互联网医疗平台问诊数据的多维行为挖掘与预测建模

赵曾援

摘要

本研究以“好大夫在线”互联网医疗平台的真实问诊数据为基础，围绕用户在病情表达与医患交互过程中的内容模式与行为特征展开综合性建模分析。通过构建涵盖文本结构分析、用户行为聚类与监督学习预测的多层次分析框架，揭示了互联网问诊用户的潜在分群结构和内容主题特征及其可预测性。首先，研究对问诊文本进行系统预处理，提高语义纯度，并分别对“病情描述”与“医患对话”进行 LDA 主题建模。提取出 8 个稳定主题，覆盖诊断、检查、用药、术后恢复等核心内容领域。随后，结合问诊记录中的“病情字数”“对话字数”“总轮数”三项结构性指标，使用 KMeans 算法实现用户分群，辅以 UMAP 与 PCA 降维可视化。聚类结果呈现出一定群体结构，轮廓系数为 0.368，说明用户在活跃度维度上具有较明显的差异性。非参数检验结果显示三类用户在字数与轮数上的差异均具有显著性（$P < 0.001$）。并且“超长对话率”在不同主题下也存在统计学显著差异（$P = 0.011332$），提示用户在部分主题内容上表现出更高的沟通意愿。在此基础上，研究构建了多个机器学习分类模型，验证仅凭结构性特征即可实现对用户群体的高精度预测。最终经 GridSearchCV 优化的 LightGBM 模型在测试集上准确率达到 97.72%。为提升模型透明度，研究引入 SHAP 方法解释特征贡献，构建 SHAP 力图直观展现模型判定机制，增强模型可解释性。总体而言，本研究提出了一套面向互联网医疗用户的内容建模与分群预测一体化框架。验证了用户在线问诊行为具有明确的结构性、可建模性与可预测性，为平台开展精细化管理与个性化医疗服务提供了理论基础与技术支持。

关键词：互联网医疗；用户行为分析；主题建模；聚类分析；降维；机器学习分类

目录

引言

随着“互联网+医疗健康”政策的持续推进，中国医疗服务体系正经历深刻的数字化转型[1]。国家层面相继出台《国务院办公厅关于促进“互联网+医疗健康”发展的意见》等政策文件，鼓励医疗机构利用互联网技术拓展服务模式，提升医疗服务的可及性和效率。尤其是在全国各行业数字化发展背景下，线上问诊、远程医疗等服务形式迅速发展，成为医疗服务的重要补充。然而，尽管数字医疗服务日益普及，用户在使用过程中的行为特征、表达方式及其背后的需求差异仍缺乏系统性的研究。本研究基于大规模在线问诊数据，采用主题建模、聚类分析、统计检验和机器学习等方法，深入挖掘用户行为特征与表达模式，旨在为数字医疗服务的个性化优化提供数据支持和理论依据。

1 文献综述

中国的数字医疗行业在国家政策支持下发展迅速[2]，极大地促进了远程医疗、线上问诊、电子处方等服务模式的普及[3]。已有研究普遍关注了互联网医疗的服务模式创新[4]与医疗资源配置优化[5]，并指出数字医疗在提升基层诊疗能力和优化患者就医路径方面发挥了重要作用[6]。

在用户行为研究领域，随着医疗平台用户规模的增长，用户行为特征的量化分析逐渐成为研究热点。部分学者采用点击流数据分析了用户在医疗平台中的搜索与咨询行为，揭示了用户需求的多样性与动态变化[7]。曹博林等研究了患者在线咨询行为与就诊决策之间的关联，指出在线咨询可以显著影响患者线下就诊倾向[8]。此外，吕艳华等从用户停留时间和交互频次等指标出发，探讨了用户黏性与医生响应特征之间的关系[9]，为平台用户管理提供了参考。

在医疗文本处理方面，自然语言处理技术被广泛应用于患者病情描述和医患对话数据的分析。主题模型（如 LDA）被广泛用于抽取医疗文本中的隐含主题，例如 Selvi 等通过对健康问答平台数据建模，发现不同疾病类别用户在提问内容上存在显著差异[10]。近年来，深度学习方法如 BERT 及其变体被引入医疗领域，大幅提升了情感分析与意图识别的精度[11]，增强了模型对医疗场景复杂语义的理解能力。

关于用户分群与聚类分析的研究，已有学者基于用户画像特征、行为序列数据等采用无监督学习方法进行群体划分。Khanmohammadi 等基于 KMeans 算法对在线医疗用户进行聚类分析，发现用户在健康意识、求医频率等维度上存在天然分层[12]。尹相森等进一步指出，不同用户群体在信息需求与咨询路径选择上表现出系统性差异[13]，提示平台可依据用户特征实施差异化运营策略。

机器学习在数字医疗领域的应用也日益丰富，特别是在疾病预测、医疗服务推荐等方向。文献表明，基于树模型（如 XGBoost、LightGBM）的分类模型在医疗预测任务中表现出色[14]，且具有良好的可扩展性与解释性。为缓解“黑盒模型”带来的不透明问题，模型可解释性方法如 SHAP（SHapley Additive exPlanations）被广泛引入，能够细粒度地量化每个输入特征对预测结果的贡献，

提升了模型在医疗领域应用的可信度[15]。特别是在患者分群预测、健康风险评估等任务中，SHAP解释有效促进了临床专家对模型决策逻辑的理解与采纳[16]。

尽管上述研究在各自领域取得了丰富成果，但综合运用主题建模、无监督聚类、统计推断与机器学习建模，系统刻画互联网医疗平台用户问诊行为特征，并验证用户行为差异性与可预测性的系统性研究仍然相对稀缺。基于此，本研究提出以下假设：

假设 H1：互联网医疗平台用户在病情描述与医患对话内容上存在可识别的主题结构。

假设 H2：用户在行为特征上存在显著群体差异。

假设 H3：基于基础行为特征，用户所属群体可以通过机器学习方法实现有效预测，且模型具备良好的可解释性。

2 数据预处理

2.1 数据来源与采集

本研究选取某国内主流互联网医疗平台“好大夫”上近 5 年（2019 年至 2024 年）的公开问诊数据。原始数据由爬取的结构化文本文件构成，包含患者提交的病情描述及与医生的完整对话过程。为确保后续分析的科学性与有效性，首先对原始数据进行了系统性清洗与标准化处理，旨在去除非结构化噪音信息并构建核心行为特征变量。

2.2 数据清洗与描述性统计

数据清洗过程中，针对问诊数据存在大量模板化内容、格式不统一及非医疗语义信息（如感谢语、问候语、医院地名等）的情况。构建了包含三十余项内容的冗余字段黑名单，并结合正则表达式对病情描述部分进行了逐句过滤。此外，为消除纯数字干扰信息，仅保留与年龄、时间单位有关的数字表达，其余均予以删除。同时，仅保留长度大于两个汉字的有效句子，以提升语料的语义密度。医患对话部分则保留原始结构，标注了对话轮数及说话者身份，为后续行为建模提供依据。在数据结构化过程中，共处理有效问诊记录 1,033,982 条，分别提取以下五项关键行为特征指标：病情描述字数、医患对话字数、病人轮数、医生轮数与总轮数。统计结果如表 1 所示。

表 1 互联网问诊记录行为特征描述性统计结果

特征名称	样本数	均值	标准差	最小值	中位数	最大值
病情字数	1,033,982	122.69	17,343.95	0	77	13,371,286
对话字数	1,033,982	154.50	32,392.71	3	65	25,399,076
病人轮数	1,033,982	2.08	401.59	0	1	306,204
医生轮数	1,033,982	2.52	359.73	0	2	282,980
总轮数	1,033,982	4.60	761.19	1	3	589,184

由表 1 可见，患者在问诊过程中平均输入 122.69 字描述自身病情，而医生与患者之间的对话平均长度为 154.50 字，总轮数约为 4.60 轮。值得注意的是，尽管大多数记录在轮数和字数上处于中等水平（如病情字数的中位数为 77 字），但最大值分布极端偏高（病情字数最大达到 1337 万

字，对话最大超过 2500 万字），说明数据中存在少量极端冗长记录。这一现象可能由于个别异常数据写入错误、系统日志嵌入、患者上传完整病历等因素所致。因此，后续建模将引入上限截断与异常值处理机制。下表是初步清理后的病情描述和医患对话示例。

表 2 病情描述和医患对话示例

id	病情描述	医患对话
0	筋膜炎 大于半年 右脚掌名指根部...	医生：穿合适的鞋子，坚持中药泡脚...
1	早发育 孩子在岁半的时候发现乳房...	病人：谢谢潘教授的回复，今天已经买了...
2	肠镜结果 非特异性结肠炎 直肠...	医生：你有什么症状？腹痛腹泻，粘液便？

3 主题建模与内容特征提取

3.1 文本预处理与词频分析

用户在互联网医疗平台中的问诊行为，首先通过其语言表达具体体现[17]。无论是最初的病情描述，还是后续与医生的交互对话，均是用户行为的重要表现形式。为深入揭示用户在线问诊时的行为特征与表达倾向，本研究采用词频统计与词云可视化方法，对问诊文本中“病情描述”与“医患对话”两个核心字段进行了系统性分析。

技术实现上，本研究使用 Python 语言进行数据处理，主要依赖 jieba 进行中文分词，collections.Counter 进行词频统计，WordCloud 与 matplotlib、seaborn 等包进行图形绘制。文本处理流程包括正则清洗文本，仅保留中文字符，去除数字、英文、标点与特殊符号。同时构建包含 130 余项的黑名单词库，清除所有寒暄、感谢、连接词、语气词、无效动词等高频冗余内容。并且进行精细化分词，仅保留长度不小于两个字、且不在黑名单中的有效词汇。

在“病情描述”部分，用户作为信息提供方，其行为表现为主观、自我导向、情境叙述倾向明显。词频排名前 30 的词汇中，最常见的为“半年”、“治疗”、“手术”、“怀孕”、“肿瘤”等，显著反映用户对疾病时间跨度、就医经历、治疗手段的高关注度（见图 1）。

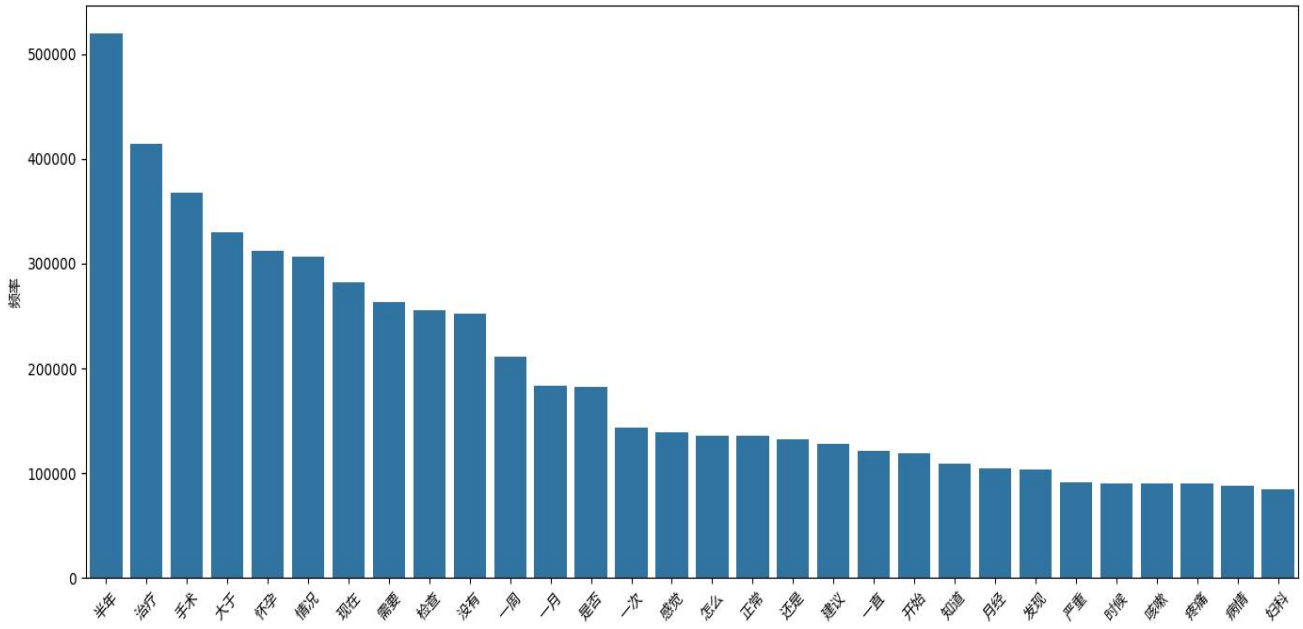

图 1 病情描述词频柱状图

从词云图来看，“怀孕”、“治疗”、“情况”、“大于半年”、“建议”等关键词具有显著视觉权重（见图 2），说明用户行为聚焦于自我诊断、症状归因及对治疗决策的急切需求。

图 2 病情描述词云图

而在“医患对话”部分，用户的行为转向互动式提问与信息获取。高频词汇如“治疗”、“检查”、“建议”、“没有”、“需要”、“手术”、“情况”等构成典型问答结构（见图 3）。

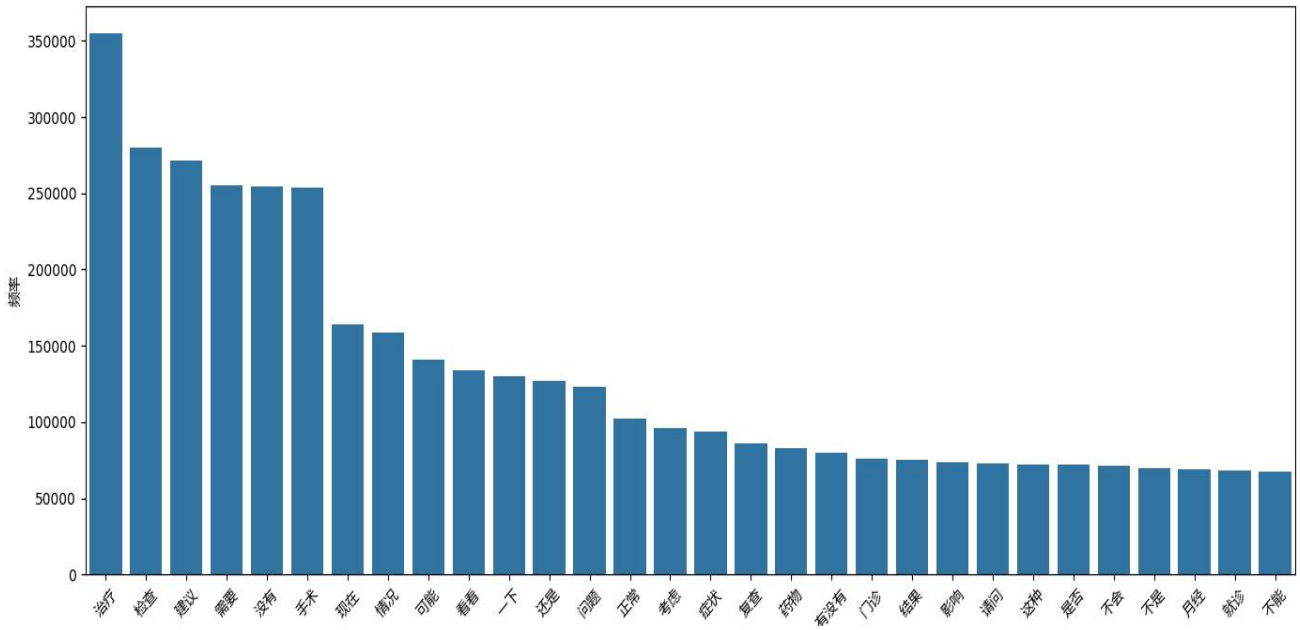

图 3 医患对话词频柱状图

这类行为体现出用户在医生引导下的行为响应、澄清与决策协商。词云图中“这种情况”、“建议”、“复查”、“药物”、“定期”等关键词尤为突出（见图 4）。反映出用户在对话中积极寻求方案、验证风险和获取后续指导的行为模式。

图 4 医患对话词云图

总体来看，病情描述部分主要表现为用户的主动表达行为，着重于描述病史、怀疑病因与希望获得的帮助。而医患对话则体现出用户的响应型行为，主要表现为对医生建议的理解、执行和反馈。这种行为转化的语言轨迹，构成用户在在线问诊过程中的信息输入-获取-反应链条，是分析行为阶段性演化的重要基础。

3.2 LDA 主题建模——病情描述

在用户提交的在线问诊中，“病情描述”是其主动表达健康状况与疑虑的首要载体。为了挖掘用户在不同类型疾病下的核心关注点，我们基于 200,000 条清洗后的病情描述文本，采用 LDA（Latent Dirichlet Allocation）模型进行主题建模[18]。为提高模型语义纯度，我们在预处理阶段引入了大规模中文停用词表与自定义黑名单，重点剔除了时间词（如“最近”、“半年”）、口语助词（如“呢”、“啊”）、常见称呼、非信息性词语等。从而确保模型训练聚焦于症状、器官、疾病名称等高信息密度的医学词汇。模型最终抽取出 8 个主题，每个主题包含 8 个主题词，分别代表了用户在不同场景下的描述倾向，如表 3 所示。

表 3 病情描述主题词

序号	主题词
1	怀孕，月经，妇科，子宫，出血，白带，阴道，输卵管，内膜，卵巢
2	术后，病情，切除，外科，化疗，检查，乳腺，骨折，囊肿，肿瘤
3	检查，服用，头晕，感觉，症状，控制，神经内科，心脏，内科，高血压，
4	眼睛，结节，甲状腺，眼科，回声，视力，右眼，左眼，淋巴结，近视，
5	疼痛，治疗，大便，走路，包皮，勃起，骨科，早泄，突出，严重
6	治疗，皮肤科，皮肤，疙瘩，脸上，严重，过敏，出现，痘痘，湿疹，
7	怀孕，检查，胎儿，出生，阳性，感染，血管瘤，小孩，宫颈，发育
8	咳嗽，感冒，发烧，鼻子，头孢，颗粒，鼻炎，耳朵，肺炎，呼吸

根据关键词，我们可以总结得出主题 1 聚焦于妇产科与生殖系统问题，主题 2 则多为外科干预后状态。主题 3 涉及慢性疾病与药物控制，主题 4 主要为眼科与甲状腺问题。主题 5 与疼痛与生殖系统功能障碍相关，主题 6 聚焦皮肤病与过敏反应。主题 7 指向孕期异常与婴幼儿发育，而主题 8 主要涉及呼吸感染与耳鼻喉常见疾病。

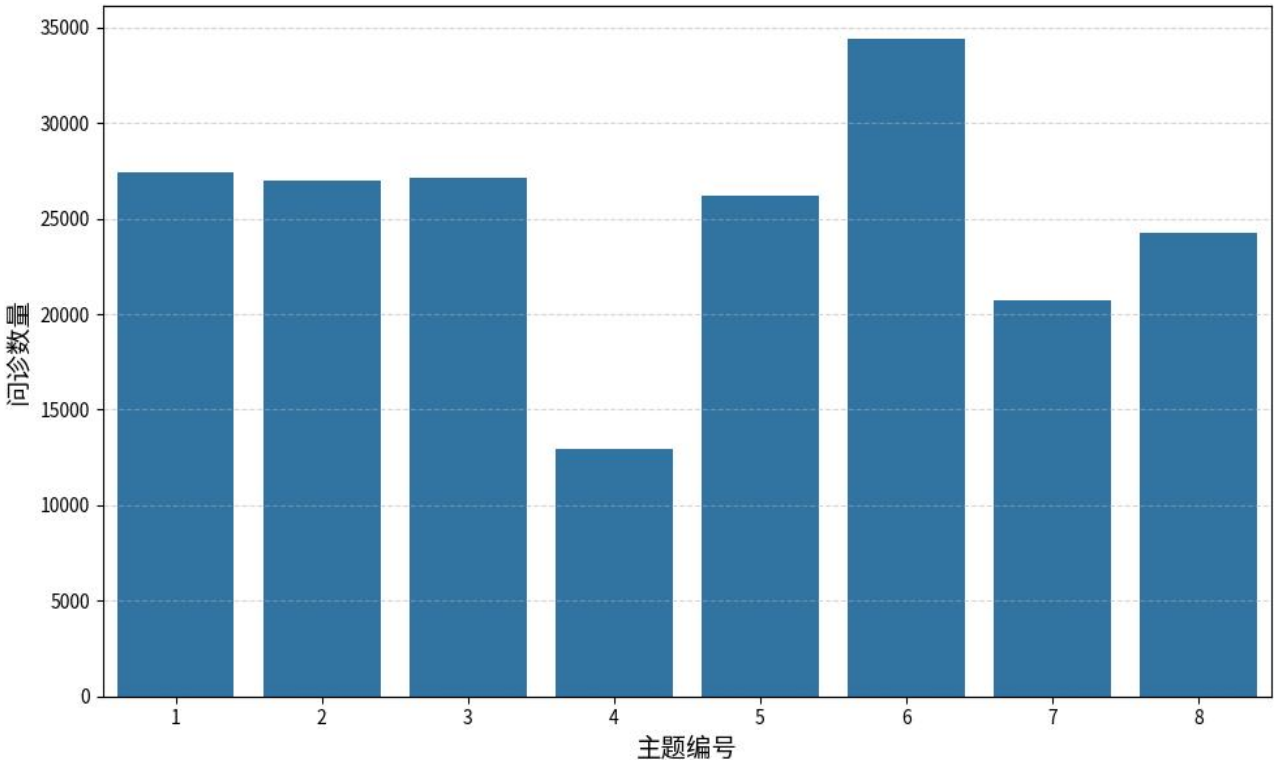

图 5 病情描述主题分布

整体来看，用户的病情描述呈现出明确的主题归类，并在不同主题下呈现出差异化的表达重点。反映出在线问诊用户在表达需求时的目标导向性。

3.3 LDA 主题建模——医患对话

与病情描述不同，医患对话部分主要包含医生的回复以及患者的补充提问或回应，结构上体现出更多功能性与交互性。我们对同样规模的 20 万条医患对话数据进行建模处理，采用相同的文本清洗与 LDA 建模流程，提取出 8 个对话主题。

表 4 医患对话主题词

序号	主题词
1	检查，月经，怀孕，正常，复查，影响，出血，子宫，问题，宫颈
2	手术，治疗，切除，术后，病理，效果，做手术，费用，化疗，肿瘤
3	治疗，药物，口服，咳嗽，感染，症状，感冒，服用，效果，检查
4	照片，感染，外用，观察，皮肤，局部，治疗，不用，湿疹，处理
5	正常，检查，大便，复查，问题，功能，化验，甲状腺，血压，抗体
6	门诊，预约，时间，检查，治疗，就诊，住院，过来，直接，面诊
7	检查，治疗，就诊，诊断，问题，明确，片子，症状，进一步，根据
8	症状，感觉，治疗，疼痛，这种，注意，平时，中药，问题，睡眠

从关键词分析来看，医生在不同主题下展现出一定程度的回应风格差异。

主题 1 主要围绕“检查-复查-妊娠影响”的临床路径建议，主题 2 侧重于“手术-术后-费用-效果”的治疗反馈。主题 3 则集中在“药物使用、感冒咳嗽、口服建议”，主题 4 多为皮肤病、感染及外用处理方法的解释。主题 5 则重在功能性异常如“化验、甲状腺、抗体”分析，主题 6 涉及“预约、住院、就诊流程”。主题 7 聚焦“诊断建议与影像说明”，主题 8 则是用户关于“症状注意事项”的频繁追问及医生中药调理建议。

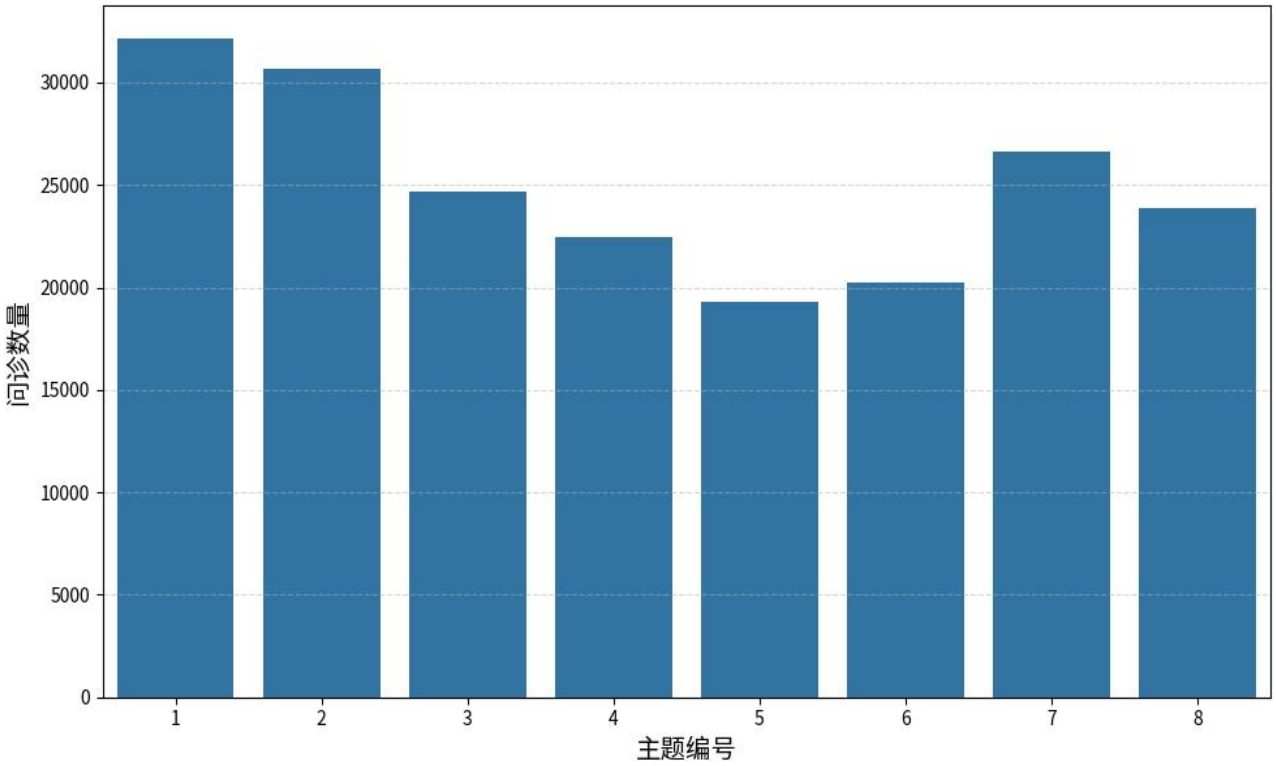

图 6 医患对话主题分布

可见医生的答复在不同问题情境中保持医疗逻辑一致的同时，也适度适配了用户的核心关注点。换言之，医生端虽具有响应规范，但仍体现出对用户问题的语义感知能力。

3.4 主题行为特征差异分析

在完成语义主题建模后，我们进一步考察用户在不同主题下的行为模式，包括“病情字数”、“对话字数”与“总轮数”三个维度，以此刻画信息输入与交互深度的变化趋势。

结果表明，用户在病情描述部分的平均字数总体集中在 105 至 107 字之间，标准差较小。说明用户在不同疾病主题下的表达意愿与习惯具有明显的一致性。这一现象可能源于平台文本输入结构引导、提问模板化提示或用户认知结构相对稳定等原因。

我们使用箱形图展示了“病情字数”、“对话字数”与“总轮数”的分布特征，如图 7 所示。

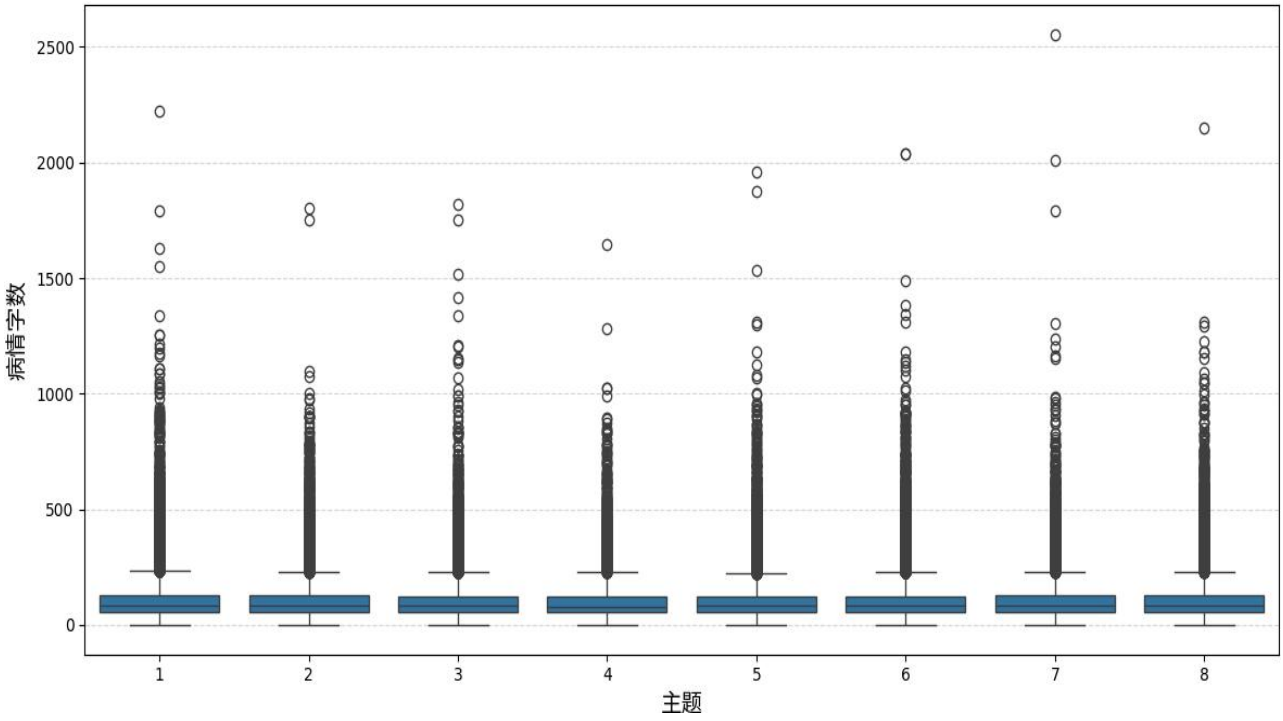

图 7 病情字数在不同主题的分布

图中显示，各主题下的箱体高度、上下四分位值极为接近，且异常值（即上须线之外的点）高度集中。说明用户在填写病情时，虽然语义内容因主题不同而有差异，但表达长度的上下限波动较小。这种模式可能受平台输入引导和用户习惯的共同作用影响。

医生在不同主题下的对话行为也呈现出高度一致的响应长度与轮数波动。对话字数均值区间 115 至 116，轮数在 3 至 4 次。反映出医生端在答复长度控制上已形成较为成熟的服务规范，不随用户提问主题而显著波动。

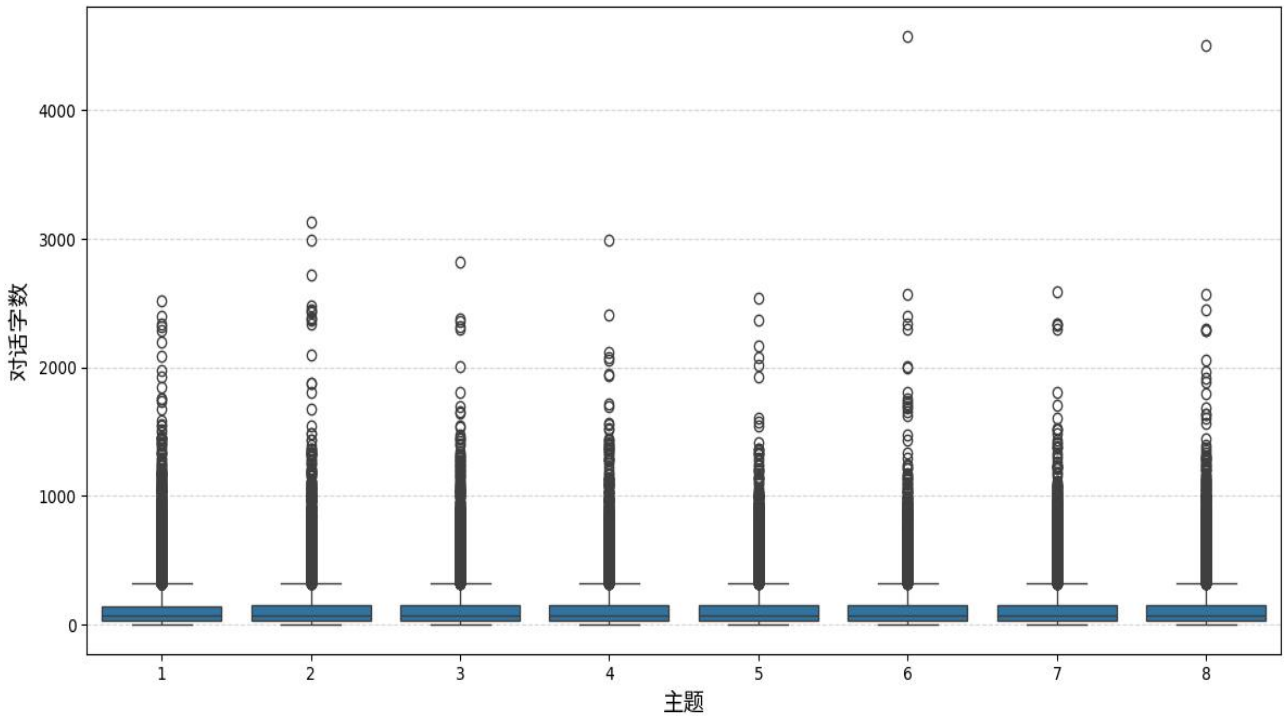

图 8 对话字数在不同主题的分布

在医患对话部分的箱形图中，医生对患者问题的回应长度与轮数更为集中，几乎呈现完全一致的分布形态。这一特征在“手术后康复”、“感冒发烧”等主题中尤为明显。整体来看，箱形图展示结果从视觉维度证实了用户与医生在“字数与轮数”上的稳定行为特征，为后续统计检验提供了直观铺垫。

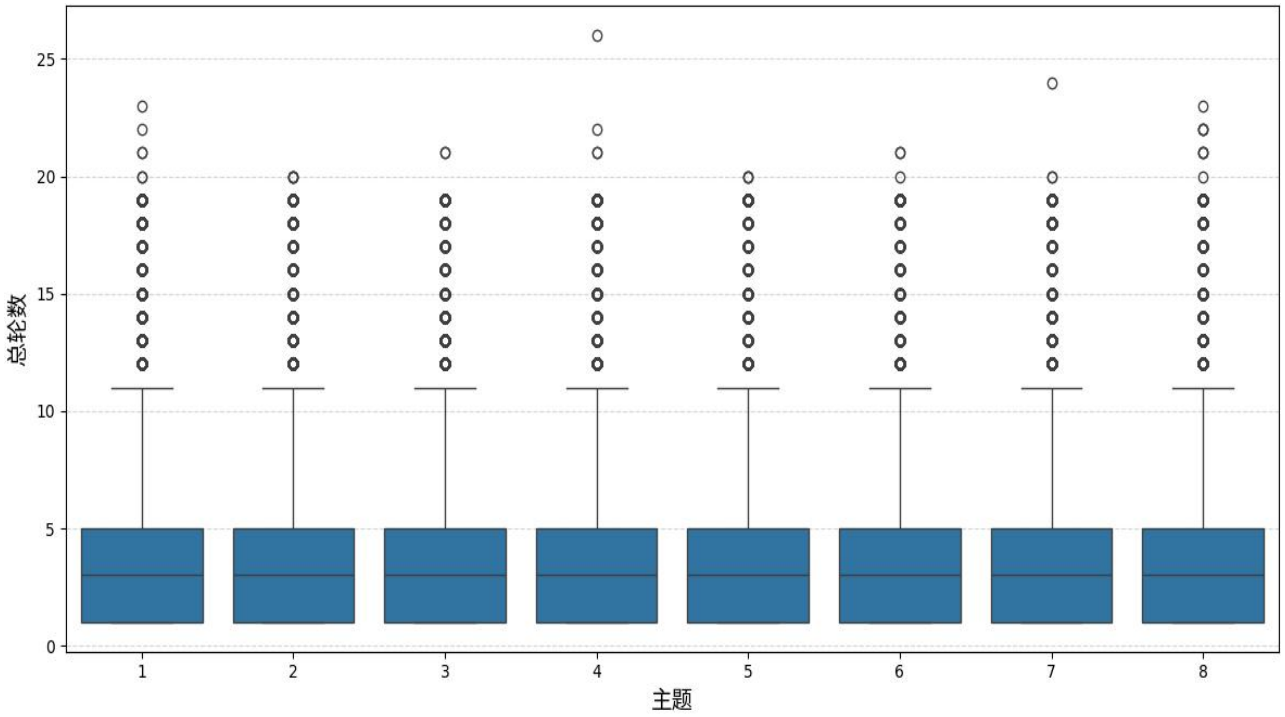

图 9 对话轮数在不同主题的分布情况

从整体统计结果来看，虽然用户问题内容在语义上存在较大差异，但在行为上却表现出跨主题的一致性和稳定性，为平台统一问答结构设计与自动答复模型训练提供了有利支撑。

3.5 主题行为一致性检验

为了进一步验证上述观察，我们对用户的“行为分布是否受主题影响”进行了统计性检验。我们构造了两个离散标签：病情描述字数是否超过 50（长文本）与对话字数是否超过 70（超长对话），并分别在病情描述与医患对话两个语料下开展卡方检验。

表 5 病情描述部分用户行为在不同主题下的卡方检验

变量	卡方统计量	自由度	P 值
长文本	0.6939	7	0.998382
超长对话	18.1451	7	0.011332

检验结果显示，在病情描述部分，“长文本”标签在不同主题间分布差异不显著（P=0.998382），但“超长对话”标签却表现出显著性差异（P=0.011332），说明尽管用户输入长度基本保持一致，但在某些主题下用户更易引发医生多轮应答，如儿童疾病、妊娠问题和术后疑难问题等主题更具“交互黏性”，体现出高焦虑感与解释需求。

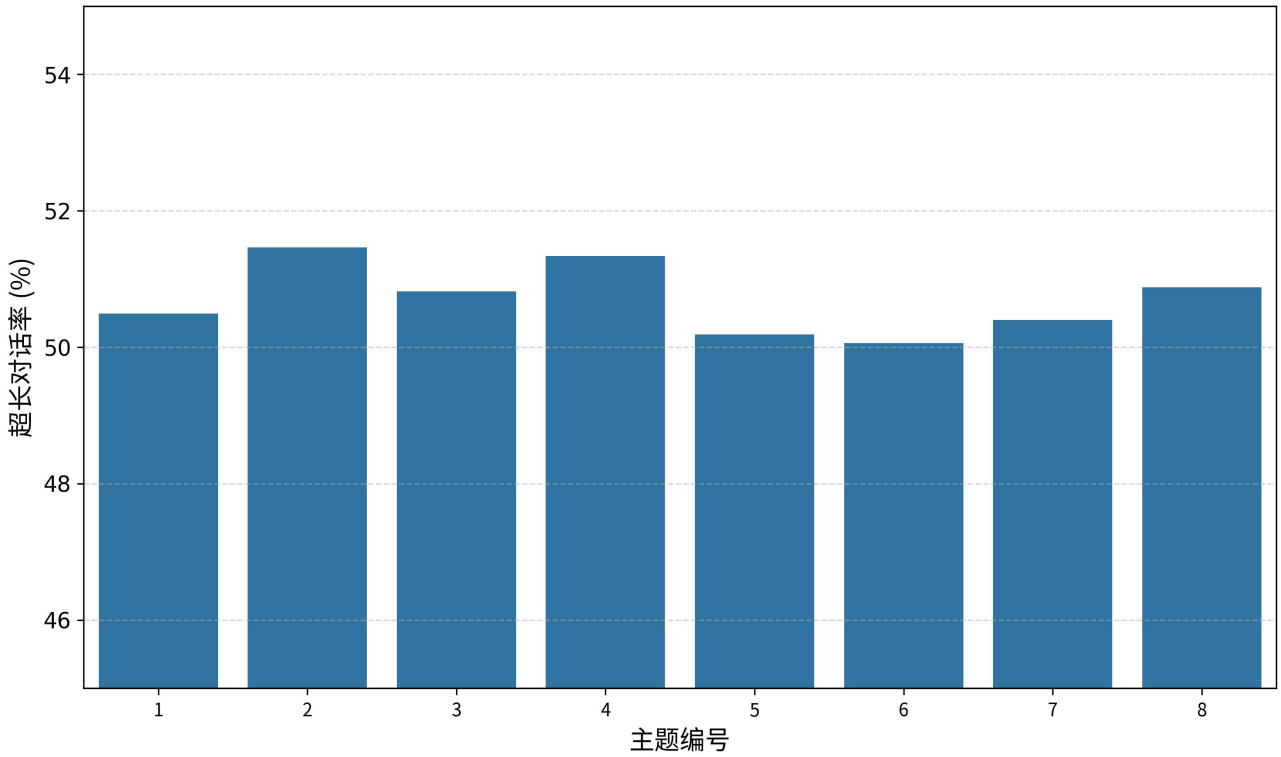

图 10 病情描述在不同主题下超长对话率的分布

为了进一步明确“超长对话”的主题差异分布，我们绘制了病情描述语料下的超长对话率柱状图（见图 10）。从图中可以看出，尽管整体比例分布相对接近，但主题 2（外科术后康复）和主题 4（眼科与甲状腺问题）下的超长对话比例略高于其他主题，分别超过 51.5%，表明这些类型的问题往往伴随着更多轮次的澄清和解读。相对而言，主题 6（皮肤病）与主题 5（功能障碍与疼痛）则对话率略低，说明其信息交互较为直接、结论明确。这些细微差异虽然在描述性统计中不明显，但通过柱状图的放大对比仍可观察到用户互动意愿与对医生说明的依赖程度随语义主题的细微波动。这一图示在宏观上支持了前述卡方检验的显著性结果，即用户在不同情境下对医生交互长度存在认知差异，而这种行为正是互联网医疗中用户感知“专业度”和“服务密度”的重要体现。

相较之下，在医患对话部分，“长文本”和“超长对话”均未呈现统计显著性（P 值分别为 0.96 与 0.36），说明医生在不同主题情境下，其响应长度和频率保持高度稳定。

表 6 医患对话部分用户行为在不同主题下的卡方检验

变量	卡方统计量	自由度	P 值
长文本	1.9928	7	0.960237
超长对话	7.6536	7	0.364126

此结果从统计学上印证了前述“用户行为随主题适度波动、医生行为高度一致”的分析结论，也表明在互联网医院场景中，用户的信息需求更多取决于其所面对的疾病情境，而医生响应则更受平台培训、服务标准化流程所约束。

本章节揭示了病情描述和医患对话在主题结构上的差异性，支持 H1。

4 用户行为聚类分析

4.1 行为特征与聚类数选取

为识别不同用户在在线医疗场景下的行为模式，本研究采用无监督聚类方法对用户进行分群分析，进一步刻画不同类型用户在交流活跃度、话题偏好与表达方式上的差异[19]。

首先，选取“病情字数”、“对话字数”和“总轮数”三项特征作为聚类依据。这些指标可分别从用户主动描述、互动过程和交流频次三个维度综合反映用户在医患沟通中的行为活跃度。

考虑到原始样本体量较大，为提高聚类效率，我们进行了 50000 条样本的随机抽样，并对选定特征进行了标准化处理。随后采用 KMeans 聚类算法对用户进行建模。为确定最优聚类数 K，分别绘制了基于误差平方和（SSE）的肘部法图以及基于轮廓系数的 K 值评估图（见图 11）。

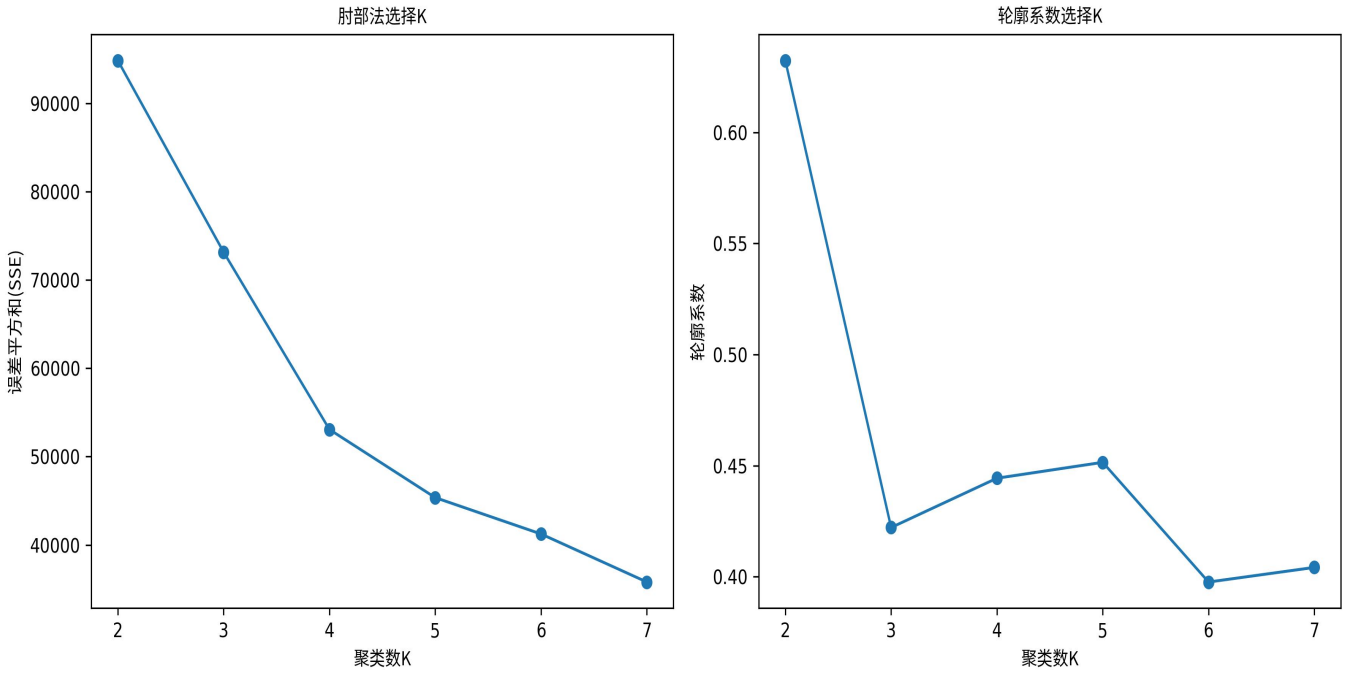

图 11 肘部法与轮廓系数评估

从结果可以看出，K=3 时 SSE 下降斜率趋缓，轮廓系数亦达到局部峰值，综合考虑模型性能与可解释性，最终选择 K=3。

4.2 聚类降维与可视化分析

我们采用主成分分析（PCA）与统一流形逼近映射（UMAP）两种降维算法对用户在三维特征空间中的分布进行可视化。PCA 作为经典线性降维方法，能够较快地保留样本的全局变异性结构[20]，但在本研究中可视效果较为集中，存在群体重叠问题（见图 12）。

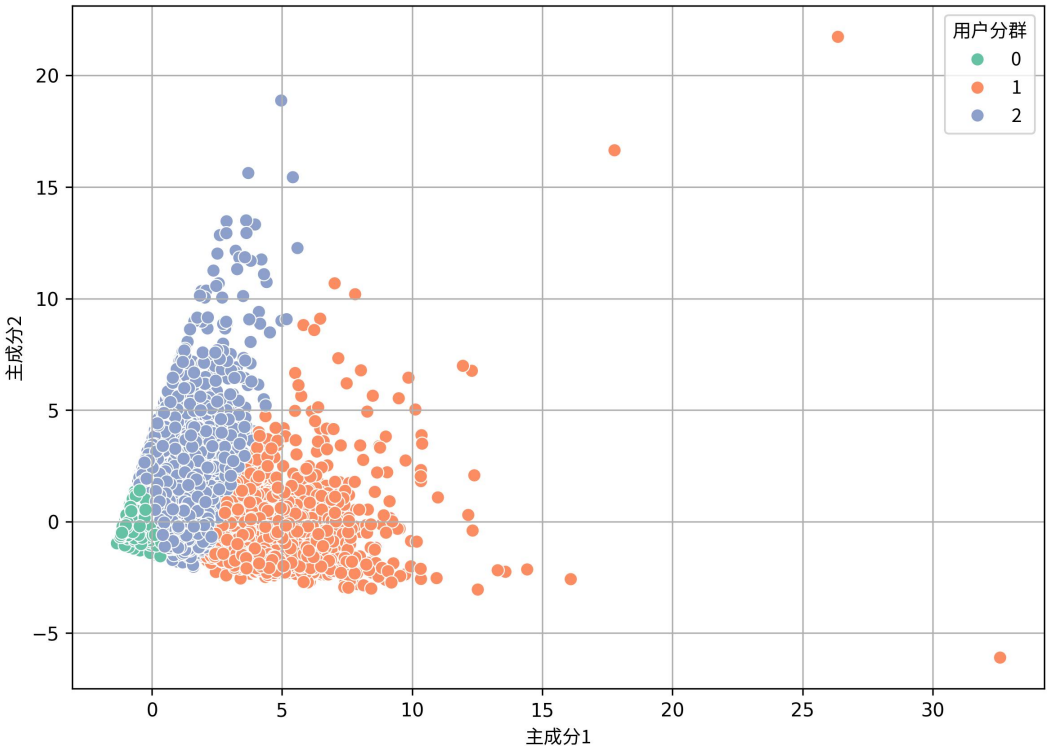

图 12 PCA 用户聚类

因此，为进一步提升聚类边界的可分性，我们引入了 UMAP 降维技术，该算法在保留数据局部结构的同时对复杂非线性边界的表示能力更强[21]。UMAP 结果显示各类用户呈现更清晰的聚类边界（见图 13），且对应的平均轮廓系数为 0.368，验证了其良好的群体划分效果。

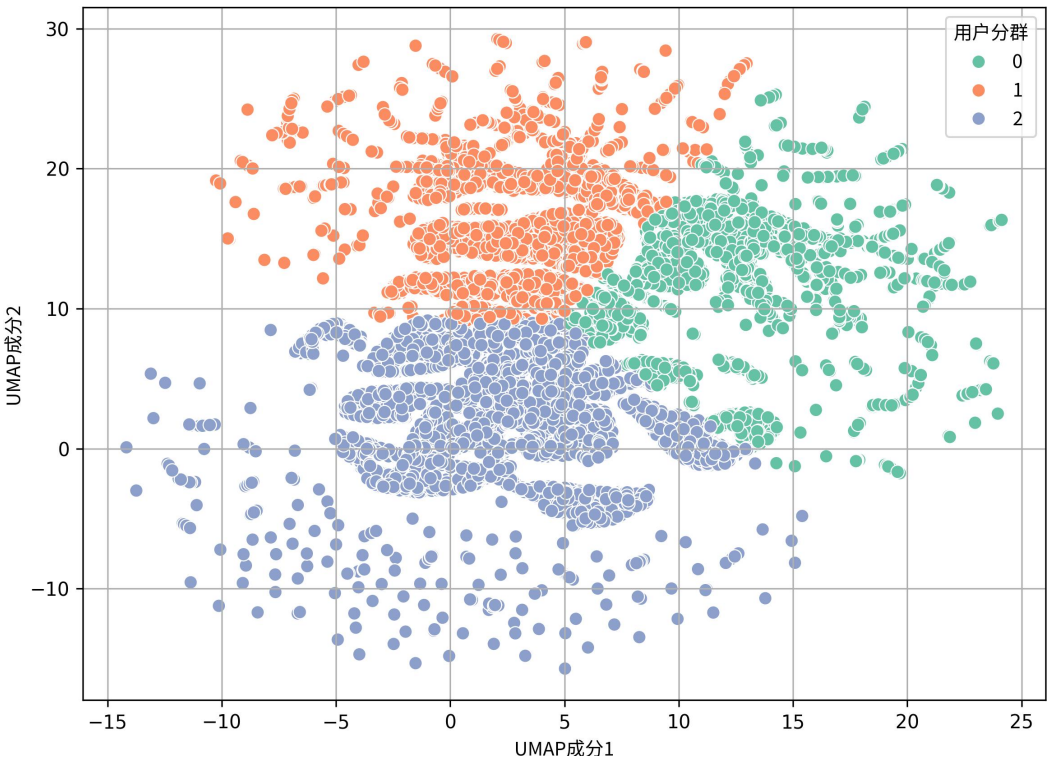

图 13 UMAP 用户聚类

4.3 用户聚类活跃度分析

在完成用户聚类之后，我们进一步围绕“病情字数”、“对话字数”与“总轮数”三项核心指标，深入分析各类用户在平台上的互动活跃程度。该部分通过对群体均值和分布的比较，揭示出不同用户在沟通密度和交流频次上的显著区别。

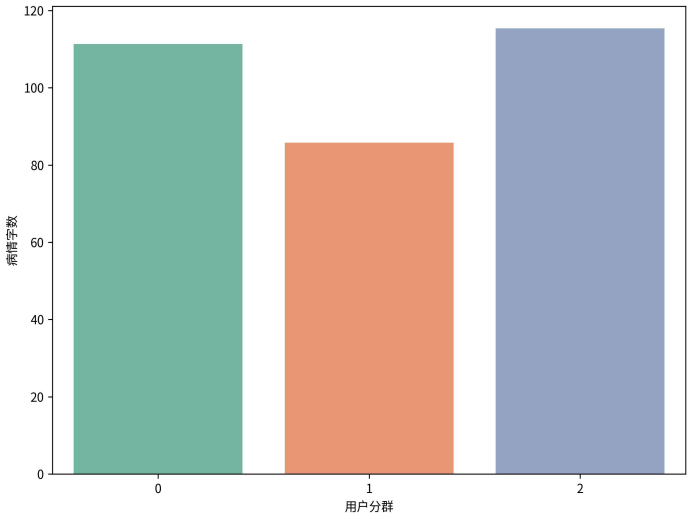

图 14 不同用户聚类的病情字数分布柱状图

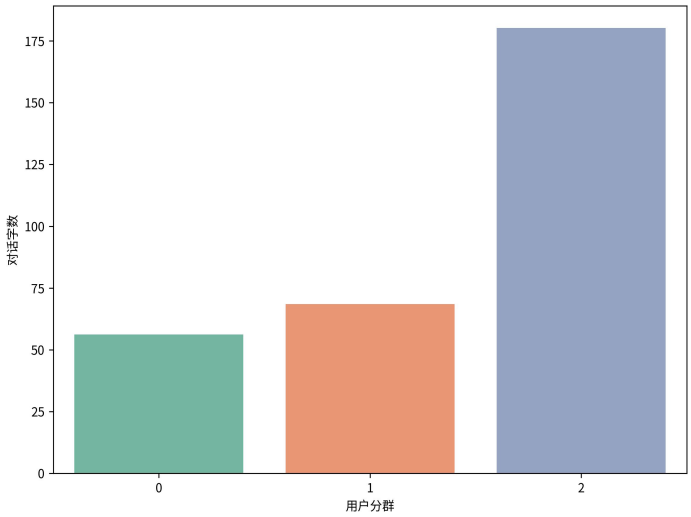

图 15 不同用户聚类的对话字数分布柱状图

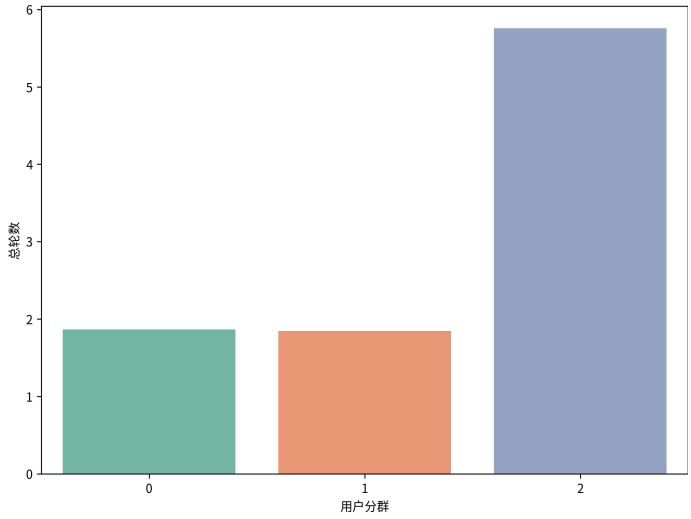

图 16 不同用户聚类的对话轮次分布柱状图

从群体均值（见 14、15、16）来看，群体 2 用户在所有活跃度维度上均显著高于其余群体，其中平均对话字数为 168.66，总轮数为 5.51，远高于群体 0 和群体 1。相较之下，群体 0 和群体 1 均属于低活跃用户群体，其中群体 1 在三项指标中均处于最低水平，表现为最为简短的自述与最少的互动轮次。

此外，箱线图分析（见图 17、18、19）进一步揭示了群体内部的分布差异。特别是在对话字数与轮数上，群体 2 呈现明显的长尾分布结构，意味着该群体内部存在一部分极高活跃用户，他们在问诊中倾向于提供更多细节并保持更长时间的交流。这类用户可能具有更强的表达能力或更高的信息诉求，进一步强调了群体间的行为异质性。

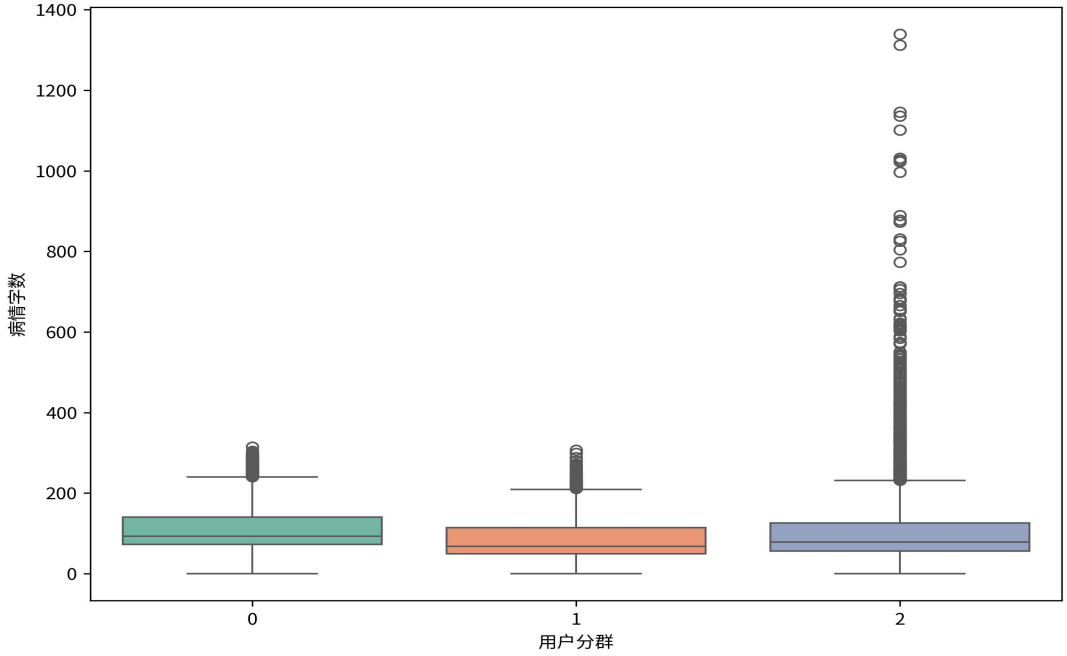

图 17 不同用户聚类的病情字数分布箱形图

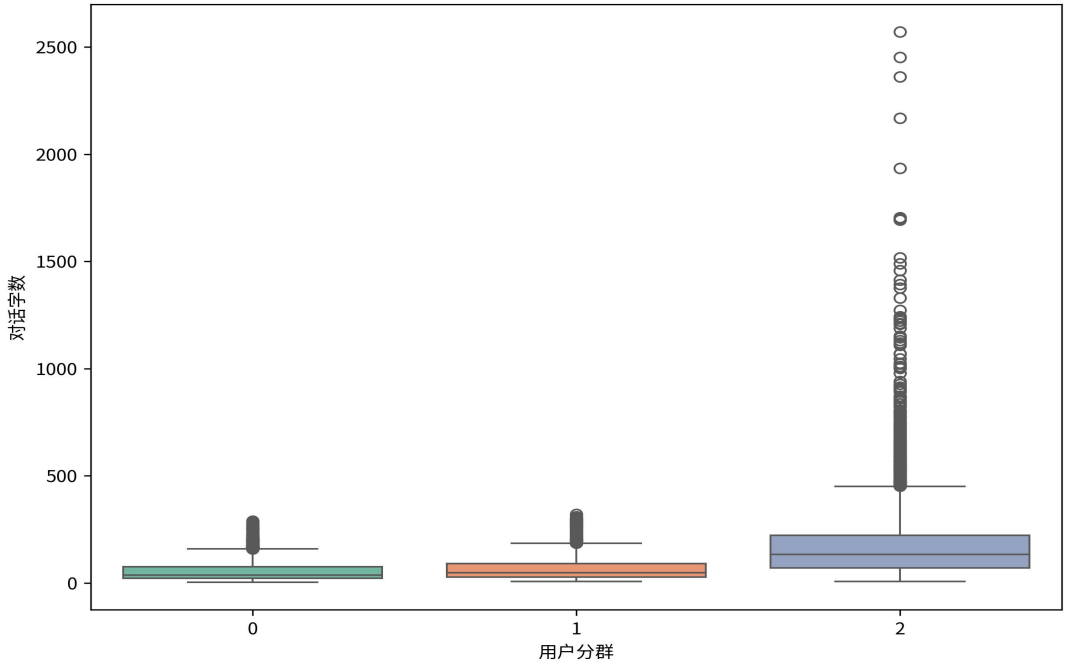

图 18 不同用户聚类的对话字数分布箱形图

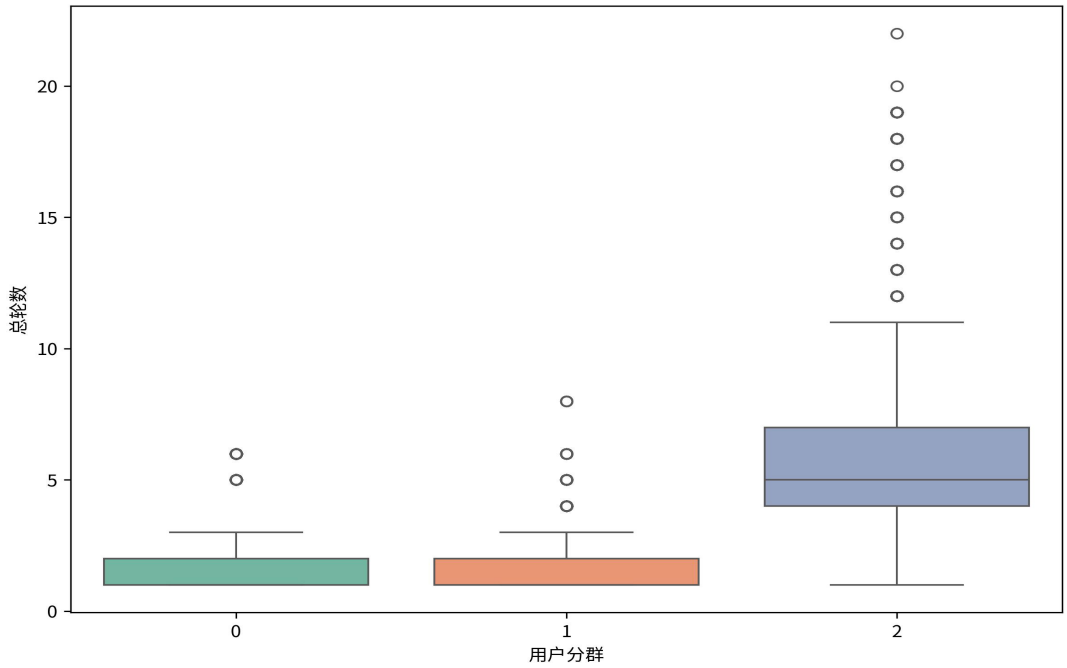

图 19 不同用户聚类的总轮数分布箱形图

综合来看，活跃度的多维对比不仅支持了聚类结果的有效性，也反映了用户在问诊平台上存在显著的参与深度差异。群体 2 可视为高活跃群体，而群体 0 和 1 则在表达密度与交流频次方面显得相对被动。

4.4 群体行为差异的显著性检验

为进一步评估聚类结果的有效性，本节对不同用户群体在“病情字数”、“对话字数”与“总轮数”三个维度上的统计差异进行了系统性检验。我们首先进行了正态性检验（Shapiro-Wilk 检验）和方差齐性检验（Levene 检验），以判断是否满足参数检验的前提条件。结果显示，三个特征在不同群体中的 P 值均远小于 0.05，表明数据在各群体之间不符合正态分布，且存在显著的方差不齐性。

表 7 正态性检验结果

变量	P 值
病情字数	7.157712022044483e-53
对话字数	8.103272907941582e-62
总论数	8.476047605202817e-77

表 8 方差齐性检验结果

变量	P 值
病情字数	0.0000
对话字数	0.0000
总论数	0.0000

在这种前提下，传统的 ANOVA 分析已不再适用。为保证检验结果的稳健性与科学性，我们选用了非参数方法——Kruskal-Wallis H 检验，该方法无需数据服从正态分布，适用于多个独立样

本的中位数比较，尤其适合处理医疗文本这类偏态分布严重的语言特征数据[22]。检验结果表明，所有指标的 P 值均小于 0.0001，达到极高显著性水平。

表 9 Kruskal 显著性差异检验

变量	P 值
病情字数	0.0000
对话字数	0.0000
总论数	0.0000

这意味着，不同用户群体在描述长度、交流字数与交流轮数这三方面均存在显著差异。这一结论与前述活跃度分析的均值比较结果一致，为聚类模型的合理性提供了进一步佐证。

总之，该部分通过严谨的统计检验流程，从方法选择到结果解释，全面揭示了群体间交流行为的显著性结构分异，也进一步支持了先前在用户分群与活跃度上的分析结论。

4.5 聚类群体对话长度特征分析

为了进一步从行为质量角度区分用户群体，我们引入“超长对话率”作为补充指标。将“对话字数”大于 70 的记录定义为“超长对话”，并统计每一群体中超长对话出现的比例，结果见图 20。

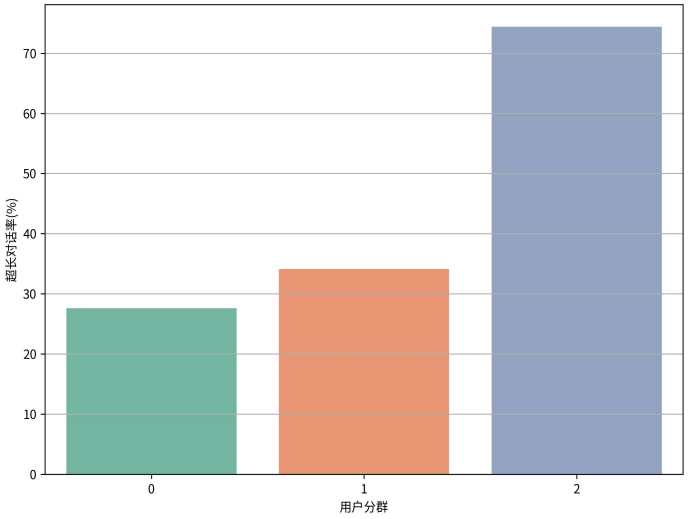

图 20 不同用户聚类的超长对话率分布柱状图

群体 2 的超长对话率高达 74.37%，远高于群体 0（27.6%）和群体 1（34.1%）。同时卡方检验统计量达 1870.10（$P < 0.0001$），显示不同用户群体在超长对话行为上存在高度显著差异。

表 10 超长对话率在不同群体间的卡方检验

卡方统计量	自由度	P 值
1870.1042	2	0.000000

这进一步验证了群体 2 用户在信息表达与主动沟通上具有更强的动机与能力，表现出明显的内容丰富性与互动意愿。

4.6 用户群体的主题一致性分析

我们基于 LDA 模型中生成的“主导主题编号”，构建了“用户分群 × 问诊主题”的交叉表（见表 11），并开展卡方检验以判断各类用户在主题分布上的偏好是否存在差异。

表 11 用户分群与主题的交叉表

	主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7	主题 8
用户群体 1	388	334	393	170	369	434	297	336
用户群体 2	331	326	281	151	321	416	245	259
用户群体 3	618	619	655	299	603	790	475	556

结果表明，P 值为 0.2709，不具显著性，说明不同活跃度用户在就诊内容的关注方向上较为一致，主题覆盖面广且分布均衡。

本章节证实了不同用户群体间的行为特征存在显著差异性，支持 H2。

5 机器学习建模与用户分类预测

5.1 建模动机与特征选择

本节旨在进一步验证前述用户分群的合理性和可区分性。具体而言，我们将尝试使用机器学习模型对用户的群体归属进行预测，考察仅基于行为层特征是否能够准确识别用户属于哪一类。这一过程不仅具备模型检验价值，也为后续在未知用户数据上实现自动标签赋予提供方法论支持。在聚类分析中，我们观察到“病情字数”、“对话字数”、“总轮数”在各用户群体间分布差异显著。尤其群体 2 表现出更高的文字量与交互轮次，体现出鲜明的活跃特征。因此，我们选择这三项行为指标作为机器学习建模的输入特征，探索它们在“区分用户群体”任务中的判别能力。

与问诊主题等语义类变量不同，这些行为特征的采集成本较低、稳定性较高，具备良好的泛化适应性。若仅凭这三项即可取得高预测准确率，则说明我们通过聚类划分出的用户群体，在行为模式层面具有较强的区隔性。

5.2 数据准备与预处理策略

我们使用已完成 UMAP 聚类标签的数据作为建模样本，并剔除未分群（标签为-1）的用户数据，确保训练集标签清晰有效。随后，将特征矩阵进行标准化处理（StandardScaler），使得各特征具有相同的尺度。以减少模型对量纲的偏倚，尤其对 SVM、Logistic 回归这类依赖距离或线性权重的模型更为关键。

数据集按 8:2 比例划分为训练集与测试集，用于模型拟合与泛化能力评估，确保结果具备一定的稳定性与可解释性。

5.3 多模型对比评估

我们选取了五种典型的监督学习分类模型开展建模实验。首先是 Logistic Regression，这是一种线性可分的基础模型，对复杂非线性边界适应性较差，存在一定欠拟合[23]。其次是 Support Vector Machine，可以通过核技巧引入非线性映射，但在多分类情境下计算成本相对较高。然后是 Random

Forest，可以集成多棵决策树，在处理高维特征和非线性关系方面效果优异。最后是 XGBoost 与 LightGBM，这是目前业界主流的梯度提升树方法，且训练效率和性能的综合表现较好[24]，是本任务的理想候选算法。

表 12 多模型运行结果对比

模型	准确率	macro F1
Logistic Regression	0.7001	0.66
SVM (RBF)	0.8242	0.81
Random Forest	0.9648	0.96
XGBoost	0.9690	0.97
LightGBM	0.9690	0.97

最终我们选择了 LightGBM 作为目标模型，进行进一步的超参数调优。

5.4 LightGBM 模型调参与验证结果

为了进一步提升模型性能，我们在 LightGBM 上进行了系统的网格搜索调参，采用 5 折交叉验证评估每组参数组合的平均表现。选定的最佳参数为“learning_rate : 0.1”, “max_depth : 20”, “n_estimators: 200”, “num_leaves: 70”。

最终，模型在交叉验证中获得 97.83% 的最优准确率。在测试集上，准确率达 97.72%，三类用户群体的 precision、recall 和 F1-score 均表现优秀，模型在不同群体上的判别能力均衡，没有出现明显偏倚。

5.5 模型可解释性分析

（1）混淆矩阵分析

我们首先绘制了真实标签与预测标签的混淆矩阵图。

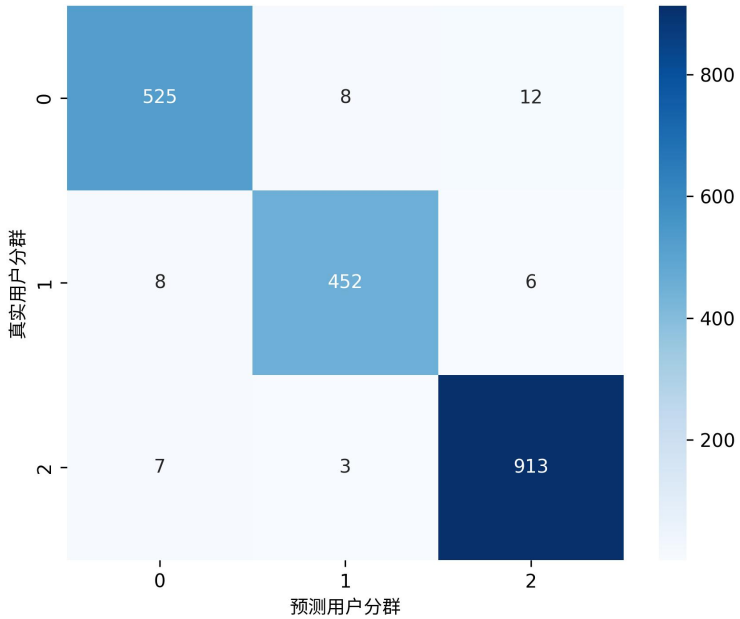

图 21 真实标签与预测标签的混淆矩阵图

结果显示群体 2 几乎全部被正确分类（913/923），说明其行为特征最具代表性。群体 0 与群体 1 在部分样本上出现混淆，可能由于这两类用户在字数与轮数上分布存在部分重叠，边界相对

模糊。

这印证了聚类分析中的发现，即群体 2 在活跃度上显著高于其他两类，而群体 0 与 1 之间的行为边界较不明确。

（2）特征重要性解读

我们对 LightGBM 模型特征重要性进行了可视化。结果显示病情字数贡献度最高，表明用户在病情自述阶段的信息量对其行为风格具有强区分力。对话字数紧随其后，进一步体现了交互内容的丰富程度也是划分用户的重要依据；总轮数作为对会话频度的量化，在一定程度上增强模型的补充判别力。

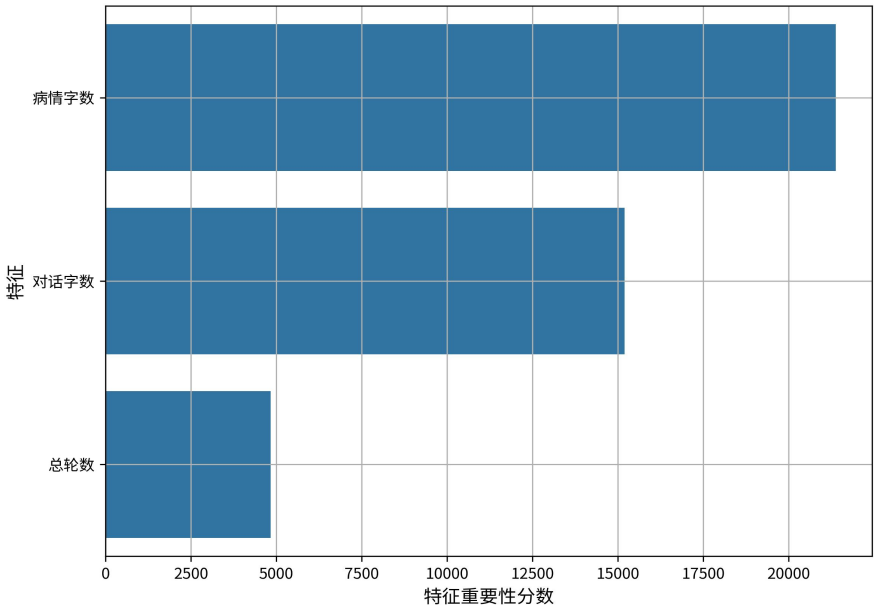

图 22 LightGBM 特征重要性分析

综上，这些活跃度相关的行为特征，不仅在聚类中体现出稳定差异，也在监督建模中展现出强可预测性。

5.5 模型预测机制分析

为了进一步剖析模型的预测机制，我们采用 SHAP（SHapley Additive exPlanations）方法对已训练的 LightGBM 分类模型进行解释性分析。SHAP 方法基于博弈论中的 Shapley 值原理，能够量化各输入特征对模型输出结果的贡献方向及强度，从而实现模型决策过程的可视化和透明化[25]。

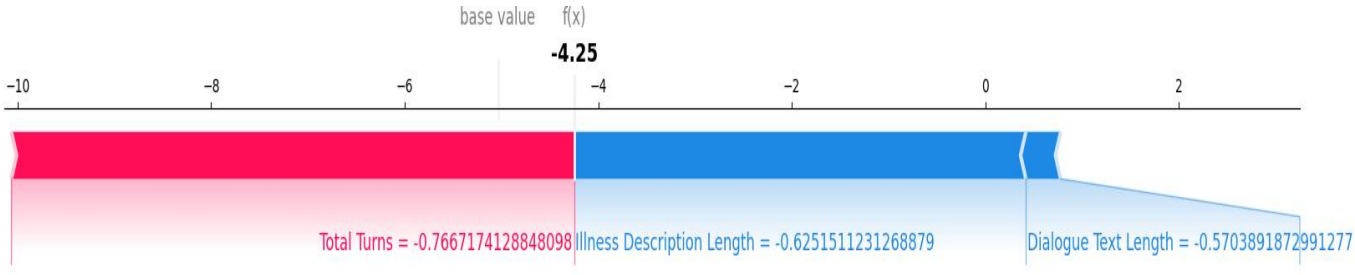

图 23 SHAP 力图（样本 10）

在本研究中，我们随机抽取了一个用户样本进行个体层面的解释，并绘制了其对应的 SHAP

力图。如图所示，模型最终对该用户样本的预测输出为 f(x) = -4.25，显著低于该分类的基准值（base value），因此模型判断该用户不属于该群体。红色区域表示会提高模型预测值的特征，在该样本中主要是“总轮数”，但其值较低，因此贡献也较弱。蓝色区域表示会降低预测值的特征，在该样本中“病情字数”与“对话字数”为主导因素，对模型输出产生了显著的负向推动作用。

整体来看，该图可清晰展示模型如何综合多个输入特征得出当前预测结果，与前文特征重要性排序结果高度一致，从侧面验证了模型学习到的逻辑是合理、可解释的。

本章节通过基础特征“病情字数”、“对话字数”和“总轮数”成功实现高准确率分类预测用户群体，支持 H3。

6 结论

本研究基于互联网医疗平台真实用户问诊数据，围绕用户内容表达、行为模式与预测建模三个维度，构建了一套完整的数据分析体系。研究综合采用词频分析、主题建模、聚类分析、降维可视化与机器学习分类等多种方法，深入挖掘用户的在线问诊行为特征，探究其结构性规律与可预测性。

在内容建模方面，研究对“病情描述”与“医患对话”分别进行了预处理和 LDA 主题建模，提取出 8 类高质量的主题结构，涵盖常见病种、检查项目、治疗策略、术后恢复等语义方向。主题词分布清晰，语义集中，反映出互联网问诊内容的结构性特征，为理解用户信息诉求提供了内容基础，该部分支持 H1。在主题分布与用户群体的关联分析中，初步卡方检验结果显示，不同聚类用户在主导主题上的分布整体差异不显著，表明内容倾向性具有一致性。然而，进一步结合行为标签分析发现，不同主题之间在“超长对话”这一行为维度上存在统计学显著差异（P = 0.011332），提示部分内容主题可能隐含着用户更高的表达活跃度需求。

在用户行为分析方面，研究基于“病情字数”、“对话字数”、“总轮数”三个核心行为指标构建行为画像，并通过 KMeans 算法完成自然聚类。在 UMAP 降维后的可视化图中，用户聚类结构清晰，轮廓系数达 0.368，初步表明行为群体划分的合理性。各群体在活跃度特征上的对比进一步揭示了系统性差异，尤其在“对话字数”与“总轮数”两个维度上表现出稳定分层，非参数检验结果显示各类差异均具显著性。超长对话比例也存在明显群体间差异，这些发现为平台进行用户分层管理和差异化服务提供了直接参考，同时支持 H2。

在预测建模方面，研究尝试构建了多个监督式分类模型，以实现对用户分群的自动预测。LightGBM 模型在训练测试中表现最优，调参后在测试集上达到 97.72%的准确率，分类各项指标均处于优秀水平，验证了基础行为特征的高度判别力，并且支持 H3。在此基础上，研究引入 SHAP 方法解释模型预测过程，从特征重要性与单样本贡献层面增强了模型的透明度。SHAP 分析结果显示，“对话字数”为模型最关键的决策依据，符合活跃度分析的全局趋势，也提升了模型在医疗应用场景下的可解释性与信任度。

参考文献

[1]国务院办公厅. 关于促进“互联网+医疗健康”发展的意见[EB/OL]. 2018-04-28. http://www.gov.cn/zhengce/content/2018-04/28/content_5286645.htm.

[2]张瑞,卢奕汝,朱瑞轩,等. 医疗公平视角下互联网医疗伦理治理路径初探[J]. 中国医学伦理学,2024,37(1):54-60. DOI:10.12026/j.issn.1001-8565.2024.01.07.

[3]徐召鹏,张录法. 空间邻近优势、互联网医疗与分级诊疗[J]. 浙江大学学报（人文社会科学版）,2025,55(3):60-75. DOI:10.3785/j.issn.1008-942X.CN33-6000/C.2023.10.112.

[4]Han, Y., Lie, R. K., & Guo, R. (2020). The internet hospital as a telehealth model in China: systematic search and content analysis. *Journal of medical Internet research*, *22*(7), e17995.

[5]Mizan, T., & Taghipour, S. (2022). Medical resource allocation planning by integrating machine learning and optimization models. *Artificial Intelligence in Medicine*, *134*, 102430.

[6]Mathews, S. C., McShea, M. J., Hanley, C. L., Ravitz, A., Labrique, A. B., & Cohen, A. B. (2019). Digital health: a path to validation. *NPJ digital medicine*, *2*(1), 38.

[7]Wang, H., Zhang, J., Luximon, Y., Qin, M., Geng, P., & Tao, D. (2022). The determinants of user acceptance of mobile medical platforms: An investigation integrating the TPB, TAM, and patient-centered factors. *International Journal of Environmental Research and Public Health*, *19*(17), 10758.

[8]曹博林,代文犊犊. 理解线上医患交流:基于"医—患—技术"三元视角透视作为传播行为的在线问诊[J]. 新闻大学,2022(11):54-66.

[9]吕艳华,王康龙,钟小云,等. 基于文本挖掘的互联网医疗平台用户画像模型构建[J]. 医学信息学杂志,2024,45(6):7-12. DOI:10.3969/j.issn.1673-6036.2024.07.002.

[10]Selvi, M., Thangaramya, K., Saranya, M. S., Kulothungan, K., Ganapathy, S., & Kannan, A. (2019). Classification of medical dataset along with topic modeling using LDA. In *Nanoelectronics, Circuits and Communication Systems: Proceeding of NCCS 2017* (pp. 1-11). Springer Singapore.

[11]Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, *19*(6), 1236-1246.

[12]Khanmohammadi, S., Adibeig, N., & Shanehbandy, S. (2017). An improved overlapping k-means clustering method for medical applications. *Expert Systems with Applications*, *67*, 12-18.

[13]尹相森,李俊儒,张云秋,等. "互联网+医疗"环境下用户行为影响机制研究——基于在线医疗平台患方用户初次使用行为的分析[J]. 价格理论与实践,2022(6):109-112,194. DOI:10.19851/j.cnki.CN11-1010/F.2022.06.298.

[14]Ogunleye, A., & Wang, Q. G. (2019). XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM transactions on computational biology and bioinformatics*, *17*(6), 2131-2140.

[15]Band, S. S., Yarahmadi, A., Hsu, C. C., Biyari, M., Sookhak, M., Ameri, R., ... & Liang, H. W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, *40*, 101286.

[16]Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, *214*, 106584.

[17]He, Y., Guo, X., Wu, T., & Vogel, D. (2022). The effect of interactive factors on online health consultation review deviation: An empirical investigation. International Journal of Medical Informatics, 163, 104781.

[18]Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia tools and applications, 78, 15169-15211.

[19]Eckhardt, C. M., Madjarova, S. J., Williams, R. J., Ollivier, M., Karlsson, J., Pareek, A., & Nwachukwu, B. U. (2023). Unsupervised machine learning methods and emerging applications in healthcare. Knee Surgery, Sports Traumatology, Arthroscopy, 31(2), 376-381.
[20]Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). Computers & Geosciences, 19(3), 303-342.
[21]McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
[22]Ostertagova, E., Ostertag, O., & Kováč, J. (2014). Methodology and application of the Kruskal-Wallis test. Applied mechanics and materials, 611, 115-120.
[23]Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. Econometric reviews, 29(5-6), 594-621.
[24]Zhang, D., & Gong, Y. (2020). The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure. Ieee Access, 8, 220990-221003.
[25]Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. Computers, Environment and Urban Systems, 96, 101845.