

基于两阶段 VaDE 聚类的跨区就诊行为分析

——以上海市 A 区为例

田文琪* 刘嘉祯 王之义 崔 欣

上海市卫生健康统计中心 上海 200040

【摘要】目的:基于门急诊数据,利用两阶段 VaDE 模型挖掘居民就诊行为的异质性结构,揭示跨区流动模式。方法:使用 2024 年上海市 A 区常住居民的门诊数据,以“每次门诊就诊记录”为分析单元,采用两阶段 VaDE 进行非线性降维与聚类,并结合跨区就诊率开展特征分析。结果:该模型聚类性能优于 K-Means 与 VAE+K-Means,识别出 7 类典型就诊模式,在跨区率、人口特征、诊疗结构和疾病谱方面差异显著。其中,簇 1 为高龄重疾型且高跨区率,簇 2 为中青年跨区通勤型,簇 3 为轻症/孕产期导向型,簇 4 为中老年低频就诊型,簇 5~7 为本地慢病或低负担型。结论:两阶段 VaDE 模型在刻画居民就诊行为异质性及识别重点跨区流动类型方面表现良好,可为推动就医服务精细化管理与区域统筹提供依据。

【关键词】两阶段变分深度嵌入聚类模型(VaDE);跨区就诊行为;变分自编码器(VAE);K-均值聚类
中图分类号:R197 文献标识码:A doi:10.3969/j.issn.1674-2982.2025.08.004

Cross-district healthcare-seeking behavior analysis using two-stage VaDE clustering approach: A case study of district A, Shanghai

TIAN Wen-qi, LIU Jia-zhen, WANG Zhi-yi, CUI Xin

Shanghai Statistics Center, Shanghai 200040, China

【Abstract】 Objective: Based on outpatient and emergency visit data, this study employs a staged VaDE model to uncover the heterogeneity structure of residents' healthcare-seeking behaviors and reveal cross-district mobility patterns. Methods: Outpatient and emergency visit records of permanent residents in District A, Shanghai, in 2024 were used. Each individual visit served as the analytical unit. A staged VaDE approach was applied for nonlinear dimensionality reduction and clustering, followed by feature analysis with cross-district visit rates. Results: The proposed model outperformed K-Means and VAE+K-Means in clustering performance, identifying seven typical healthcare-seeking patterns with significant differences in cross-district rates, demographic characteristics, care structures, and disease profiles. Cluster 1 comprised elderly patients with severe conditions and high cross-district rates; Cluster 2, young and middle-aged cross-district commuters; Cluster 3, mild-case or maternity-oriented visits; Cluster 4, infrequent visits by middle-aged and elderly patients; and Clusters 5~7, local chronic disease or low-burden visits. Conclusion: The staged VaDE model demonstrates strong capability in characterizing behavioral heterogeneity and identifying key cross-district mobility types, providing evidence for refined healthcare service management and regional coordination.

【Key words】 Two-stage VaDE; Cross-district healthcare-seeking; VAE; K-Means

1 引言

居民的跨区就诊行为往往反映出本地医疗服务

未能充分满足实际需求,提示区域间医疗资源配置存在不平衡问题。推动资源公平可及一直是政府关注的重点。尽管上海整体医疗资源水平处于全国领

* 基金项目:上海市卫生健康委员会卫生行业临床研究专项(202340272)

作者简介:田文琪(1987年—),女,硕士,中级统计师,主要研究方向为医疗卫生统计。E-mail:tianwenqi@wsjkw.sh.gov.cn

通讯作者:崔欣。E-mail:cuixin@wsjkw.sh.gov.cn

先地位,但优质资源集中在中心城区,郊区相对薄弱。^[1]深入分析郊区居民的跨区就诊现状,有助于识别医疗供需特征,为优化资源配置提供依据。

已有研究在跨区就诊行为分析方面进行了诸多探索。传统分析方法主要采用统计描述结合可视化手段,如桑基图可直观呈现跨区就诊的区域和机构分布情况。^[2-3]在行为影响因素建模方面,回归类方法被广泛应用,如Logit^[4-5]、Lasso^[6]、Probit^[7]模型,尤其是逻辑回归^[8-10],可量化变量对跨区就诊倾向的影响方向与程度。近年来,分类器模型结合特征重要性评估工具(如SHAP)也被用于医疗行为预测与解释^[11-13],但其更关注变量对预测准确率的贡献,难以揭示结构性差异。

聚类分析作为揭示对象潜在异质性的重要工具,广泛应用于疾病分型^[14]、用户画像^[15]及公共卫生决策^[16]等领域。然而,传统方法多基于原始变量或线性降维处理,难以有效刻画行为数据中存在的复杂非线性结构,从而限制了对模式特征的深度解析。变分自编码器(Variational Auto-Encoder, VAE)正好可以克服以上不足,进行非线性特征学习。其中,基于变分深度嵌入聚类模型(Variational Deep Embedding, VaDE)将VAE的非线性编码能力与高斯混合模型(Gaussian Mixture Model, GMM)的概率聚类机制相结合,能够在低维潜在空间中挖掘更为稳定且具有区分力的人群结构。

本文使用上海市A区常住居民门诊就诊数据,

利用聚类算法识别就诊模式,并对不同簇的跨区率进行分析,通过比较不同跨区率簇之间的特征差异,发现跨区就诊行为的潜在结构,为医疗资源调配与精准服务提供依据。

2 数据与研究方法

2.1 数据来源与预处理

本研究数据来源于《上海市卫生资源与医疗服务统计调查制度》中的门诊病人调查表。共获取上海市A区常住居民2024年度门急诊就诊数据约365万条,就诊人数约57万名。依据就诊发生时记录的相关信息,对原始数据进行结构化处理,整理得到39个特征变量,并将其划分为三大类(表1)。

表1 输入数据信息类别

字段类型	代表性变量示例
患者属性	性别、年龄、居住地所属街道、患有慢性病情况、截至就诊日是否跨过区
就医情况	总花费、诊断疾病系统分类
医院属性	所属区域、机构类型、机构等级

数据中仅费用字段存在缺失,且缺失率为1.8%。考虑到存留的数据量达到365万人次,样本充足且缺失比例极低,故对缺失记录采取直接剔除的方式处理。为消除变量量纲对聚类距离度量的影响,本研究对连续数值型变量进行Z-score标准化,对类别型变量进行独热编码化(表2)。在聚类阶段不放入“是否跨区”。

表2 变量预处理方法

类型	变量名	变量含义	变量取值	标准化方法
目标	y	是否跨区	是=1,否=0	无需处理
个人属性	x1	性别	女=1,男=0	无需处理
	x2	年龄	数值	z-score
	x4	高血压	是=1,否=0	无需处理
	x5	糖尿病	是=1,否=0	无需处理
诊疗信息	x8	截至就诊日是否跨过区	是=1,否=0	无需处理
	x3	门诊总费用	数值	z-score
	x6	今年至今就医次数	数值	z-score
	x7	距上次就医间隔天数	数值	z-score
就诊机构属性	x9	市属	是=1,否=0	无需处理
	x10	三级	是=1,否=0	无需处理
	x11	二级	是=1,否=0	无需处理
	x12	社区	是=1,否=0	无需处理
	x13	专科	是=1,否=0	无需处理
居住区域	x14~x23	A区的10个街道或镇 (为避免特征间相互表达,已舍去1个)	是=1,否=0	无需处理

表2 变量预处理方法 (续)

类型	变量名	变量含义	变量取值	标准化方法
就诊疾病分类	x24	恶性肿瘤	是=1,否=0	无需处理
	x25	其他肿瘤	是=1,否=0	无需处理
	x26	内分泌	是=1,否=0	无需处理
	x27	精神类	是=1,否=0	无需处理
	x28	眼疾病	是=1,否=0	无需处理
	x29	循环系统	是=1,否=0	无需处理
	x30	呼吸系统	是=1,否=0	无需处理
	x31	消化系统	是=1,否=0	无需处理
	x32	皮肤类	是=1,否=0	无需处理
	x33	骨骼肌肉类	是=1,否=0	无需处理
	x34	妊娠类	是=1,否=0	无需处理
	x35	P类疾病	是=1,否=0	无需处理
	x36	Q类疾病	是=1,否=0	无需处理
	x37	R类疾病	是=1,否=0	无需处理
	x38	Z类疾病	是=1,否=0	无需处理

2.2 VaDE 聚类方法与评价指标

为获取最优的超参数配置,本文采用两阶段的模型训练流程,第一阶段通过遍历潜在维度与聚类数的组合,选取性能最佳的维度数和聚类数;第二阶段则基于该参数进行两阶段 VaDE 模型训练。模型整体训练流程如图 1 所示。

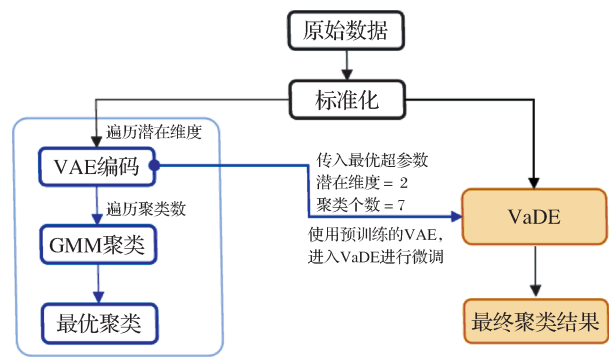


图1 模型整体训练架构

聚类完成后,选取以下三项常用指标进行聚类效果的评估:(1)轮廓系数:衡量样本在聚类中的一致性与分离度,值越高表示聚类效果越佳;(2) Davies-Bouldin 指数(DBI):衡量簇间相似度与簇内紧密度,值越小表示聚类质量越高;(3) Calinski-Harabaz 指数(CHI):反映簇间方差与簇内方差的比值,值越大表示聚类结构越显著。

2.3 模型参数设置与训练配置

为保证模型训练过程的稳定性及实验结果的可重现性,本文对两阶段 VaDE 模型进行了如下参数设

置:编码器与解码器均采用两层全连接神经网络结构,分别为“输入维度→256→潜在维度”与“潜在维度→256→输入维度”,激活函数为 ReLU。潜在空间维度设为 2,聚类类别数设为 7,相关选择依据详见 3.1 与 3.4 部分。模型训练过程中,采用 AdamW 优化器,学习率设为 1×10^{-3} ,批次大小为 1 024,权重衰减设为 1×10^{-5} ,并在梯度更新中使用梯度裁剪(阈值 1.0)以提升训练稳定性。最大训练轮数为 100,采用早停策略(patience=5)防止过度拟合。

模型初始化方面,固定随机种子为 42 以确保实验结果的可重现性。数据划分方面,训练集与验证集按 8:2 比例划分。

2.4 潜在维度敏感性分析方法

为评估潜在维度设置对模型聚类性能的影响,本研究开展了潜在维度敏感性分析实验。在保持其他超参数不变的条件下,将潜在维度分别设定为多个不同取值,并在相同训练配置下独立训练模型。

在模型训练完成后,提取各潜在维度下的样本表示向量,并采用 GMM 算法进行聚类,且对聚类性能进行评估。通过对比不同潜在维度设置下的各项聚类评价指标,确定最优的潜在维度取值,为后续实验提供合理参数设定依据。

2.5 潜在维度语义解析方法

为揭示模型中潜在空间中各维度的语义含义,本研究采用定量相关性分析与定性插值分析相结合的方式,对训练得到的潜变量进行系统解析。

2.5.1 相关性分析

在模型训练完成后,首先提取所有样本在潜在空间中的均值编码向量,并计算每一潜在维度与原始特征之间的皮尔逊相关系数。

2.5.2 插值分析

对潜在维度进行插值实验。具体方法为:在潜在空间中选定某一潜变量维度,在固定其它维度编码值的情况下,将该维度在区间 $[-3,3]$ 内按等距步长平滑插值,生成一系列潜在向量,并通过解码器重构得到对应样本的特征值。通过比较不同插值点下各特征的重构值变化趋势,分析该潜在维度对重构特征的影响程度及方向。

该方法能够定量评估潜在维度与原始特征的相关性,并结合重构变化趋势进行定性解释,从而揭示模型在潜在空间中所捕捉到的主要变异模式。

3 实验结果

3.1 潜在维度敏感性实验

按照2.4节所述的方法,对不同潜在维度设置下的模型进行了聚类性能评估。表3展示了潜在维度 $\in \{2, 5, 10, 15, 20, 25, 30\}$ 时,轮廓系数、Calinski-Harabaz指数、Davies-Bouldin指数的变化趋势。实验结果表明,当潜在维度为2时,轮廓系数和CHI均达到最大值,DBI达到最小值,聚类性能最优。

该设定兼顾信息保留与降维压缩,后续实验均采用潜在维度=2作为模型参数配置。

表3 潜在维度敏感性实验结果

潜在维度	聚类数	轮廓系数	CHI	DBI
2	29	0.304 0	3 965.53	1.00
5	28	0.107 6	622.45	1.42
10	21	0.042 0	224.85	2.22
15	20	0.015 6	125.76	2.57
20	8	0.007 4	86.23	3.11
25	25	0.010 6	69.15	3.38
30	8	0.006 8	56.26	3.73

注:聚类指标中,轮廓系数越大、CHI指数越大、DBI指数越小,聚类效果越好。

3.2 潜在维度含义解析

潜变量与原始特征的相关性分析结果(表4)显示, z_1 与年龄($r=-0.758\ 2$)、高血压($r=-0.383\ 1$)、社区机构就诊($r=-0.480\ 0$)及循环系统疾病($r=-0.379\ 4$)均呈显著负相关,提示该维度主要反映老年慢病群体在社区及基层医疗机构就诊的特征。 z_2 则与年龄($r=0.402\ 1$)及今年至今就医次数($r=0.809\ 7$)呈显著正相关,与距上次就医间隔天数呈显著负相关($r=-0.614\ 0$),表明该维度更侧重刻画就诊频率高、就诊间隔短的人群特征。部分变量在某一潜变量上的相关系数缺失(NaN),说明该维度对该特征的刻画作用不显著或数值分布趋于稳定。

表4 潜变量与原始变量的相关系数矩阵

潜变量	年龄	高血压	今年至今就医次数	距上次就医间隔天数	社区	循环系统疾病
z_1	-0.758 2	-0.383 1	NaN	NaN	-0.480 0	-0.379 4
z_2	0.402 1	NaN	0.809 7	-0.614 0	NaN	NaN

在潜变量插值分析中(图2), z_1 维度的变化主要影响变量序号2~4与8、12对应的特征,其中变量3在 z_1 取负值时重构值显著升高(峰值约4),而在正值区间则明显下降;变量6在极端正负取值下均表现出幅度较大的负向波动。相比之下,其他特征在 z_1 变化过程中波动较小,说明该维度主要刻画少数关键特征的显著差异。结合特征对应的实际意义, z_1 维度的变化对性别、年龄、高血压状态以及就医频率等特征具有显著影响。

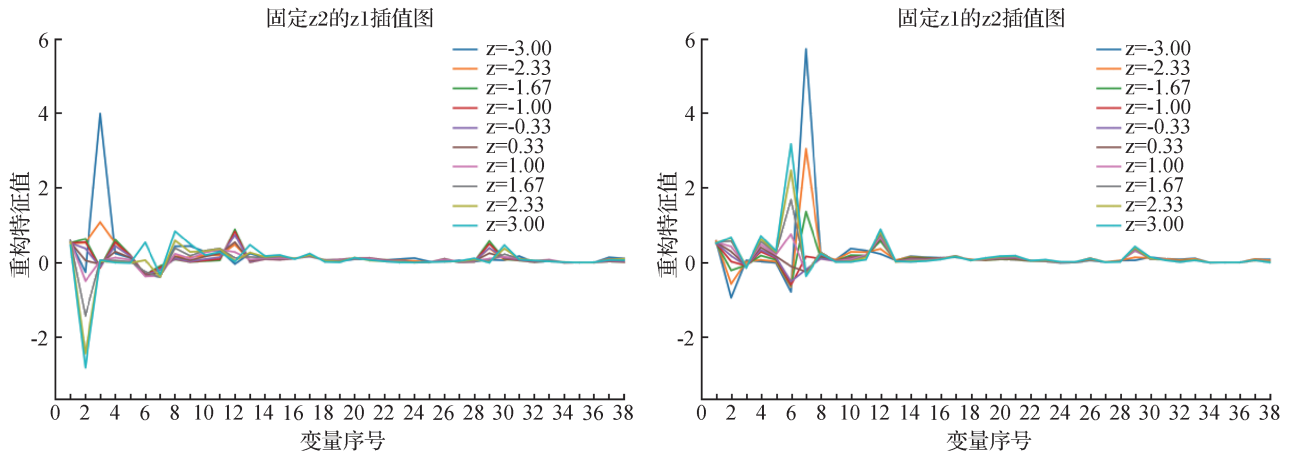
在 z_2 维度插值中,变量6、7是变化最为剧烈的特征,在 z_2 负值区间重构值显著高于正值区间;同时变量3和7在正向取值时整体重构值下降,反映该维度同样与少数高敏感度特征强相关。结合特征对

应的实际意义, z_2 维度的变化主要影响年度累计就医次数与就诊间隔天数,提示该维度高取值更多对应低频、间隔长的就诊行为模式。

潜变量的相关性分析与插值分析结果均表明,两维潜变量分别侧重于刻画患者人群属性差异(尤其是老年慢病就医)与就医行为节奏差异(高频 vs. 低频就诊)。该结果表明,在捕捉主要变异结构的要求下,2个潜变量维度已能够支撑高质量的聚类划分,为聚类结果的可解释性提供了结构化依据。

3.3 聚类方法对比实验

为验证两阶段VaDE方法的有效性,本文选取传统的K-Means聚类与基于VAE编码的K-Means聚类(VAE+K-Means)作为对照,三种方法均在相同数据



注:横轴是原始变量的序号,数字对应变量名可参照上文表2;纵轴是潜变量生成(解码)的样本特征值。

图2 潜在空间维度变化对特征重构的影响插值路径图

集上进行训练与评估。

从聚类质量指标来看,传统K-Means的轮廓系数最低(0.068 8)、DBI值最高(2.338 8),整体聚类效果最差;VAE+K-Means的轮廓系数(0.311 2)略高于两阶段VaDE(0.306 0),但在DBI上,两阶段VaDE的得分为0.841 2,低于VAE+K-Means的0.872 1,表明其在簇内紧凑性和簇间分离度方面具有优势。两阶段VaDE的最终聚类数为7类,相比VAE+K-Means的9类,具有更高的结构简洁性(表5)。

表5 方法比较结果

方法	聚类数	轮廓系数	DBI
传统K-Means	10	0.068 8	2.338 8
单独VAE+K-Means	9	0.311 2	0.872 1
本方法两阶段VaDE	7	0.306 0	0.841 2

结合本研究所关注的“能否有效识别高跨区就医人群”这一目标,对VAE+K-Means所得到的聚类结果的跨区就诊率进行分析后发现,其9个聚类的跨区率分别为:3.9%、22.6%、2.8%、6.7%、7.4%、2.4%、13.9%、8.0%、21.3%。可以看出,高跨区人群被分散在多个簇中,难以形成清晰的簇结构,不利于后续的精准管理。相比之下,两阶段VaDE所划分的聚类结构更为集中,其中簇1、簇2、簇3的跨区就诊率分别为42.0%、24.1%和22.7%,显著高于整体平均水平(8.9%),展现出更好的聚焦性与实际应用价值。

3.4 模型训练与聚类优化

从VAE预训练中可看出,10个epoch后模型基本收敛,且损失曲线变化平稳(图3),说明模型能够有效捕捉诊疗数据的潜在非线性结构,并在较短的训练周期内达到稳定状态。

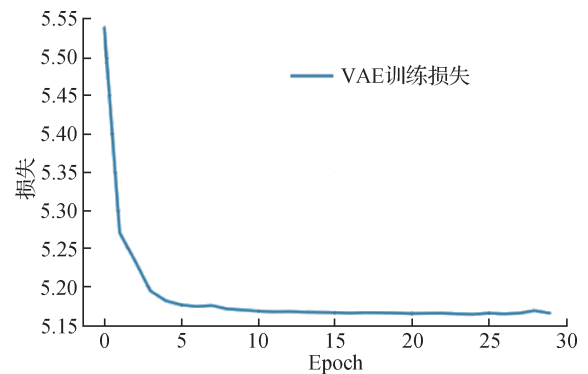


图3 VAE训练损失图

基于预训练的潜在表示,利用GMM模型对聚类数进行评估,BIC指标在聚类数为7时达到最小值(图4)。

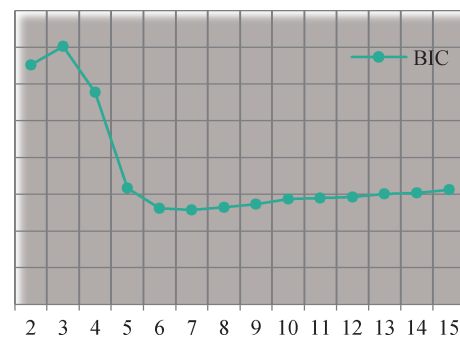


图4 聚类数和BIC结果图

在此基础上进行两阶段VaDE微调,训练过程的总损失在前期快速下降后趋于平稳,并在若干epoch出现轻微波动(图5),表明潜在聚类结构的微调对模型性能仍有细微影响,但整体波动幅度较小。总体而言,模型在预训练和微调阶段均表现出良好的稳定性和收敛性。

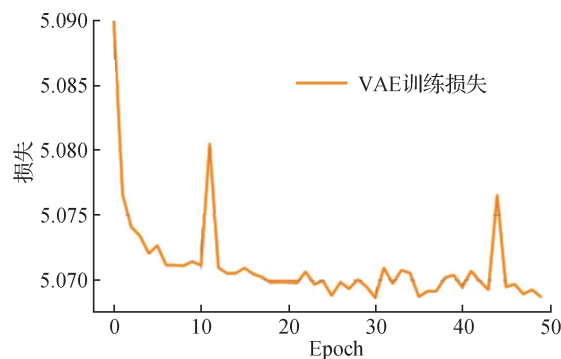


图5 VAE微调损失图

3.5 聚类稳定性分析

为评估两阶段 VaDE 模型聚类结果的稳定性,在不同随机种子下进行了5次独立运行,计算了聚类性能指标的均值、标准差和 CV 值。结果显示,轮廓系数、CHI、DBI 的 CV 值均小于 0.3(表 6),说明该算法在多次独立运行中能够保持较高的一致性与稳定性。

表 6 稳定性实验结果

运行编号	随机种子	轮廓系数	CHI	DBI
Run 1	1	0.312	4 988.348	0.826
Run 2	2	0.316	4 887.748	0.891
Run 3	3	0.304	4 778.551	0.838
Run 4	4	0.308	4 905.792	0.840
Run 5	5	0.314	4 951.847	0.822
均值	—	0.311	4 902.457	0.843
标准差	—	0.004	71.270	0.025
CV 值	—	0.013	0.015	0.030

鉴于模型在多次运行中表现较高的稳定性(各聚类指标 $CV < 0.3$),单次运行结果即可较好地代表整体性能。基于该结论,本文选取其中一次聚类结果作为后续就诊行为分析与政策建议的依据。这一做法在深度聚类实践中具有实践可行性与合理性。

3.6 结合跨区率的聚类结果

本研究基于两阶段 VaDE 模型对门诊记录进行聚类,识别出 7 类具有代表性的就诊行为簇(簇 1~7)(表 7)。各簇在跨区率、人口学特征、诊疗结构及疾病谱方面呈现出显著差异。

簇 1:高龄重疾就诊行为,强烈跨区倾向(就诊量占 0.3%)。该簇的跨区率最高,达 42.0%。其中 82.4% 的跨区就诊发生在市中心医院。平均年龄为 67.0 岁,居住区域主要集中在市中心(60.0%),高于总体水平(52.1%)。疾病谱显示:恶性肿瘤(18.6%)、

其他肿瘤(3.4%)、R 类疾病(12.1%)、Z 类疾病(11.6%)的就诊占比均为最高。次均费用为 2 342.6 元,为各簇最高。该簇可能对应于居住在市中心老年慢病患者或重疾患者,为获取更高质量的诊疗资源而倾向跨区就诊的就诊行为。

簇 2:中青年跨区通勤型就诊行为,偏好不相邻郊区机构(就诊量占 3.1%)。跨区率为 24.1%,平均年龄 34.6 岁。三级医疗机构就诊占比 33.8%,为各簇最高。专科医疗机构就诊量占比亦较高(15.6%)。主要疾病包括眼疾病(6.0%)、消化系统(12.2%)和皮肤类疾病(7.4%)。次均费用 952.0 元。该部分就诊簇主要来源居住地位于区域中间(52.2%)。其跨区行为中,前往不相邻郊区机构占比 28.2%,在各簇中位列第二,而前往市中心的占比(56.0%)低于总体(60.0%)。该簇反映出一类可能受工作位置影响的通勤背景下的就诊行为,倾向于二、三级机构。

簇 3:轻症/孕产期导向的年轻人就诊行为,兼顾医院等级与距离(就诊量占 13.0%)。该簇跨区率 22.7%,平均年龄 29.9 岁。女性占比 58%,为最高。三级医院就诊量占 33.8%、专科医院就诊量占比 15.7%,均为最高,二级医院就诊量占比也较高(35.5%)。呼吸系统疾病(27.8%)及妊娠类疾病(1.0%)为典型该簇典型疾病。需求量集中来自居住在中间区域(50.2%)。跨区行为中,前往不相邻郊区机构的比例为 34.4%,为最高,而市中心医院比例最低(49.6%)。次均费用 271.0 元。该簇体现的是一种年轻就医群体中对服务等级有所偏好但不盲目追求远距离中心机构的行为特征。

簇 4:中老年低频就诊行为,适度跨区(就诊量占 3.7%)。跨区率为 13.7%,平均年龄 62.1 岁。就诊频率偏低(均值为 5.9 次),次均费用偏高(645.2 元)。跨区行为中,前往市中心医院的就诊量占比较高(73.2%)。该簇表现出的是中老年人群中较低就医频率、但在选择跨区机构时高度集中于市中心医院的行为特征。

簇 5:高龄慢病管理型就诊行为,以本地社区为主(就诊量占 12.1%)。该簇跨区率 5.8%,平均年龄 70.8 岁,为最高。就医数频次高(22.5 次),社区卫生服务机构就诊占比 72.3%。高频就诊的疾病包括内分泌疾病(12.4%)、骨骼肌肉类疾病(6.7%)。就诊需求来自区域内侧占比 59.9%。该簇或对应高龄慢病群体的就诊,有过跨区史,就诊频次高但多在区域内社区,跨区行为有限。

簇 6:本地依赖型基础就诊行为,主力服务来源(就诊量 67.4%)。跨区率 5.6%,跨区史 13.8%。社区卫生服务中心就诊量占比 70.1%。次均费用低(184.3 元),就医频率适中(7.9 次)。该簇构成了基层医疗卫生服务的主要承载部分,其就诊行为高度依赖本地社区医疗资源。

簇 7:低频低负担封闭型就诊行为(就诊量占

0.4%)。跨区率最低(3.4%),平均年龄 66.6 岁,次均费用最低(148.9 元)。就医频率低(6.3 次),社区门诊量占比高(83.7%)。循环疾病为主要就诊疾病(57.3%)。该簇可能代表了一种低复杂度就诊需求下的本地医疗服务使用行为,对机构等级和跨区选择均不敏感。

表 7 各聚类人群特点划分(列示部分变量,以 CV 值≥0.3 为主)

簇编号	簇 1	簇 2	簇 3	簇 4	簇 5	簇 6	簇 7	整体
人次占比(%)	0.3	3.1	13.0	3.7	12.1	67.4	0.4	100.0
是否跨区(%)	42.0	24.1	22.7	13.7	5.8	5.6	3.4	8.9
个人属性								
年龄(岁)	67.0	34.6	29.9	62.1	70.8	66.9	66.6	61.4
性别(女占比%)	55.0	57.0	58.0	55.0	56.0	53.0	50.0	55.0
高血压(%)	39.5	3.3	2.9	23.7	54.3	42.0	40.5	36.5
糖尿病(%)	17.2	1.4	1.7	8.2	24.8	16.6	12.7	14.9
就诊信息								
费用(元)	2 342.6	952.0	271.0	645.2	193.4	184.3	148.9	244.5
跨区史(%)	57.6	30.8	33.4	18.9	28.7	13.8	10.6	19.0
就医次数(次)	20.3	5.1	6.2	5.9	22.5	7.9	6.3	9.3
机构属性(%)								
市属	37.8	19.6	18.7	10.8	4.1	3.8	1.9	6.6
三级	15.6	33.8	33.8	15.9	5.1	9.6	6.2	13.2
二级	43.1	32.3	35.5	23.6	18.5	16.5	8.2	20.0
社区	3.5	14.3	11.9	49.6	72.3	70.1	83.7	60.2
专科	14.5	15.6	15.7	5.9	2.3	2.5	1.2	4.8
疾病类型(%)								
恶性肿瘤	18.6	1.5	0.4	2.8	1.2	0.5	0.2	0.8
其他肿瘤	3.4	1.2	1.1	1.1	0.8	0.6	0.4	0.7
内分泌	4.5	4.4	4.2	8.0	12.4	10.4	10.2	9.6
精神类	2.2	2.1	2.4	1.5	1.4	1.2	0.8	1.4
眼疾病	2.8	6.0	5.4	3.2	2.3	2.2	1.6	2.8
循环系统	7.1	3.7	3.6	27.1	34.3	37.9	57.3	31.5
呼吸系统	4.0	19.3	27.8	9.4	10.2	12.1	6.9	14.0
消化系统	11.5	12.2	9.6	10.0	8.0	7.0	4.7	7.7
皮肤类	2.1	7.4	5.6	3.7	2.4	2.5	2.1	3.1
骨骼肌肉类	3.6	4.7	4.2	5.6	6.7	5.9	3.4	5.7
妊娠类	0.1	0.6	1.0	0.1	0.0	0.1	0.0	0.2
R 类	12.1	9.4	7.4	8.6	4.5	4.7	3.1	5.3
Z 类	11.6	9.6	9.3	5.1	2.9	3.0	1.9	4.2
居住区域(%)								
内侧	60.0	44.6	46.6	48.3	59.9	52.3	49.2	52.1
中间	34.5	52.2	50.2	46.9	33.7	41.7	45.6	42.3
外侧	5.5	3.2	3.2	4.7	6.4	6.0	5.2	5.5
跨区偏好(%)								
市中心	82.4	56.0	49.6	73.2	66.4	65.1	58.8	60.0
接壤郊区	7.5	15.8	16.0	13.4	17.4	18.4	20.7	16.8
不相邻郊区	10.1	28.2	34.4	13.4	16.1	16.5	20.5	23.1

4 结论与展望

4.1 研究结论

4.1.1 两阶段 VaDE 提升就诊行为聚类性能与可解释性

首先,两阶段 VaDE 方法在就诊行为建模方面展现出优越性能。与 K-Means、VAE+K-Means 等传统聚类方法相比,两阶段 VaDE 在聚类性能指标(如 Davies-Bouldin Index)上取得更优结果,且可生成结构紧凑、语义清晰的7类就诊行为簇,有利于结果解释与政策制定。

其次,模型能够有效提炼原始 39 维特征中的核心行为差异,仅用 2 个潜在维度即可实现高质量的聚类划分,说明就诊行为的主要变异性集中在少数关键因素上,如年龄、居住区域与疾病类别。

4.1.2 聚类揭示就诊行为类型与跨区特征

模型识别的 7 类就诊行为在跨区率、人口特征、机构等级偏好及疾病谱方面表现出显著差异,具体包括:

簇 1 为高龄重疾就诊行为,强烈跨区倾向,呈现出高费用、高等级机构偏好,集中前往市中心医院;簇 2 为中青年跨区通勤型就诊行为,偏好不相邻郊区机构,反映部分跨区发生于通勤路径上的实际需求;簇 3 为轻症/孕产期导向的年轻人就诊行为,兼顾医院等级与距离,以呼吸类、妊娠类疾病为主;簇 4 为中老年低频就诊行为,适度跨区,就医频率较低,但存在对高等级机构的倾向;簇 5 为高龄慢病管理型就诊行为,以本地社区为主,频次高,服务重心集中在基层医疗卫生机构;簇 6 为本地依赖型基础就诊行为,主力服务来源,为占比最大的簇,体现出社区卫生服务机构的广泛服务覆盖;簇 7 为低频低负担封闭型就诊行为,呈现较低就诊频次与单一病种,为行为收缩型簇。

各类就诊行为在空间分布上并未显著集中于某一街道或镇,表明在上海市 A 区范围内,各类需求广泛分布,就诊行为类型的多样性在空间上具有普遍性特征。

4.2 针对不同就诊行为簇的干预建议

4.2.1 高跨区就诊行为的管理引导

第一,对高龄重疾簇(簇 1),建议强化分级诊疗机制,依托家庭医生签约服务,引导患者在完成核心治疗后,回转本地开展康复与随访;第二,对孕产期导向簇(簇 3),可提升区内妇幼健康服务能力,构建

标准化孕产期服务路径,提升服务透明度,增强就近建档、产检吸引力;第三,对跨区通勤型簇(簇 2),可鼓励在单位附近建立健康档案,推进“通勤带”节点设置移动诊室或医联体服务点,缓解表面上的跨区就医率。

4.2.2 对低跨区簇的服务保障与防范外流

第一,对簇 5、6、7 低跨区簇,应继续提升社区服务能力、药品供应与慢病管理体系,防止“因服务不足而被动跨区”;第二,对部分倾向外流,建议对标市中心服务能力补齐短板,加强区域品牌医院建设;第三,针对居住区域偏外侧(接壤其它省份的区域)、医疗服务薄弱地带,建议探索“区级远程诊疗中心”模式,接入市属优质资源,缓解“医疗洼地”问题。

4.3 讨论

本研究所采用的两阶段 VaDE 模型体现出对就医行为的高质量识别能力、对非线性复杂数据的建模能力。相较于传统聚类方法,两阶段 VaDE 模型能在降维空间中学习出代表性更强、边界更清晰的就诊行为簇,有助于揭示就医行为的潜在结构特征。

在区域医疗资源配置层面,基于两阶段 VaDE 聚类结果识别出的跨区就诊簇,结合人口属性和空间分布等特征分析,可揭示区域内医疗资源供需之间的结构性失配问题,为补齐关键医疗服务能力短板提供数据支持,从而提升资源配置与管理决策的精准性。具体应用潜力体现在以下几个方面:

第一,识别重点服务人群:如本研究中识别出的“高龄重病高跨区簇”“跨区通勤簇”等,有助于明确对特定资源存在高度依赖的人群类型。

第二,指导资源投放策略:结合行为簇的空间分布特征,可识别区域服务能力薄弱环节,辅助决策层针对妇幼、慢病、康复等领域开展精准投入,实现“资源随需下沉”。

第三,优化分级诊疗与转诊路径:对倾向跨区就医的行为簇,探索如何通过家庭医生、医联体、远程医疗等机制进行合理引导,提升本地就医吸引力。

第四,支持动态评估与模拟优化:未来可结合多时点或时间序列数据,利用两阶段 VaDE 持续更新行为分布,为模拟资源调整后的服务改善效果提供基础。

尽管本研究在就诊行为建模和聚类分析方面取得了一定进展,但仍存在以下不足之处,有待后续研究完善:第一,特征构建维度仍待拓展:目前主要采

用诊疗行为相关特征(如就诊频次、机构类型、疾病分类等),后续可引入交通可达性等地理位置类的外部变量,提升模型的综合解释力;第二,缺乏就诊路径或序列结构建模:本文聚类以单次就诊为单位,未考虑就诊序列的演变过程。未来可结合就诊轨迹、医疗利用链等序列特征,采用序列聚类或生成模型深入刻画就诊行为模式。综上所述,两阶段 VaDE 聚类方法在刻画多样化门诊就诊需求、识别重点跨区流动类型方面表现出良好的适应性和拓展性,未来有望在居民就诊行为建模、区域医疗资源管理等实际场景中得到广泛应用。

作者贡献:田文琪负责数据分析、论文撰写与修订;刘嘉祯负责论文审校、资源协调;王之义负责资料整理、数据收集与加工;崔欣负责确定选题、框架设计与指导。

作者声明本文无实际或潜在的利益冲突。

参 考 文 献

- [1] 张中浩,孙诗萌,汪雪,等.特大城市医疗公共服务可达性的空间格局及其影响因素分析:以上海市为例[J].地理科学,2022,42(4):622-630.
- [2] 路凤,聂瀚林,董兆敏,等.2013—2022年北京地区住院患者流动就诊情况分析[J].中国卫生政策研究,2023,16(10):64-70.
- [3] 聂瀚林,石学峰,董兆敏,等.2013—2022年北京市儿科出院患者流动就医情况分析[J].中国卫生政策研究,2024,17(7):35-42.
- [4] 郑莉莉.医疗保险改变了居民的就诊行为吗?:来自我国CHNS的证据[J].财政研究,2017(2):84-97.
- [5] 李子豪.医防融合视角下的家庭医生有偿签约服务与公众就诊行为[J].江西财经大学学报,2024(4):73-86.
- [6] 胡晓辉,刘艳飞,蒋蕾.基于Post-LASSO方法的就医需求多控制变量选择[J].系统工程理论与实践,2018,38(10):2659-2667.
- [7] 周慧,吴宇涵,岳希明.医疗报销比例与居民就医行为:基于CHIP2018的实证分析[J].南开经济研究,2024(1):205-220.
- [8] 吴雪,高摘星,路凤,等.北京地区住院患者跨区就医影响因素分析[J].中国医院,2025,29(5):63-66.
- [9] 贺安琦,喻妍,郑思,等.基于分层模型流动人口慢性病患者就医行为影响因素分析[J].中国公共卫生,2022,38(1):75-79.
- [10] 朱兴华.基于Logistic回归模型的我国城市社区居民就医选择的影响因素研究[J].应用数学进展,2021,10(10):3239-3246.
- [11] 赵兴赞,张强,杨方社,等.基于XGBoost-SHAP方法的陕西省PM_{2.5}影响因素分析[J].环境科学研究,2025,(5):990-999.
- [12] 陈小昆,左航旭,廖彬,等.融合XGBoost与SHAP的冠心病预测及其特征分析模型[J].计算机应用研究,2022,39(6):1796-1804.
- [13] 陈永福,莫显博,刘冬东.单操作孔胸腔镜手术治疗纵隔肿瘤临床效果的影响因素及XGBoost模型的构建[J].微创医学,2025(1):102-107.
- [14] 张晓博,杨燕,李天瑞,等.基于医疗文本数据聚类的帕金森病早期诊断预测[J].计算机应用,2020,40(10):3088-3094.
- [15] 吴江,侯绍新,靳萌萌,等.基于LDA模型特征选择的在线医疗社区文本分类及用户聚类研究[J].情报学报,2017,36(11):1183-1191.
- [16] 徐娜,夏卫东,刘长娜,等.2016年天津市医疗卫生机构法定传染病报告特征聚类分析[J].职业与健康,2019,35(1):124-127.

[收稿日期:2025-07-07 修回日期:2025-08-14]

(编辑 刘博)