F35DB46D55707EA4$ 用户：

表 5　异常案例

appid	app_class	start_day	start_time	end_day	end_time	duration
4803	a	19	00:52:46	19	07:47:59	24912
18478	c	19	07:48:25	19	07:48:38	12
	：		：		：	
˙6192	NaN	19	20:46:11	19	20:46:29	18
3309	f	19	23:14:49	19	23:16:38	109

凌晨零时至早上七时的记录是不符合生活常态不可持续的，猜测是应用后台驻留、系统故障或用户因故未关闭应用。不过，是否存在异常行为可以作为一个新的特征。因此，本文对持续时间超过 9159 的认定为无效使用时长，不能真实反映用户的行为特征。

此外，以第一天为例，不同用户在 $appid$ 个数、$duration$ 总时长、up_flow、$down\_flow$ 总流量、日志行数层面有所差异，分布如图 5 所示。

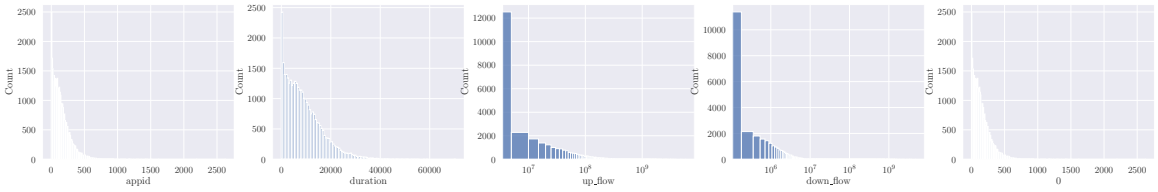


图 5　直方图（使用 APP 数量、使用有效总时长、消耗上下行流量、日志行数）

可以观察到某些用户较依赖手机，日志数、使用时长、APP 多样性、消耗流量偏多。

2.3.2 天数变化

探索发现，可为每名用户绘制使用 APP 总数/时长/流量/日志随天数变化的折线图，以展现用户在月（周）级别的变化趋势与独潜在规律，并据此将用户群归类。

2.3.3 变量关联

更细粒度地，对于每名用户/每天，可绘制其各类 APP 的个数/时长/流量/日志情况，以了解用户对不同类型 APP 的适用情况、青睐程度；更进一步，还可以绘制 24 小时各类 APP 使用情况，这样便于了解用户的作息、通勤、活跃时段等信息。

2.4 直觉小结与用户量化

2.4.1 数据直觉

概括来说，赛题数据 A 给出 4000 多个常用 APP 所属类别：这是 $1:1$ 的数据；而赛题数据 B 则记录了每名用户（uid）每日每时使用各款 APP（$appid$）的起始时间，使用时长，上下流量等信息：这是 $1:N$ 的资料。可以使用统计量进行归纳。

具体而言，监测数据蕴藏大量用户行为特征，例如：

- 用户一天使用多长时间的手机（可间接反映依赖程度、年龄）
- 用户平均多长时间看一次手机（可间接反映依赖程度、年龄、工作）
- 距离上一次上线，隔了几天（可间接反映依赖程度、年龄、工作）
- 在什么星期几的时间段最常用什么类型 APP（可反映工作、生活）
- 用户早、晚各使用什么类 APP（可反映工作、生活、作息）
- 用户周末、工作日各使用什么类 APP（可反映工作、生活、作息）
- 用户最早什么时候开始使用手机、什么时候结束使用（可反映工作、作息）
- 用户最常用什么类型 APP（可反映喜好）
- 哪一类 APP 使用最频繁、哪一类使用时长最多（可反映喜好）
- 用户一共安装了多少个 APP（可反映保守程度、对多样性的接纳程度）
- 系统预装与自行安装的比例（可反映保守程度、对多样性的接纳程度）
- 用户的每月流量使用情况（可间接反映财富程度、年龄）

2.4.2 用户模型

用户量化是指将现实生活中的“用户实体”进行抽象，采用不同维度的量化指标建模，即将其视为 n 维空间的一个点，使用形如 $X=[x_1,x_2,\cdots,x_n]$ 的数学符号表示。

基于对数据集的深入探索及理解，提出简易用户模型：

表 6　简易用户模型

符号	意义	维度
uid	用户的唯一标识	1
$DU_{d,h,c}$	该用户在第 d 天 h 时内使用 c 类 APP 时，投入的总计时长	$d\times h\times c$
$UF_{d,h,c}$	该用户在第 d 天 h 时内使用 c 类 APP 时，消耗的上行流量	$d\times h\times c$
$DF_{d,h,c}$	该用户在第 d 天 h 时内使用 c 类 APP 时，消耗的下行流量	$d\times h\times c$
$NO_{d,h,c}$	该用户在第 d 天 h 时内使用 c 类 APP 时，记录的日志行数	$d\times h\times c$

三、 问题一：聚类分析与用户画像

3.1 特征工程与评价指标

3.1.1 特征选择与数据降维

聚类指将数据样本对象划分成若干类（簇、标签）并尽可能的保证“类内紧凑”、“类间独立”[3]。不同的量化指标、不同的相似度量（距离定义），往往会带来迥异的聚类结果。一般来说，量化指标维度数目越多，算法运行时间越长、结论可解释性越弱。

关于用户画像的量化指标，陈 [5] 等人、成 [6] 等人从各类日均屏幕使用时间切入；武 [7] 等人从 APP 数量、阅读时间、消费等特征对阅读类 APP 使用人群进行聚类解读；侯 [8] 针对每日手机使用时长、使用频次、使用偏好等特征对用户进行建模；韦 [9] 基于“安装数量”、“打开次数”、“使用时长”、“工作日使用时长”、“周末使用时长”构建用户特征。

首先，尝试选择前 7 日各类 APP 使用时长、使用频次、上行流量、下行流量之和作为量化特征，共计 $20*4=80$ 维；对于不同的量纲特征，分别扣除均值，除以标准差以进行数值标准化。Pearson 相关热力图 6 显示，各类 APP 时长、频次、流量成弱正相关。

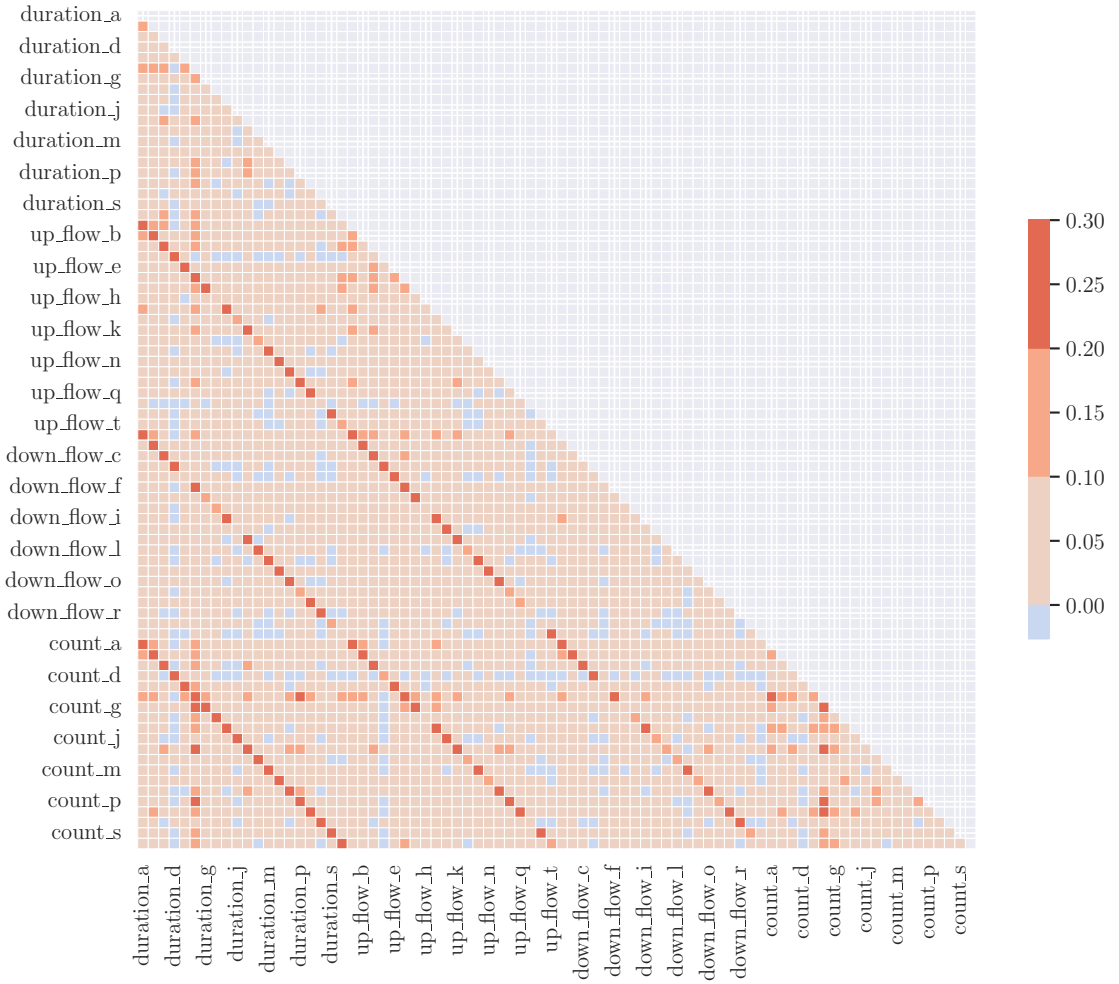


图 6　皮尔逊相关系数热力图

为增加数据易用性，降低计算开销，增强视觉理解，而后采用主成分分析对特征进行变换，并按方差排序表示各维度重要程度，如图 7 ，选定阈值将维度压缩至 3 维。

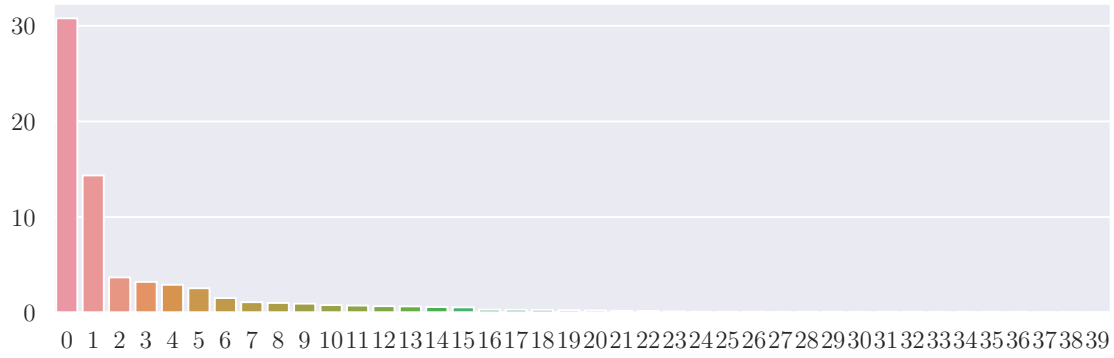


图 7　PCA 特征方差柱状图

3.1.2 评价指标

聚类“好坏”不存在绝对的客观的标准 [1]；聚类数目设定是否“合理”也往往依赖人工先验知识 [2]。聚类数目设定过低，划分粒度不够细腻；聚类数目设定过高，宏观结论的可解释性又受到限制。常用选择聚类数目方法是人为观察聚合系数折线图，大致估计最优聚类数量 K。相关定义如下：

定义 1 *各簇畸变程度：该簇重心与其内部成员位置距离的平方和；假设一共将 n 个样本划分到 K 个簇中，用 C_k 表示第 k 簇，该簇重心记为 u_k，则第 k 簇的畸变程度为：*

$$\sum_{i \in C_k} |x_i - u_k|^2$$

定义 2 *聚合系数：*

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} |x_i - u_k|^2$$

此外，还有 Calinski-Harabasz 系数 [11]、Davies-Bouldin 指数 [12]、Silhouette 轮廓系数 [13] 可用于度量某些聚类目的下的结论性能。

3.2 算法概述与 K 值选择

注：本小节所使用算法及评价指标均采用 scikit_learn[14] 开源库实现。

3.2.1 原型聚类：K-Means++

K-Means 是一种简单、高效的聚类算法，假设聚类结构能通过一组“原型”刻画，算法的主要思想是通过迭代过程把数据集划分为不同的类别，流程如图 8 。K-means++ 优化“初始化 K 个聚类中心”，要求初始的聚类中心之间的相互距离要尽可能的远，在“孤立点数据敏感性”方面优于 K-Means 算法。默认采用欧式距离、重心法进行相似度量。

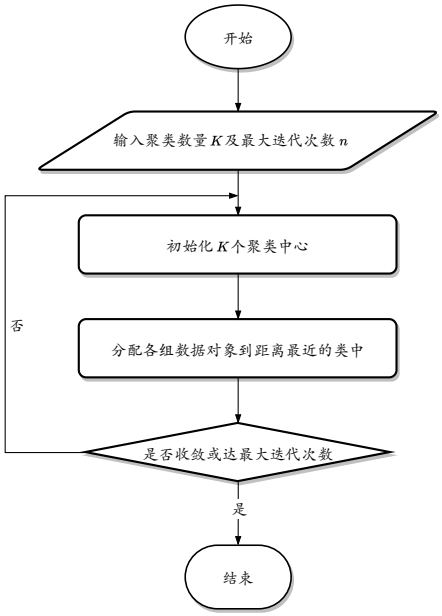


图 8　KMeans 算法流程图

将最大迭代次数设置为 1000，选择 K 等于 2～50 绘制聚合系数与卡林斯基-哈拉巴斯指数折线图。根据图 9 ，K 值从 2 到 13 时，畸变程度变化最大；超过 6 畸变程度变化显著降低：因此根据肘部法则，可将聚类数量 K 设定为 5；从来看，应将聚类数量设定为 6 以下。该结论符合卡林斯基-哈拉巴斯指数峰值，故将聚合数目设定为 5。

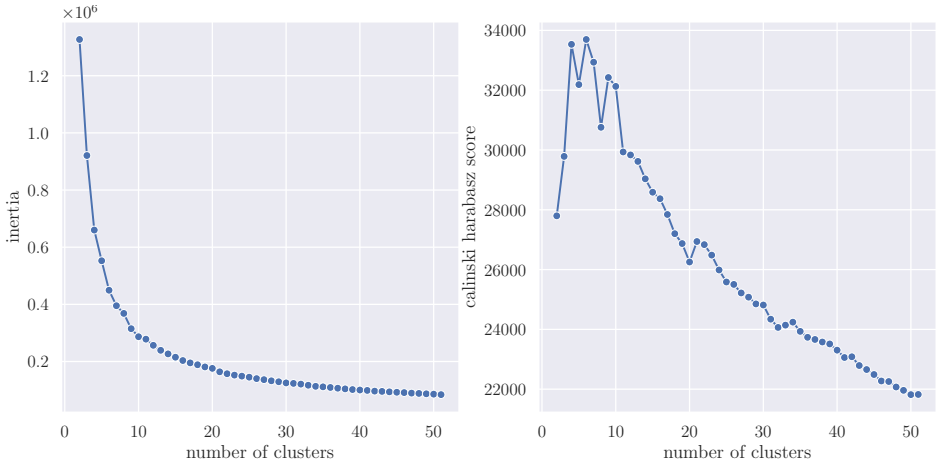


图 9　K-Means 聚合系数与卡林斯基-哈拉巴斯指数

3.2.2 层次聚类：AGNES

AGNES 算法（Agglomerative Nesting），以自底向上方式，不断重复合并，产生不同粒度（层次）的聚类结果，一般最终预设聚类数目为 1。该算法可通过聚类谱系图（dendrogram）可视化，算法执行流程如下：

Algorithm 1: AGNES 算法

input : 样本集 $D = \{x_1, x_2, \ldots, x_m\}$;
聚类簇距离度量函数 d;
聚类簇数 k。

output: 簇划分：$\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

```
# 先将每个样本视作一个初始簇构造;
# 构造 M 个类，每个类仅包含一个样本;
for j = 1, 2, ..., m do
    C_j = x_j
# 两两计算距离;
for i = 1, 2, ..., m do
    for j = i + 1, ..., m do
        M_{i,j} = d(C_i, C_j);
        M_{j,i} = M_{i,j}
# 当前类个数大于预设簇数;
while q > k do
    合并距离最近的两个聚类簇 C_{i*} = C_{i*} ∪ C_{j*};
    for j = j* + 1, j = j* + 2, ..., q do
        将聚类簇 C_j 重编号为 C_{j-1}
    删除距离矩阵 M 的第 j* 行与第 j* 列;
    # 重新计算距离矩阵;
    for j = 1, 2, ..., q - 1 do
        M_{i*,j} = d(C_{i*}, C_j);
        M_{j,i*} = M_{i*,j}
    q=q-1
```

默认采用“欧式距离”进行度量样本距离，采用“离差平方和”（ward linkage）作为簇距离度量函数。该算法执行结果见下页：

3.2.3 密度聚类：DBSCAN

DBSCAN 算法从样本密度的角度来考察样本之间的可连接性，要求聚类空间中的以 eps 为半径的邻域内所包含对象的数目不小于某一给定阈值 $min\_samples$，并基于可连接样本不断扩展生长聚类簇以获得最终的聚类结果。DBSCAN 不需要预先输入要划分的聚类个数，但是对 eps、$min\_samples$ 参数敏感。记特征维度数目 $N = 3$、$K = 2N-1$，按照以下经验值确定超参数 [15, 16]：$min\_samples = 2N = 6$，将数据集各点与 K-最近邻算法分类标签的距离排序，观察图 8 拐点 y 坐标确定 $eps = 2$。

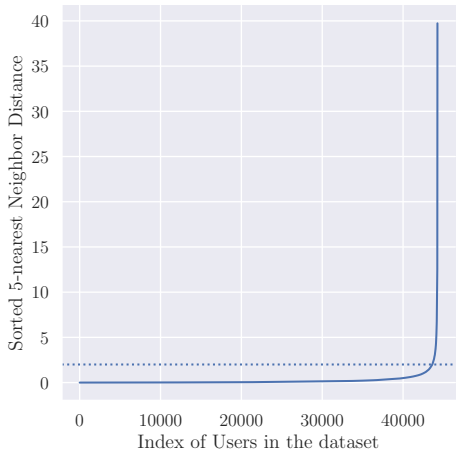


图 10　数据集各点 6-最近邻距离（排序）

运行结果：聚类数量为 7，噪点用户 539 名。

3.3 算法比较与用户画像

共性。

四、 问题二：未来使用情况预测

4.1 问题分析与流程思路

XGBoost、ResNet、

调整参数

ARIMA

固定窗

滑动窗

验证

4.2 你好

要使用 LaTeX 来完成建模论文，首先要确保正确安装一个 LaTeX 的发行版本。

- Mac 下可以使用 MacTeX
- Linux 下可以使用 TeXLive；
- windows 下可以使用 TeXLive 或者 MikTeX；

具体安装可以参考 Install-LaTeX-Guide-zh-cn 或者其它靠谱的文章。另外可以安装一个易用的编辑器，例如 TeXstudio 。

使用该模板前，请阅读模板的使用说明文档。下面给出模板使用的大概样式。

```
\documentclass{cumcmthesis}
%\documentclass[withoutpreface,bwprint]{cumcmthesis} %去掉封面与编号页

\title{论文题目}
\tihao{A}          % 题号
\baominghao{4321} % 报名号
\schoolname{你的大学}
\membera{成员A}
\memberb{成员B}
\memberc{成员C}
\supervisor{指导老师}
\yearinput{2017}  % 年
\monthinput{08}   % 月
\dayinput{22}     % 日

\begin{document}
\maketitle
\begin{abstract}
    摘要的具体内容。
    \keywords{关键词1\quad 关键词2\quad 关键词3}
\end{abstract}
\tableofcontents
\section{问题重述}
\subsection{问题的提出}
\section{模型的假设}
\section{符号说明}
\begin{center}
    \begin{tabular}{cc}
```

```
        \hline
        \makebox[0.3\textwidth][c]{符号} & \makebox[0.4\textwidth][c]{意义}
            \\ \hline
        D                              & 木条宽度（cm）                    \\
            \hline
    \end{tabular}
\end{center}
\section{问题分析}
\section{总结}
\begin{thebibliography}{9}%宽度9
    \bibitem{bib:one} ....
\end{thebibliography}
\begin{appendices}
    附录的内容。
\end{appendices}
\end{document}
```

根据要求，电子版论文提交时需去掉封面和编号页。可以加上 `withoutpreface` 选项来实现，即：

```
\documentclass[withoutpreface]{cumcmthesis}
```

这样就能实现了。打印的时候有超链接的地方不需要彩色，可以加上 `bwprint` 选项。

另外目录也是不需要的，将 `\tableofcontents` 注释或删除，目录就不会出现了。

团队的信息填入指定的位置，并且确保信息的正确性，以免因此白忙一场。

编译记得使用 `xelatex`，而不是用 `pdflatex`。在命令行编译的可以按如下方式编译：

```
xelatex example
```

或者使用 `latexmk` 来编译，更推荐这种方式。

```
latexmk -xelatex example
```

下面给出写作与排版上的一些建议。

五、 图片

建模中不可避免要插入图片。图片可以分为矢量图与位图。位图推荐使用 `jpg,png` 这两种格式，避免使用 `bmp` 这类图片，容易出现图片插入失败这样情况的发生。矢量图一般有 `pdf,eps` ，推荐使用 `pdf` 格式的图片，尽量不要使用 `eps` 图片，理由相同。

注意图片的命名，避免使用中文来命名图片，可以用英文与数字的组合来命名图片。避免使用`1,2,3` 这样顺序的图片命名方式。图片多了，自己都不清楚那张图是什么了，命名尽量让它有意义。下面是一个插图的示例代码。

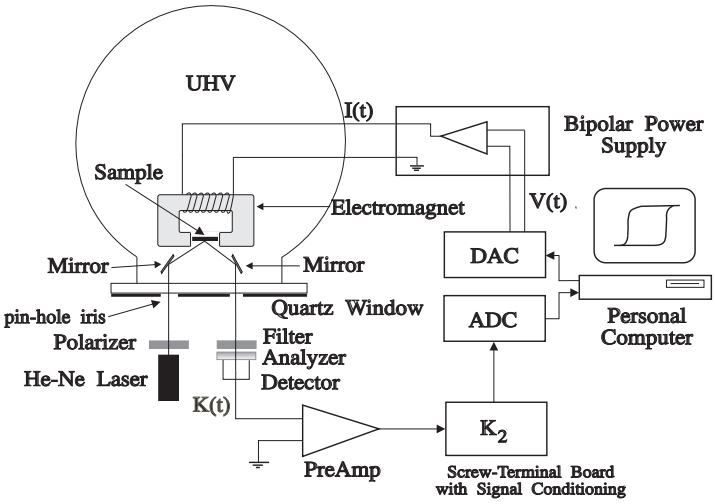


图 11　电路图

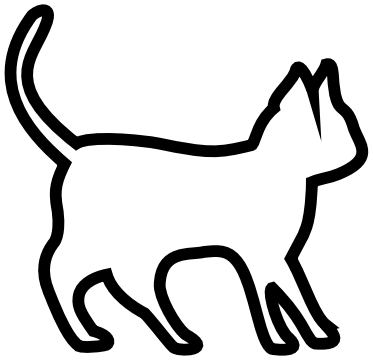
注意 `figure` 环境是一个浮动体环境，图片的最终位置可能会跑动。`[!h]` 中的 `h` 是 here 的意思，`!` 表示忽略一些浮动体的严格规则。另外里面还可以加上 `btp` 选项，它们分别是 bottom, top, page 的意思。只要这几个参数在花括号里面，作用是不分先后顺序的。page 在这里表示浮动页。

`\label{fig:circuit-diagram}` 是一个标签，供交叉引用使用的。例如引用图片 `\cref{fig:circuit-diagram}` 的实际效果是图 11。图片是自动编号的，比起手动编号，它更加高效。`\cref{label}` 由 `cleveref` 宏包提供，比普通的 `\ref{label}` 更加自动化。`label` 要确保唯一，命名方式推荐用图片的命名方式。

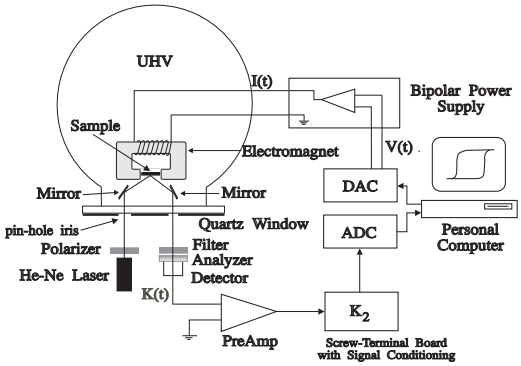
图片并排的需求解决方式多种多样，下面用 `minipage` 环境来展示一个简单的例子。注意，以下例子用到了 `subcaption` 命令，需要加载 subcaption 宏包。

这相当于整体是一张大图片，大图片引用是**??**，子图引用别分是**??**、**??**、**??**。

如果原本两张图片的高度不同，但是希望它们缩放后等高的排在同一行，参考这个例子：



(a) 一只猫



(b) 电路图

图 12　多图并排示例

六、 绘制普通三线表格

表格应具有三线表格式，因此常用 booktabs 宏包，其标准格式如表 7 所示。

表 7　标准三线表格

D(in)	P_u(lbs)	u_u(in)	β	G_f(psi.in)
5	269.8	0.000674	1.79	0.04089
10	421.0	0.001035	3.59	0.04089
20	640.2	0.001565	7.18	0.04089

其绘制表格的代码及其说明如下。

```
\begin{table}[!htbp]
   \caption[标签名]{中文标题}
   \begin{tabular}{cc...c}
      \toprule[1.5pt]
      表头第1个格  & 表头第2个格  & ... & 表头第n个格          \\
      \midrule[1pt]
      表中数据(1,1) & 表中数据(1,2) & ... & 表中数据(1,n)     \\
      表中数据(2,1) & 表中数据(2,2) & ... & 表中数据(2,n)     \\
      ..................................................... \\
      表中数据(m,1) & 表中数据(m,2) & ... & 表中数据(m,n)     \\
      \bottomrule[1.5pt]
   \end{tabular}
```

```
\end{table}
```

table 环境是一个将表格嵌入文本的浮动环境。tabular 环境的必选参数由每列对应一个格式字符所组成：c 表示居中，l 表示左对齐，r 表示右对齐，其总个数应与表的列数相同。此外，`@{文本}`可以出现在任意两个上述的列格式之间，其中的文本将被插入每一行的同一位置。表格的各行以\\分隔，同一行的各列则以 & 分隔。`\toprule`、`\midrule` 和 `\bottomrule` 三个命令是由 booktabs 宏包提供的，其中 `\toprule` 和 `\bottomrule` 分别用来绘制表格的第一条（表格最顶部）和第三条（表格最底部）水平线，`\midrule` 用来绘制第二条（表头之下）水平线，且第一条和第三条水平线的线宽为 1.5pt ，第二条水平线的线宽为 1pt 。引用方法与图片的相同。

七、 公式

数学建模必然涉及不少数学公式的使用。下面简单介绍一个可能用得上的数学环境。

首先是行内公式，例如 θ 是角度。行内公式使用 `$ $` 包裹。

行间公式不需要编号的可以使用 `\[ \]` 包裹，例如

$$E = mc^2$$

其中 E 是能量，m 是质量，c 是光速。

如果希望某个公式带编号，并且在后文中引用可以参考下面的写法：

$$E = mc^2 \tag{1}$$

式 (1) 是质能方程。

多行公式有时候希望能够在特定的位置对齐，以下是其中一种处理方法。

$$\begin{aligned} P &= UI && (2)\\ &= I^2R && (3) \end{aligned}$$

`&` 是对齐的位置，`&` 可以有多个，但是每行的个数要相同。

矩阵的输入也不难。

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{pmatrix}$$

分段函数这些可以用 case 环境，但是它要放在数学环境里面。

$$f(x)=\begin{cases}0 & x\text{为无理数,}\\ 1 & x\text{为有理数.}\end{cases}$$

在数学环境里面，字体用的是数学字体，一般与正文字体不同。假如要公式里面有个别文字，则需要把这部分放在 text 环境里面，即 \text{文本环境} 。

公式中个别需要加粗的字母可以用 $\bm{math symbol}$ 。如 $\alpha a\boldsymbol{\alpha a}$ 。

以上仅简单介绍了基础的使用，对于更复杂的需求，可以阅读相关的宏包手册，如 amsmath。

希腊字母这些如果不熟悉，可以去查找符号文件 symbols-a4.pdf ，也可以去 detexify 网站手写识别。另外还有数学公式识别软件 mathpix 。

下面简单介绍一下定理、证明等环境的使用。

除了 definition 环境，还可以使用 theorem 、lemma、corollary、assumption、conjecture、axiom、principle、problem、example、proof、solution 这些环境，根据论文的实际需求合理使用。

定理 1 这是一个定理。

由定理 1我们知道了定理环境的使用。

引理 1 这是一个引理。

由引理 1我们知道了引理环境的使用。

推论 1 这是一个推论。

由推论 1我们知道了推论环境的使用。

假设 1 这是一个假设。

由假设 1我们知道了假设环境的使用。

猜想 1 这是一个猜想。

由猜想 1我们知道了猜想环境的使用。

公理 1 这是一个公理。

由公理 1我们知道了公理环境的使用。

定律 1 这是一个定律。

由定律 1我们知道了定律环境的使用。

问题 1 这是一个问题。

由问题 1我们知道了问题环境的使用。

例 1 这是一个例子。

由例 1我们知道了例子环境的使用。

证明 1 这是一个证明。

由证明 1我们知道了证明环境的使用。

解 1 这是一个解。

由解 1我们知道了解环境的使用。

八、 参考文献与引用

参考文献对于一篇正式的论文来说是必不可少的，在建模中重要的参考文献当然应该列出。LaTeX 在这方面的功能也是十分强大的，下面进介绍一个比较简单的参考文献制作方法。有兴趣的可以学习 `bibtex` 或 `biblatex` 的使用。

LaTeX 的入门书籍可以看《LaTeX 入门》[**?**]。这是一个简单的引用，用 `\cite{bibkey}` 来完成。要引用成功，当然要维护好 bibitem 了。下面是个简单的例子。

参考文献

[1] 周志华. 机器学习 [M]. 北京. 清华大学出版社. 2016. 197-219

[2] 何宏. 高维数据的聚类分析 [M]. 上海. 上海交通大学出版社. 2022. 1-16

[3] 陈志泊, 韩慧, 王建新, 孙俏, 聂耿青. 数据仓库与数据挖掘 [M]. 北京. 清华大学出版社. 2009

[4] 常乐. 基于用户行为分析的用户画像系统设计与实现 [D]. 北京邮电大学. 2020

[5] 陈纯, 龙瀛, 黄贵恺. 屏幕使用时间与步行活动关系的探索性研究 [J]. 景观设计学 (中英文). 2021. 9(04):68-81

[6] 成雪, 于冬梅, 赵丽云等. 2016—2017 年中国各省中小学生电子屏幕使用现状 [J]. 卫生研究. 2023,52(03):382-387

[7] 武慧娟, 赵天慧, 孙鸿飞等. 基于支付意愿的数字阅读用户画像聚类研究 [J]. 情报科学. 2022. 40(05)

[8] 侯金凤. 移动互联网下手机用户使用行为特征的研究 [J]. 电脑知识与技术. 2016,12(07)

[9] 韦磊. 基于移动终端数据的用户画像模型研究 [D]. 江苏科技大学. 2021

[10] Garg, R., & Barpanda, S. Machine Learning Algorithms for Time Series Analysis and Forecasting[Z/OL]. arXiv preprint arXiv:2211.14387. https://arxiv.org/abs/2211.14387

[11] T. Caliński, J Harabasz. A dendrite method for cluster analysis[J]. Communications in Statistics. 1974. 3:1, 1-27

[12] Davies D L , Bouldin D W. A Cluster Separation Measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979. PAMI-1(2):224-227

[13] Peter R J . Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational & Applied Mathematics. 1987. 20

[14] Swami A , Jain R. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research. 2013, 12(10):2825-2830

[15] Sander J , Ester M , Kriegel H P ,et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications[J]. Data Mining & Knowledge Discovery. 1998. 2(2):169-194

[16] Schubert E , Sander J , Ester M ,et al. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN[J]. ACM Transactions on Database Systems, 2017, 42(3):1-21

附录 A　环境依赖与使用说明

表 8　依赖罗列

模板中已经加载的宏包				
amsbsy	amsfonts	amsgen	amsmath	amsopn

以上宏包都已经加载过了，不要重复加载它们。

附录 B　2.1 源代码

附录 C　2.2 源代码

附录 D　2.3 源代码

```
kk=2;[mdd,ndd]=size(dd);
while ~isempty(V)
[tmpd,j]=min(W(i,V));tmpj=V(j);
for k=2:ndd
[tmp1,jj]=min(dd(1,k)+W(dd(2,k),V));
tmp2=V(jj);tt(k-1,:)=[tmp1,tmp2,jj];
end
tmp=[tmpd,tmpj,j;tt];[tmp3,tmp4]=min(tmp(:,1));
if tmp3==tmpd, ss(1:2,kk)=[i;tmp(tmp4,2)];
else,tmp5=find(ss(:,tmp4)~=0);tmp6=length(tmp5);
if dd(2,tmp4)==ss(tmp6,tmp4)
ss(1:tmp6+1,kk)=[ss(tmp5,tmp4);tmp(tmp4,2)];
else, ss(1:3,kk)=[i;dd(2,tmp4);tmp(tmp4,2)];
end;end
dd=[dd,[tmp3;tmp(tmp4,2)]];V(tmp(tmp4,3))=[];
[mdd,ndd]=size(dd);kk=kk+1;
end; S=ss; D=dd(1,:);
```

附录 E　规划解决程序–lingo 源代码

```
kk=2;
[mdd,ndd]=size(dd);
while ~isempty(V)
    [tmpd,j]=min(W(i,V));tmpj=V(j);
for k=2:ndd
    [tmp1,jj]=min(dd(1,k)+W(dd(2,k),V));
    tmp2=V(jj);tt(k-1,:)=[tmp1,tmp2,jj];
end
    tmp=[tmpd,tmpj,j;tt];[tmp3,tmp4]=min(tmp(:,1));
if tmp3==tmpd, ss(1:2,kk)=[i;tmp(tmp4,2)];
else,tmp5=find(ss(:,tmp4)~=0);tmp6=length(tmp5);
if dd(2,tmp4)==ss(tmp6,tmp4)
    ss(1:tmp6+1,kk)=[ss(tmp5,tmp4);tmp(tmp4,2)];
else, ss(1:3,kk)=[i;dd(2,tmp4);tmp(tmp4,2)];
end;
end
    dd=[dd,[tmp3;tmp(tmp4,2)]];V(tmp(tmp4,3))=[];
    [mdd,ndd]=size(dd);
    kk=kk+1;
end;
S=ss;
D=dd(1,:);
```