Operations Research and Fuzziology 运筹与模糊学, 2023, 13(5), 5045-5055
Published Online October 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.135506

基于K-Means聚类的消费者直播购物偏好研究

韩　晨

上海工程技术大学管理学院，上海

收稿日期：2023年8月9日；录用日期：2023年10月9日；发布日期：2023年10月18日

摘　要

本文针对消费者直播购物行为数据展开分析，采用K-means算法进行聚类，将具有相似购买行为的样本聚为一组，得到三种类别的直播购物偏好风格，进一步分析了各个偏好风格的群体特征，并为电商平台实施直播互动提供启示。

关键词

直播，购物偏好，K-Means聚类

Research on Consumers' Preference for Live Shopping Based on K-Means Clustering

Chen Han

School of Management, Shanghai University of Engineering Science, Shanghai

Received: Aug. 9th, 2023; accepted: Oct. 9th, 2023; published: Oct. 18th, 2023

Abstract

Based on the analysis of consumers' live shopping behavior data, this paper uses K-means algorithm for clustering, gathers samples with similar purchasing behaviors into a group, obtains three types of live shopping preference styles, further analyzes the group characteristics of each preference style, and provides inspiration for e-commerce platforms to implement live interaction.

Keywords

Live Streaming, Shopping Preference, K-Means Cluster

文章引用: 韩晨. 基于 K-Means 聚类的消费者直播购物偏好研究[J]. 运筹与模糊学, 2023, 13(5): 5045-5055.
DOI: 10.12677/orf.2023.135506

1. 引言

随着互联网技术的快速发展，直播的互动、社交、娱乐等属性深刻吸引着当下的消费者。他们在直播间可以获得更直观的商品展示和较为优惠的产品价格，尤其流量主播带货时，粉丝会在购物过程中获得更愉悦的体验。直播带货已然重塑了我们的生活方式，并且在逐渐加速改变我们的购物习惯。与传统网络购物渠道相比，直播带货的强互动性和实时反馈等特点，以最大程度吸引着潜在消费者。电商平台大多和流量主播合作进行直播带货，这类主播往往擅于采用销售话术和带货技巧来吸引消费者下单。平台将合作商家的产品交给主播排序展示，可以短时间内促成大量订单。2019 年“双十一”当天，李佳琦直播间粉丝数量为 3683.5 万。在 2020 年双十一期间，抖音合作主播罗永浩累计销售额超 3 亿，快手合作主播辛有志累计销售额超 32 亿。商家则更倾向于培训员工作为自家直播间主播进行带货。在商家直播间里，主播有更充裕的时间展示产品、与消费者即时互动，即刻以专业的知识回答消费者提出的各种疑问，增强消费者对产品的了解和体验感；也可以在展示产品间隙讲述品牌故事，以提升消费者对商家的信任程度。近年来，商家开通直播渠道的比例逐年上升，目前这种直播方式已经成为众多品牌的主要销售场景之一。因此，本文在引入直播渠道的基础上，探讨消费者对直播方式、直播产品、直播时间等直播因素的偏好程度。

与本文相关的研究主要集中在 K-means 算法优化和社交平台消费者行为、直播购物偏好等方面。罗雲潇等[1]以车辆历史运行物理参数为研究对象，首先基于 K-means 聚类优先识别出九种行驶工况，随后利用因子分析对数据降维，并通过 SOM-Kmeans 聚类模型识别出司机的驾驶风格。田俐[2]利用 PCA 算法降维后，采用 K-means 聚类算法获取政务服务数据分类，更加快速地提取出当前热点问题。刘国华[3]采用优化的 K-means 算法，通过计算得出最佳的聚类系数 K，将学生主要的在校行为特征作为聚类的维度，并加上量纲的矩阵系数对学生的行为特征进行聚类分析。凌玉龙[4]等采用马氏距离代替欧氏距离以适用于具体的校园消费数据应用场景，在高密度样本集合中选择相距最远的 k 个样本作为初始聚类中心的改进方法，有效地区分了不同行为特征的学生并很好地刻画学生的消费画像。杨尊琦和张倩楠[5]利用 K-means 聚类将具有相似兴趣的微博用户聚为一组，结合现实情况分析各类别之间的相似性和区别，挖掘用户关注兴趣的隐性信息，并对微博用户推荐兴趣提出建议。易茹[6]在传统 K 均值聚类算法分析的基础上，提出了改进的 K 均值聚类算法来建立数字媒体推荐模型，并应用于微博推荐中。袁海霞和黄丽雯[7]分析平台类型对于互动模式作用机制的调节作用并进行假设检验，为直播营销提供了建议。韩琮师[8]对传统的 K-means 算法进行改进，提出采用 HC-Kmeans 算法对唯品会中消费者的消费行为进行分析，将消费者进行聚类后划分为不同群体并提出针对性的营销策略。但上述研究未结合直播带货对消费者行为进行探讨。Jin 和 Lee [9]针对直播电商网红与消费者形象一致性之间的关系展开研究，并进一步探究其对消费者品牌偏好及购买意愿的影响。Chen 等[10]探讨消费者在直播间购买扶贫产品前后的购物偏好变化情况。Liu 等[11]认为在冲动消费的情形下，主播专业水平对消费者购物偏好存在影响。然而，这些研究并未结合消费者在直播间购物的群体行为数据进行深入分析。

基于此，本文针对消费者在电商平台的购物行为数据，利用 K-means 聚类分析消费者直播购物的偏好行为，探讨多种购物渠道并存的情形下不同背景消费者的直播购物偏好风格，为电商平台实施直播互动提供启示。

2. K-means 聚类概述

在聚类分析中，类别的个数及个体标签本身并不存在，只是根据个体特征的相似性形成合理的聚集，并无正确答案参考，属于无监督学习。在层次聚类中，一旦个体被分入一个族群，它将不可再被归入另一个族群。而 K-means 聚类作为最常用的分割法，可以根据样本之间的相似度分为不同的簇，需要的计算量较少，且更容易理解，尤其适用于大样本。K-means 聚类是由麦奎因 1967 年首次使用，其聚类思想是：假设数据中有 p 个变量参与聚类，并且要聚类为 k 个簇，则需要在 p 个变量组成的 p 维空间中，首先选取 k 个不同的样本作为聚类种子，然后根据每个样本到达这 k 个点的距离的大小，将所有样本分为 k 个簇，在每一个簇中，重新计算出簇的中心(每个特征的平均值)并作为新的种子，再把所有的样本分为 k 类。如此下去，直到种子的位置几乎不发生改变为止。

3. 数据准备

3.1. 数据收集

本研究通过发放调查问卷采集数据。首先，通过筛选观看过直播带货的消费者来确定此次调研对象。在问卷中设置测试题目“您观看直播带货的频率如何”，将选择选项为“从不”的问卷判定无效，不作为调研对象。本文遵循随机抽样原则，采用三种渠道多阶段发放问卷：通过风铃系统在多个社交网站上发布问卷；在商场随机拦截路人发放问卷；采用滚雪球方式请求被调研对象邀请其他符合条件的消费者填写问卷。问卷填写从 2023 年 5 月下旬持续到 2023 年 6 月中旬，总计回收问卷 212 份。为确保问卷数据质量，本文剔除未通过测试题目的无效问卷和答卷时间过短的问卷，最终获得有效问卷 175 份，有效回收率为 82.55%。

3.2. 数据描述与展示

(1) 样本均值向量：对随机样本 $\{Y_1,\cdots,Y_n\}$ (其中 $Y_i=\left(y_{i1},\cdots,y_{ip}\right)'$)，$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i=\left(\overline{y}_1,\cdots,\overline{y}_p\right)'$，其中 $\overline{y}_j=\frac{1}{n}\sum_{i=1}^{n}y_{ij}$。样本均值 $\overline{y}$ 是总体均值 μ 的无偏估计，即 $E\left(\overline{Y}\right)=\mu$。如表 1 所示。

Table 1. Sample mean
表 1. 样本均值

样本均值	
性别	50.17143
年龄段	12.98857
职业	14.76000
地区	14.01143
月收入水平	14.24000
常用电商平台	37.57143
月网购消费水平	16.28000
渠道偏好	15.89143
观看直播频率	23.12571
网店购物方式偏好	33.21143
直播购物方式偏好	33.79429
直播商品偏好	30.09143
直播类型偏好	28.72571
直播时间偏好	30.72571

(2) 欧氏距离与马氏距离：在多元的情形中，对于两个 p 维向量 $Y_1=\left(y_{11},\cdots,y_{1p}\right)'$ 和 $Y_2=\left(y_{21},\cdots,y_{2p}\right)'$ 之间的距离，存在两种测度方式，即欧氏距离和马氏距离。

欧氏距离 $\left\|Y_1-Y_2\right\|=\sqrt{\left(Y_1-Y_2\right)'\left(Y_1-Y_2\right)}=\sqrt{\sum_{j=1}^{p}Y_i\left(y_{1j}-y_{2j}\right)^2}$，但这个方式没有考虑到变量之间的相关性以及不同变量变化的尺度存在差异。如表 2 所示。

Table 2. Euclidean distance
表 2. 欧氏距离

	1	2	3	4
2	70.00714			
3	96.08330	40.80441		
4	110.71134	50.99020	20.07486	
5	110.45361	54.50688	39.94997	42.03570

马氏距离 $d=\sqrt{\left(Y_1-Y_2\right)'S^{-1}\left(Y_1-Y_2\right)}$，其中样本方差 S 能够将所有变量标准化成相同方差，并且消除变量之间的相关性。方差更大的变量对应了更小的权重，而且两个高度相关的变量对统计距离的贡献小于两个相关性相对较低的变量的贡献。事实上，马氏距离是两个经过变换的向量 $S^{-1/2}Y_1$ 和 $S^{-1/2}Y_2$ 之间的欧氏距离。如表 3 所示。

Table 3. Mahalanobis distance
表 3. 马氏距离

	1	2	3	4
2	4.48			
3	6.04	4.73		
4	6.33	5.26	2.56	
5	7.35	5.60	5.38	5.47

(3) 多元散点图：是借助两变量散点图的作图方法，它可以看作一个大的图形方阵，其每一个非主对角元素的位置上是对应行的变量与对应列的变量的散点图，而主对角位置上是各变量名，这样可以清晰地看到所研究的多个变量两两之间的相关关系。两两散点图矩阵用以观察多元数据中数值变量两两之间的关系。以前五个变量为例画出散点图，如图 1 所示。

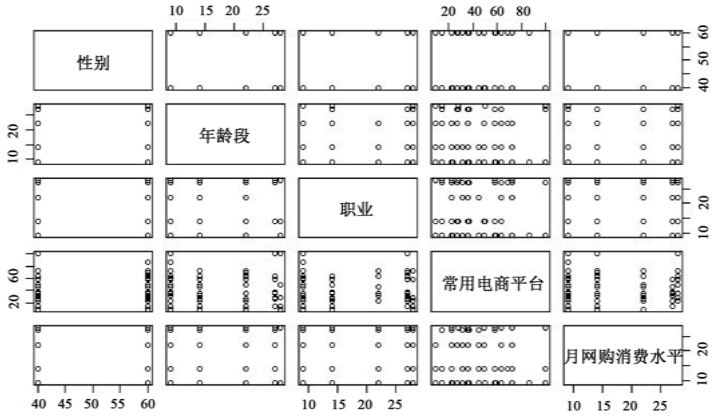

Figure 1. Scatter diagram
图 1. 散点图

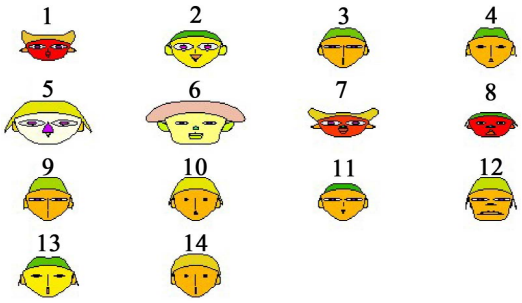

Figure 2. Face graph
图 2. 脸谱图

(4) 脸谱图：由美国统计学家 H.切尔诺夫(H. Chernoff)于 1970 年首先提出，采用脸谱来表达多元数据的样品。将观测的 p 个变量分别用脸的某一部位的形状或大小来表示，一个样品可以画成一张脸谱。取 14 个样本画出脸谱图，如图 2 所示。

(5) 雷达图是目前应用较为广泛的对多元资料进行作图的方法，利用雷达图可以很方便地研究各样本点之间的关系并对样品进行归类。本文选取聚类中心点，如图 3 所示。

(6) 星图：每一个观测对应一张星图。取 14 个样本画出星图，如图 4 所示。

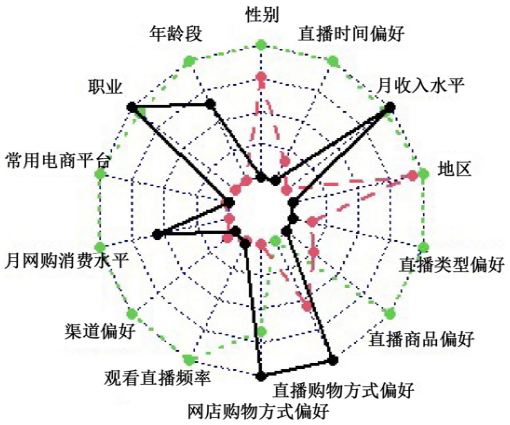

Figure 3. Radar map
图 3. 雷达图

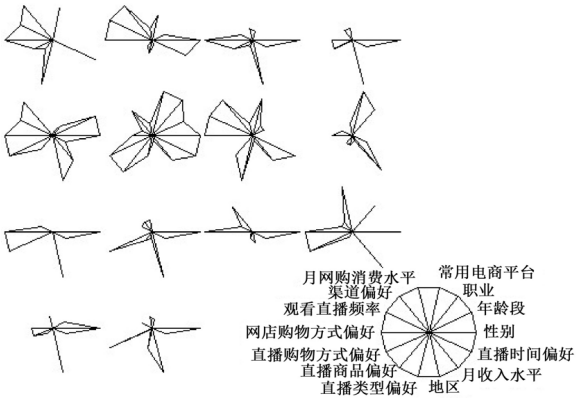

Figure 4. Star chart
图 4. 星图

4. 结果分析

4.1. 主成分分析

主成分分析主要用于构造综合指标来区分目标群体。主成分是原变量的线性组合，通常采用少于原始变量数的主成分来描述尽可能多的数据差异，特别是当原始变量维度很高时，可达到降维目的。

首先，判断保留的主成分个数，输出如图 5 所示。常用的判断方法有三种。常用的 cattell 碎石检验，即绘制主成分数与特征值相关的碎石图，图形变化最大处之上的主成分都可以保留。Kaiser-harris 准则要求保留特征值大于 1 的主成分。平行分析则是模拟与初始矩阵相同大小的随机数据矩阵，若基于正式数据的某个特征值大于一组随机数据矩阵相应的平均特征值，则保留该主成分。此时，结合三种方法，保留 5 个主成分即可。在此基础上，主成分分析可视化输出如图 6 所示。

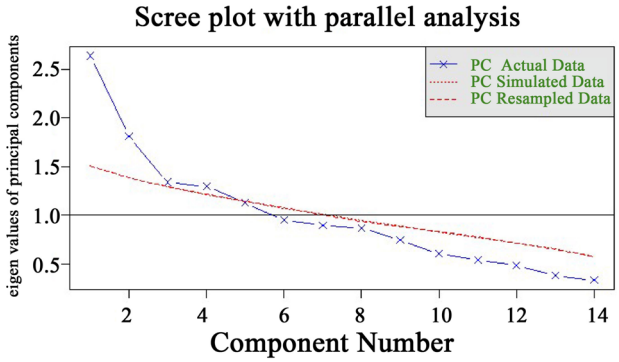

Figure 5. The number of principal components discriminant diagram
图 5. 主成分个数判别图

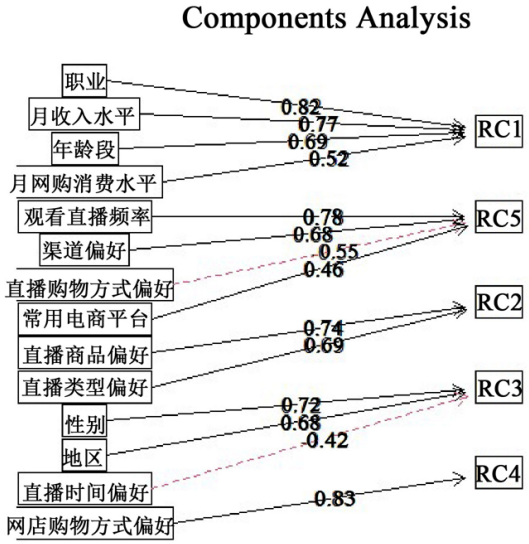

Figure 6. Visualization of principal component analysis
图 6. 主成分分析可视化

然而，从方差贡献率来看，保留 5 个主成分并不合适。各主成分的方差贡献率如图 7 所示。进行主成分分析的目的之一是减少变量的个数，通常取主成分的个数使得累积贡献率达到 85%以上为宜。据此，应该保留 10 个主成分。

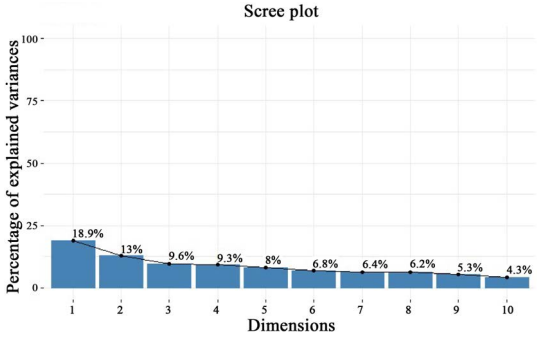

Figure 7. Principal component variance contribution rate
图 7. 主成分方差贡献率

事实上，图中所得主成分解释原始变量总差异的效果并不理想。以第一主成分为例，其方差贡献率为 18.9%，是保留的特征根占所有特征根的和的比值，由此可见第一主成分解释原始变量总差异的效果不好。第二个主成分的方差贡献率为 13%，这个相对第一主成分贡献率更低。主成分分析的关键在于能否对主成分赋予新的意义，并给出合理的解释，这个解释应根据主成分的计算结果结合定性分析来进行。由于分析效果不理想，此处不再赘述。

4.2. 因子分析

因子分析模型是主成分分析的推广，同样利用降维的思想。相比主成分分析，因子分析更倾向于描述原始变量之间的相关关系，因此，因子分析的出发点是原始变量的相关矩阵。因子分析旨在用一个变量(公因子)代替原始高度相关的某几个变量，原变量表示为因子的线性组合，其主要目的是找出不可观测的潜在变量作为公共因子，并解释公共因子含义来探讨数据内部结构。假设可观测随机向量 $Y=\left(y_1,\cdots,y_p\right)'$ 的均值为 μ。假定 Y 线性依赖于 m 个不可观测公共因子 $f=\left(F_1,\cdots,F_m\right)'$ 和 p 个不可观测的特殊因子 $\varepsilon=\left(\varepsilon_1,\cdots,\varepsilon_p\right)'$，通常 $m<p$：

$$y_1-\mu_1=l_{11}F_1+l_{12}F_2+\cdots+l_{1m}F_m+\varepsilon_1 \tag{1}$$

$$y_2-\mu_2=l_{21}F_1+l_{22}F_2+\cdots+l_{2m}F_m+\varepsilon_2 \tag{2}$$

$$\cdots$$

$$y_p-\mu_p=l_{p1}F_1+l_{p2}F_2+\cdots+l_{pm}F_m+\varepsilon_p \tag{p}$$

系数 l_{jk} 称为第 j 个变量在第 k 个因子上的载荷，体现了该公共因子对此变量的解释力。

首先，判断提取的公共因子数，输出如图 8 所示。此时，提取 4 个公共因子最合适。因子分析可视化输出如图 9 所示。

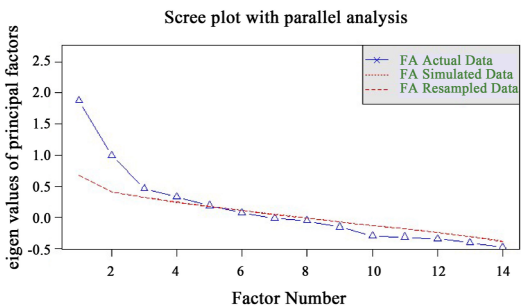

Figure 8. The number of common factors discriminant diagram
图 8. 公共因子个数判别图

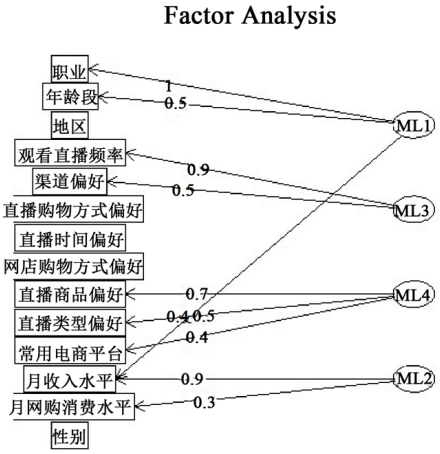

Figure 9. Visualization of factor analysis
图 9. 因子分析可视化

根据因子载荷矩阵，职业、年龄段在第一因子上载荷较大；月收入水平和月网购消费水平在第二因子上载荷较大；观看直播频率和渠道偏好在第三因子上载荷较大；直播商品偏好、直播类型偏好和常用电商平台在第四因子上载荷较大。另一种可视化输出如图 10 所示。

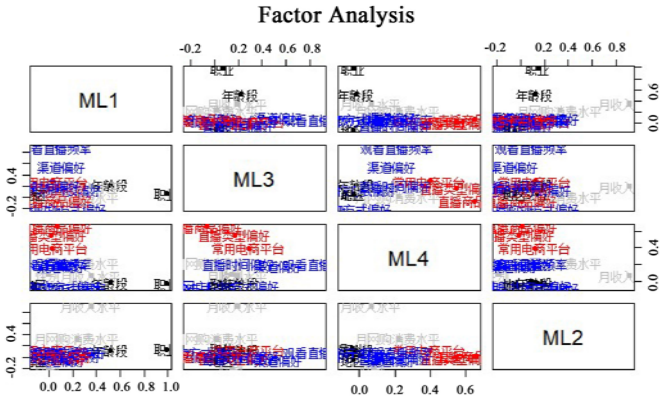

Figure 10. Visualization of factor analysis
图 10. 因子分析可视化

可以看出，第一因子主要由工作性质和年龄阶段决定，能够反映出消费者的社会背景。第二因子主要由收入水平消费水平决定，则说明了消费者的资金能力。第三因子主要由观看直播频率和购物渠道的偏好决定，反映消费者的时间充裕度。第四因子主要由常用购物平台和消费者偏好的直播风格、直播商品决定，反映消费者的直播关注点。综上，称它们为消费者背景因子、资金因子、时间因子、直播偏好因子，在此基础上进行聚类分析。

4.3. K-means 聚类

本文采用三种判别方法来确定 k 值。首先执行肘部判别法，输出如图 11 所示，展示了在不同的K值下，聚类结果的组内平方和的变化情况。曲线的形状类似于人的手肘，且随着聚类个数的增加，组内平方和在减少。但由于曲线的变化趋势未出现明显的改变，难以确定聚类数目 k 值，因此继续尝试其他判别方法。

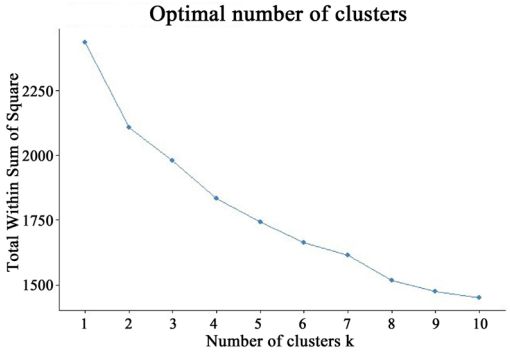

Figure 11. Elbow discrimination diagram
图 11. 肘部判别图

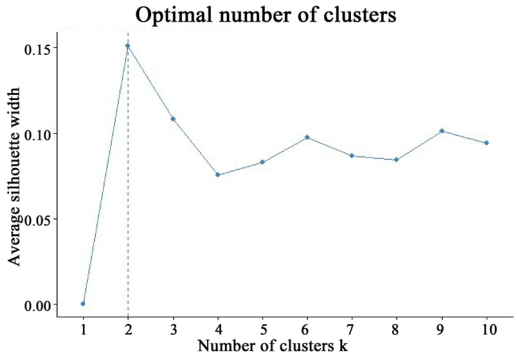

Figure 12. Contour discrimination diagram
图 12. 轮廓判别图

其次，执行轮廓判别法，判别结果为 $k=2$ 或 3，输出如图 12 所示。

随后又执行了差距统计法进行判别，判别结果为 $k=2$，3，4。输出如图 13 所示。

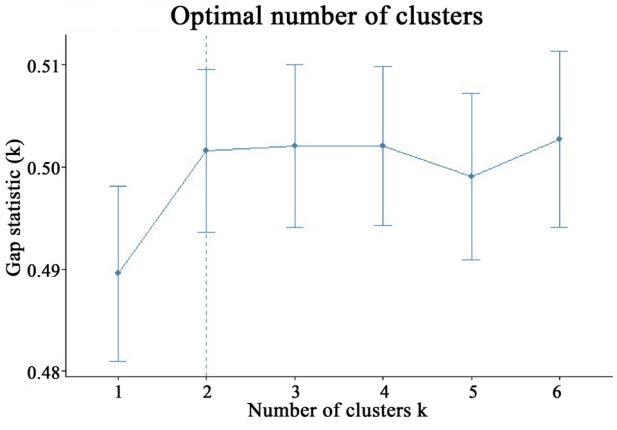

Figure 13. Difference statistical discriminant diagram
图 13. 差异统计判别图

最后，取 $k=3$，使用 kmeans ( )函数将数据聚为 3 类，聚类可视化输出如图 14 所示。从程序的输出结果可以发现，3 个簇的样本数目分别为 29、90、56，并且程序还输出了每个簇的聚类中心(cluster means)以及每个样本的所属类别。分类中心点统计特征数据如表 4 和表 5 所示，表中数据为权值赋值结果。主

要数据特征表格中可以看出，类型 1 从年龄特征、职业特征、直播时间偏好、消费与收入水平、直播商品偏好与类型偏好、常用电商购物平台偏好等维度和其余两者区别较明显；类型 2 从年龄特征、直播时间偏好、消费与收入水平、直播商品偏好与类型偏好、常用电商购物平台偏好以及地区分布等特征和其余两者区别较为明显；类型 3 从年龄特征、直播时间偏好、消费与收入水平、直播商品偏好与类型偏好、渠道偏好、常用电商购物平台偏好等特征和其余两者区别较为明显。

Table 4. Clustering center characteristics (a)
表 4. 聚类中心特征(a)

聚类类别	性别	年龄段	职业	直播时间偏好	月网购消费水平	月收入水平	观看直播频率
1	50.89	9.79	9.39	29.61	14.19	9.56	22.47
2	48.21	15.3	20.75	27.66	17.55	19.36	22.61
3	51.72	18.45	19.86	40.1	20.31	18.9	26.17

Table 5. Clustering center characteristics (b)
表 5. 聚类中心特征(b)

聚类类别	网店购物方式偏好	直播购物方式偏好	直播商品偏好	直播类型偏好	地区	渠道偏好	常用电商平台
1	32.72	33.66	29.33	26.82	14.38	14.79	34.14
2	33.86	34.82	24.73	24.18	13.18	14.05	33.55
3	33.48	32.24	42.79	43.41	14.48	22.86	55.97

将分类结果反标到数据表格表 4 和表 5 中，针对反标的分类标签进行进一步分析。第一类的样本数目为 29。从背景来看，该类消费者大多是大龄已工作者和年轻在读学生，从事个体经营和无工作的消费者也大都在此类中。他们选择多个时段观看直播，时间较为充裕，并且有足够的耐心。但这类消费者往往收入水平不高，在购物时倾向于货比三家，因此他们常用的电商平台种类较多，也更倾向于留意头部主播的待售商品预告以及时在直播间抢购低价日用品，或在商家直播间开展优惠活动时前去“薅羊毛”。同时也存在相当一部分消费者，被主播的带货风格所吸引，将观看直播作为日常娱乐消遣，仅在对产品感兴趣时下单。故称该类群体为“闲多钱少型”消费者。

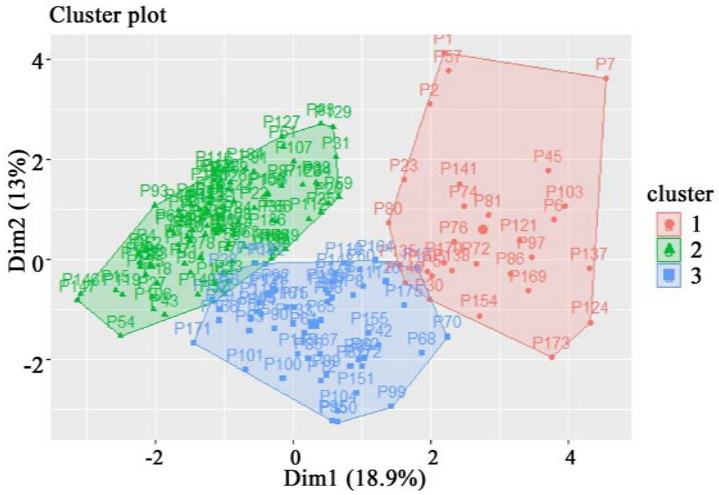

Figure 14. Visualization ofcluster analysis
图 14. 聚类分析可视化

第二类的样本数目为90，该类消费者多为来自西南、华东、华北地区的上班族，其中代表城市有重庆、上海、北京等，均为一线城市。该类消费者往往生活节奏快，只有在晚上才有少量时间观看直播，但却是传统网购的主力军。他们的日常生活用品多从常用品牌的旗舰店购买，对价格敏感度较低。受益于大都市的便捷高效，这类消费者的服饰等时尚用品多于当季从品牌专柜购买，无需通过直播了解款式细节。故称这类消费者为“快节奏型”消费者。

第三类的样本数目为56，该类消费者主要是供职于企业的年轻人，在电商平台上的月消费水平属于此次被调研对象中的中等偏上水平。他们在传统网商渠道购物时多选平台自营店，对平台的信任程度要高于对商家的信任程度；但在直播渠道购物时，他们更倾向于商家直播，原因在于想要在头部主播的直播间买到目标商品需要付出更多等待时间，但这类消费者无法在此投入大量精力。此外，他们经常利用碎片时间进行购物，如不定时访问销售同类产品的店铺、刷短视频时被主播的服饰展示吸引进直播间即兴下单等，且观看直播的时间集中于晚饭时间。故将这类群体归为“钱多闲少型”消费者。

5. 结论

本文基于问卷数据研究消费者在直播间购物的偏好，通过K-means聚类对被调研消费者的购物行为进行识别，最终将其划分为：闲多钱少型、钱多闲少型和快节奏型三种直播购物偏好风格，对调整电商平台的直播产品推送和商家直播产品种草等活动具有实践意义。对于闲多钱少型消费者，平台应多向其推送互动性强、带货风格多样的主播，同时以物美价廉、方便实用为噱头引起消费者对产品的兴趣。对于钱多闲少型消费者，商家应在社交媒体上加强产品直播前的种草频率和直播时配套产品同时出镜的频率，以增进此类消费者对店内配套产品的了解从而提高销售额。对于快节奏型消费者，则应注重产品质量，商家可以将直播间销售成绩亮眼、反馈优秀的产品推送给此类消费者，通过优化产品和服务质量来增强客户粘性。

参考文献

[1] 罗雲潇, 张海瑞, 张振京, 宋亚栋, 屈亚祥. 基于SOM-Kmeans算法的司机驾驶风格研究[J]. 时代汽车, 2023(8): 189-192.

[2] 田俐. 基于Kmeans的12345问题热点分析[J]. 电子技术与软件工程, 2023(7): 244-247.

[3] 刘国华. 基于Kmeans算法的学生行为分析系统的设计与实现[D]: [硕士学位论文]. 石家庄: 河北科技大学, 2014.

[4] 凌玉龙, 张晓, 李霞, 张勇. 改进Kmeans算法在学生消费画像中的应用[J]. 计算机技术与发展, 2021, 31(10): 122-127.

[5] 杨尊琦, 张倩楠. 基于K-means算法的微博用户推荐功能研究[J]. 情报杂志, 2013, 32(8): 142-144+131.

[6] 易茹. 基于K均值聚类算法的数字媒体推荐方法研究[J]. 长春工程学院学报(自然科学版), 2020, 21(4): 99-102.

[7] 袁海霞, 黄丽雯. 电商直播互动模式对消费者购买意愿的影响研究[J]. 哈尔滨商业大学学报(社会科学版), 2022(6): 19-30.

[8] 韩琮师. K-means聚类算法优化及其在电商平台精准营销中的应用研究[D]: [硕士学位论文]. 青岛: 山东科技大学, 2020.

[9] Jin, C. and Lee, S. (2022) The Effect of Self-Image Congruence with Live Commerce Influencer on Consumer Fanship and Brand Preference and Purchase Intention. *Journal of Product Research*, **40**, 9-20.

[10] Chen, T., Tang, S.D., Shao, Z.J., *et al*. (2023) Doing Well by Doing Good: The Effect of Purchasing Poverty-Alleviation Products on Consumers' Subsequent Product Preference in Live Streaming Shopping. *Computers in Human Behavior*, **144**, Article 107753. https://doi.org/10.1016/j.chb.2023.107753

[11] Liu, X.H., Wang, D.H., Gu, M., *et al*. (2023) Research on the Influence Mechanism of Anchors' Professionalism on Consumers' Impulse Buying Intention in the Livestream Shopping Scenario. *Enterprise Information Systems*, **17**, 920-940. https://doi.org/10.1080/17517575.2022.2065457