计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(6):99−106 [doi: 10.15888/j.cnki.csa.009122]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合领域知识的药店用户画像建模及应用①

杨雨晨, 李　涛, 谢君臣, 常　远

(湖北省智能信息处理与工业实时系统重点实验室 (武汉科技大学 计算机科学与技术学院), 武汉 430070)
通信作者: 李　涛, E-mail: litao@wust.edu.cn

摘　要: 用户画像是对用户形象的勾勒与描述, 现已广泛应用于睡眠会员唤醒, 用户到店预测, 个性化推荐等典型零售场景, 药品不同于普通商品, 包含较强的语义知识, 现有用户画像主要从消费属性和静态属性出发, 不能完全适用于药店销售和预测领域. 本文提出了一种针对药品领域的用户画像模型 UPP (persona of pharmacy user), 在现有画像的基础上嵌入医药知识, 利用规则, 聚类, 统计, 实体识别等方法提取慢病、疾病、特殊病类、活动敏感度、用户价值、价格偏好等新标签. 将所有标签融入一种基于聚类的群体划分方法, 形成用户画像. 实验表明, 该模型相较于现有的用户画像模型, 在消费行为预测场景下精准率提高了 13%, 更加适用于药店营销场景.

关键词: 用户画像; 群体划分; 聚类; 标签提取; 语义树

引用格式: 杨雨晨,李涛,谢君臣,常远.融合领域知识的药店用户画像建模及应用.计算机系统应用,2023,32(6):99–106. http://www.c-s-a.org.cn/1003-3254/9122.html

Modeling and Application of Pharmacy User Persona Based on Domain Knowledge

YANG Yu-Chen, LI Tao, XIE Jun-Chen, CHANG Yuan

(Hubei Provincial Key Laboratory of Intelligent Information Processing and Real-time Industrial System (College of Computer Science and Technology, Wuhan University of Science and Technology), Wuhan 430070, China)

Abstract: The user persona is a sketch and description of the user image, which has been widely used in typical retail scenarios such as the wake-up of sleeping members, prediction of users arriving at the store, and personalized recommendations. Drugs are different from ordinary commodities, and they contain strong semantic knowledge. The existing user persona mainly starts from the consumption attribute and static attribute and is not completely applicable to the pharmacy marketing and prediction field. This study proposes a persona of pharmacy user (UPP) model for the drug field, which embeds medical knowledge on the existing persona and uses methods such as rules, clustering, statistics, and entity recognition to extract new labels including chronic diseases, diseases, special diseases, activity sensitivity, user value, and price preference. All labels are integrated into a clustering-based group division method to form the user profile. The experiment shows that the accuracy of this model is 13% higher than the existing user persona model in the consumer behavior prediction scenario, so the proposed model is more suitable for the pharmacy marketing scenario.

Key words: user persona; group division; clustering; label extraction; semantic tree

用户画像在现有零售领域已经得到广泛的应用, 成为刻画用户形象和行为习惯, 进行精准营销的重要研究方向. 用户画像可以应用于睡眠会员唤醒, 用户到店预测, 个性化推荐等场景. 宋美琦等[1]、李锐[2]、汪倩等[3]、刘学太等[4] 从用户画像的内涵定义、模型方法和应用领域等角度对用户画像的研究做了较详细的综述.

不同于普通商品, 药品具有较强的领域知识, 涵盖了疾病特征、适用范围、配伍禁忌、不良反应、用药

① 基金项目: 国家自然科学基金 (61702383); 湖北省教育厅重大项目 (17ZD014); 武汉市重点研发计划 (2022012202015070)
收稿时间: 2022-11-27; 修改时间: 2023-01-06; 采用时间: 2023-01-13; csa 在线出版时间: 2023-04-14
CNKI 网络首发时间: 2023-04-18

© 中国科学院软件研究所 http://www.c-s-a.org.cn

周期等丰富的语义, 现有的用户画像建模方法主要通过用户静态属性和消费行为数据刻画用户特征, 不能满足药店运营的需要.

本文针对上述问题, 提出了一种针对药品领域的用户画像模型 UPP (persona of pharmacy user). 采用实体识别模型挖掘出文本中的药品、疾病、症状实体, 并利用规则, 聚类, 统计等方法提取新的标签, 在现有画像的基础上生成了新的嵌入医药知识的疾病, 慢病, 特殊病类, 结合用药周期的近期价值, 药品价格偏好, 活动敏感度等标签. 提出一种将标签进行融合聚类的方法, 用于实现用户群体的划分, 形成用户画像. 实验表明, 该模型相较于现有的用户画像模型, 在典型应用场景上具有更好的效果, 更加适用于药店营销场景.

1 相关研究

用户画像的核心是用户数据的描述和分析, 而大数据、深度学习等技术的兴起, 在用户画像的内涵、标签体系、特征建模、领域应用上带来了新的机遇和挑战.

大数据驱动用户画像标签体系由简单走向复杂. 如: 王冬羽[5]、杨欧亚等[6]、李斯[7] 则通过结合实际的业务提出不同的多层次构建模型. 尤明辉等[8]、吴彦文等[9]、Chikhaoui 等[10]、Francisco 等[11] 提出用户画像应该包含用户的行为特征, 刘海鸥等[12]、安璐等[13] 提出融入情境属性, 把用户的自然属性、社交属性、能力属性等融入用户画像中.

用户画像特征建模方法由基本的统计模型逐步转向机器学习和深度学习为主的隐式特征挖掘. 从早期的简单的统计到现在的基于深度学习技术的应用, 国内外越来越多的研究者通过借助机器学习的方法构建用户画像. Zigoris 等[14] 通过使用贝叶斯模型学习用户的兴趣特征, 以此构建用户画像, 最终将用户画像特征应用于物品推荐上, 有效地缓解了冷启动问题; 赵建建[15] 运用 TF-IDF 算法计算标签的权值来构建用户画像, 巨星海等[16]、张秋平[17] 立足于实体识别的技术构建用户画像. 邹京甫[18] 提出了基于文本语义规则的实体抽取方法, 通过挖掘药品说明书中的实体构建用户画像的标签, 为线下药店零售领域的用户画像研究打开了思路.

用户画像概念衍生到群体, 形成群体的划分, 围绕产品或活动, 对人群进行细分, 勾勒目标人群的细致轮廓. Massanari[19] 将用户画像定义为一个利用用户群体的属性特征、兴趣偏好等用户标签来构建的用户群体模型; Iglesias 等[20] 通过聚类方式展示各个群组画像; Wang 等[21] 根据用户的点击行为来获取用户群体行为, 以此构建了基于在线用户行为的点击流群体模型; 周林兴等[22] 利用大数据技术捕获群体的兴趣偏好、属性特征以及需求概况, 然后建立了服务于用户画像的信息服务路径.

用户画像逐渐融合领域知识, 为不同领域的研究提供全方位、整体性, 领域知识的挖掘. 刘宝等[23] 通过知识图谱建立化学品危险评估知识图谱用于有效管理和应用化学品信息. 李阳等[24] 从知识关联的角度出发构建医药知识图谱, 支持智慧医疗的发展方向. 韩珊珊等[25] 结合信息提取方法对雷公藤安全性研究的现状及发展趋势进行分析, 以发现研究热点、演化路径及发展趋势.

用户画像可以用于实现精准营销, 一个典型的应用就是到店预测. 预测算法逐渐由传统的单模型算法研究过渡到集成模型算法研究. Dahiwade 等[26] 提出了基于患者症状的一般疾病预测模型, 采用 K 近邻和卷积神经网络机器学习算法来准确预测疾病. 贾志强等[27] 建立了针对用户消费决策的混合预测模型.

2 UPP 用户画像模型构建

对于大型连锁药店, 进行用户画像对效精准度要求很高, 本文提出的用户画像构建模型流程如图 1.

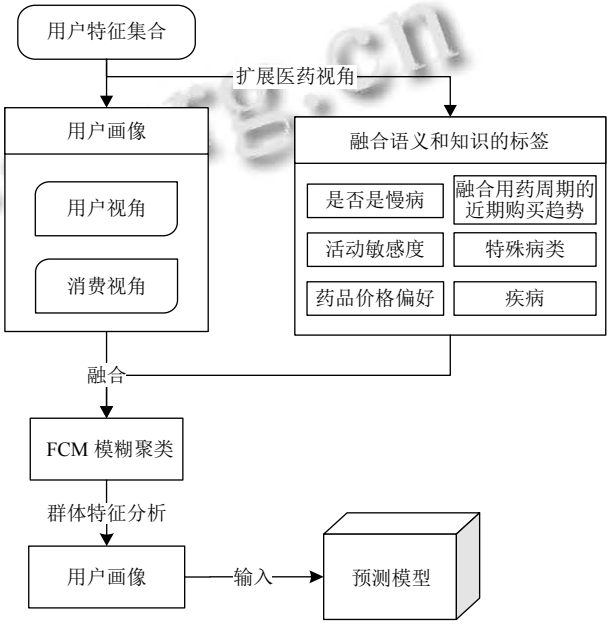

图 1 UPP 用户画像模型

2.1 融合药品属性与消费行为的画像标签设计

药店会员画像与现有用户画像在消费行为, 用户

© 中国科学院软件研究所 http://www.c-s-a.org.cn

属性等特征上存在很大重合, 但他又存在其商品特殊性与专业性, 所以与现有用户画像模型具有一定差异性. 基于医药大数据药店商品的特殊性以及用户活动特点, 本文在用户画像标签体系基础上进行了新的标签设计, 加入疾病标签, 融合用药周期的近期购买趋势标签, 特殊病类标签, 慢病标签形成多视域画像模型.

用户画像的基本模型构建参考文献 [28], 本文在原有画像模型中加入医药主题, 基于医药视角对原有模型进行扩充. 用户画像标签体系如表 1.

表 1 用户画像标签体系

主题	标签	描述
医药维度	HAV__RESUM	是否是医保用户
	BRANG__FAV	品牌偏好
	SENSTY2ACTY	活动相应敏感度
	DISEASE	疾病
	MEDIC_CLASS	疾病大类
	IS_OTC	是否常购处方药
	SPECIAL	特殊病类
	MEDIC_TYPE	药品种类
	VALS	融合用药周期的近期购买趋势标签
会员维度	SEX	性别
	CREATE_TIME	注册时间
	POINT	会员卡积分
	AGE	年龄
	ADDRESS	地址
	REG_DAYS	注册至今天数
	CARD_TYPE	会员卡类型
	IS_SMS	是否能发送短信
	PAY_WAY	支付方式
	DEPART	归属门店
	PHARMACIST	归属药师
消费维度	CONSUME_LAST	最近一次消费金额
	TIME_LAST	最近一次消费时间
	CONSUME_1_MONTH	近1个月消费金额
	TIMES_1_MONTH	近1个月消费次数
	COMSUME_3_MONTH	近3个月消费金额
	TIMES_3_MONTH	近4个月消费次数
	CONSUME_6_MONTH	近6个月消费金额
	TIME_6_MONTH	近6个月消费次数
	CONSUME_1_YEAR	近1年消费金额
	TIMES_1_YEAR	近1年消费次数
	TIMES	平均每月消费次数
	CONSUME	平均每月消费金额
	CONSUME_TIME	主要消费时间
	BUY_LAST_DRUG	最近一次购买药品
	BUY_MOST_DRUG	购买次数最多的药品
	PAY_SUM	累计支付次数
	MCARD_SUM	使用医保卡支付次数
	COMSUME_SUM	累计贡献额
	CIRCLE	生命周期阶段

药品本身蕴含着极强的知识专业性, 基于实体识别提取出其包含的知识. 本文将药品知识与消费行为标签进行融合, 进行新的标签设计.

药品本身带有一定的属性, 通过用户的购买行为, 部分属性会转移给用户. 在药品属性的提取和融合过程中, 需要基于药师和医学专家的经验, 满足一定的规则. 本文采用实体识别技术提取药品的疾病属性, 是否是慢病属性, 是否是处方药属性, 用药周期属性, 药品类别属性, 适用人群属性.

定义 1. 是否为慢病患者 *IS_CHRONIC*

慢病购药具有很强的持续性和规律性, 是否是慢病患者对于药品推荐有很重要的意义, 慢病标签基于规则定义. 令 *HBC*(*Med*): *Med* 为所购药品集合, 属于慢性药品合集, *ts* 为购药次数.

是否为慢病患者的规则如式 (1).

$$IS\_CHRONIC = HBC(Med) \vee (ts \geqslant 2) \tag{1}$$

定义 2. 疾病标签 *DISEASE*

药品所属疾病由实体识别技术提取, 利用药品说明书, 用户电子病例等非结构化文本数据.

本文采用 ALBERT+BiLSTM+CRF 神经网络模型挖掘出文本中的药品、疾病、症状实体, 作为医药领域的标签. 在传统的 BiLSTM-CRF 模型为核心的基础上, 结合 ALBERT 中文预训练模型, 将输出层词向量作为 BiLSTM 网络的辅助分层输入, 依靠网络层的主分类模型捕获序列的有效信息, 采用 CRF 模型的目标是通过邻近实体的关系获得一个最优的预测序列, 提取除药品疾病对应关系. 大致提取流程如图 2.

疾病标签由用户购买的药品提取出的疾病症状标签转移给用户, 疾病标签并不是所有疾病都具有标识意义, 所以疾病标签分为长期标签和短期标签. 短期标签设有一定的时间阈值, 例如感冒标签一般在 10 天左右就会自动取消. 长期疾病标签则长期持有. 疾病标签的规则如式 (2).

$$DISEASE = \begin{cases} LONG\_DIS, DISEASE \notin IS\_CHRONIC \wedge \\ \qquad DISEASE \notin SPECLAL \\ SHO\_DIS \text{ with } TIME, \text{ else} \end{cases} \tag{2}$$

定义 3. 融合用药周期的近期价值标签 *VALS*

部分药品具有购买周期较为规律的特点, 所以将提取出的购药周期与反应消费价值的 RFM 模型相结合, 提取用户近期价值标签 *VALS*. RFM 模型只需客户

© 中国科学院软件研究所 http://www.c-s-a.org.cn

交易数据, 容易收集, 适应零售行业特点. R (recency) 为最近一次消费时间, F (frequency) 为消费频率, 用观测期顾客消费的总次数代替, M (monetary) 为观测期消费金额. 药品零售中有部分群体会周期性购药, 例如受慢性病用药周期, 医保发放周期等因素的影响, 部分特征会呈现周期性变化, 因此在模型构建时进行优化, 加入 C (cycle) 表示为距离购药周期的下一次可能购入时间距离现在时间的时间差, C 的计算方式如式 (3).

$$C = |(t_l + T) - t_n| \tag{3}$$

其中, t_l为用户上次购买时间, t_n为现在的时间, T 为一个周期, 计算方式如式 (4).

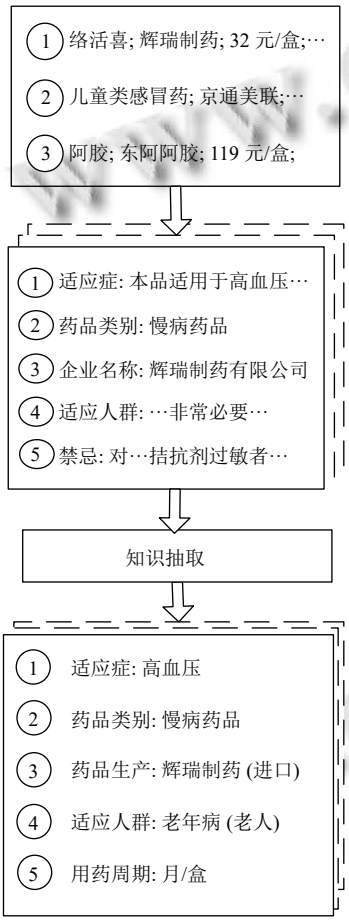

图 2 基于实体识别的药品疾病标签提取流程图

$$T = \begin{cases} \dfrac{t_n - t_f}{n-1}, & n \neq 1 \\ \dfrac{\sum_{i=1}^{m} T_i}{m}, & n = 1 \end{cases} \tag{4}$$

其中, t_f为用户第 1 次购入该商品的时间, n 为购买次数, 当该用户的购买次数为 1 时, T 为该药品所有非购买一次的用户的平均购买周期. 将 R, F, M, C 按如下公式进行标准化.

标准化公式: F、M 用式 (5), R、C 用式 (6).

$$N_{ij} = \frac{N_{\max} - N_j}{N_{\max} - N_{\min}} \tag{5}$$

$$N_{ij} = \frac{N_j - N_{\min}}{N_{\max} - N_{\min}} \tag{6}$$

其中, N_j为第 j 项指标值, $N_{\max}$为第 j 项指标的最大值, 同理, $N_{\min}$为最小值, N_{ij}为标准化值.

$$p_{ij} = \frac{N_{ij}}{\sum_{i=1}^{m} N_{ij}},\ 0 \leqslant p_{ij} \leqslant 1 \tag{7}$$

熵值法确定各指标的权重计算公式如式 (8).

$$W_j = \frac{1 + k \times \sum_{i=1}^{m} p_{ij} \ln p_{ij}}{\sum_{i=1}^{m} \left(1 + k \times \sum_{i=1}^{m} p_{ij} \ln p_{ij}\right)},\ k = \frac{1}{\ln m} \tag{8}$$

最终基于 RFM 改进的 RFMC 模型表示为式 (9), W_j为第 j 项的权重. 客户的价值与 R、C 成反比, 与 F、M 成正比.

$$VALS = \frac{W_1 F + W_2 M}{W_3 R + W_4 C} \tag{9}$$

定义 4. 特殊病类标签 *SPECIAL*

是否有老年病/儿童病/职业病标签 *GERIAT/CHD/OCUPAT*.

根据药品购买情况及其年龄特征, 特殊病类标签的规则如式 (10). 令 *HBCHD*(*Med*): *Med* 为所购药品集合, 属于儿童用药药品合集, *TS*(*Med*) 为购药次数, *TS_Max*(*Med*) 某一种药的购药次数.

$$SPECLAL = \begin{cases} GERIANT,\ DISEASE \subset IS\_CHRONIC \wedge \\ \qquad age > 50 \\ OCUPAT,\ TS\_Max(Med) \geqslant 3 \wedge age < 50 \end{cases} \tag{10}$$

定义 5. 药品价格偏好标签 *PROFIT*

用户对药品的价格偏好在消费领域本来就具有一定的意义, 在药品这个特殊领域, 同一种药品往往具有多个不同品牌, 不同价格的商品, 有些人认为大品牌值得信赖, 有些人认为都通过了药品质量检查都具有相同的作用, 有时候也会反应药店消费者的经济水平. 高

© 中国科学院软件研究所 http://www.c-s-a.org.cn

价药品偏好客户主要为利润较高和偏好买同种较贵药品的顾客. 规则为式 (11).

$$PROFIT=\begin{cases} HIGH\_PRO, \left(\dfrac{Price-avg(Meds)}{avg(Meds)}>0.5\right)\vee \\ \qquad\qquad \left(\dfrac{Profit}{Price}>0.5\right) \\ LOW\_PRO, \text{ else} \end{cases} \tag{11}$$

定义 6. 活动敏感度标签 *ACTIVE_SEN*

活动敏感度可以反映用户愿意响应活动的程度, 故采用 K-means 聚类将消费行为相似的用户群体用购买特征作为衡量指标打上活动敏感度标签, 再使用实际活动响应情况进行调整, 具体如算法 1.

算法 1. 活动敏感度标签提取算法

输入: 训练数据 *X*、簇数目 *K*、训练数据样本数目 *m*、最大的更新次数 *max_iter*
输出: 每个样本的活动敏感度标签

1. *X.ACTIVE_SEN* = LOW_QUA //初始化所有样本的敏感度标签都为低敏感度
2. centers = *X*[:*K*]/ /随机产生 *K* 个样本作为簇中心点
3. *num_iter* = 0
4. while *num_iter* < *max_iter* and changeFlag = true do //当小于迭代次数且任意一个点的簇分配结果发生改变
5. *num_iter* += 1 //迭代次数增加
6. for *i* in *X*: //对数据集中每个数据点
7. if *i.buy_count* > 1:
8. *i.ACTIVE_SEN* = HIGH_QUA //对于活动响应次数超过 2 次的样本, 直接设为高活动敏感度标签
9. center_idx = calc_min_dist_center(*i*, centers) //计算当前样本到 *K* 个簇中心点的距离, 并获取距离最小的簇对应的下标
10. center_2_dict[center_idx].append(*i*) //认为当前样本属于该簇
11. if *i.cluster* != center_idx :
12. changeFlag = false;
13. end if
14. *i.cluster* = center_idx //记下样本点所属的簇
15. for *i* in *K*: // 基于簇的隶属关系更新簇中心点
16. *tmp_x* = *X*[center_2_dict[*i*]] //获取当前簇的所有样本
17. centers[*i*] = np.mean(*tmp_x*) //求均值
18. for *i* in *K*: //对每一个簇计算平均消费价格, 平均消费次数, 参与活动 1 次的人数, 参与活动 2 次以上的人数
19. act_cluster1, act_cluster2 //初始化
20. point_avg[*i*] = center_2_dict[*i*].price.sum()/center_2_dict[*i*].times.sum() //平均消费价格/平均消费次数
21. act_1_count = calc_1_activePerson(center_2_dict[*i*])/center_2_dict[*i*].size()//参与活动 1 次的人数/该簇样本数
22. if point_avg < min_avg : //找出平均消费金额低且购买次数多的客户群体
23. act_cluster1 = *i*
24. min_avg = point_avg
25. if act_1_count > max_count: //找出之前活动响应次数较多的客户群体
26. act_cluster2 = *i*
27. max_count = act_1_count
28. Make_high_qua(center_2_dict[act_cluster1]) //将找到的群体打上高活动敏感度标签
29. Make_high_qua(center_2_dict[act_cluster2])
30. return *X*, *X.ACTIVE_SEN*//返回每个样本的活动敏感度标签

2.2 一种基于语义树的模糊聚类方法

根据设计的标签收集数据, 对提取出的标签进行聚类, 实现群体划分. 将所属群体及其隶属度作为特征输入预测模型. 具体步骤如下.

(1) 特征选择和形成特征向量.

会员主题和消费主题的特征基于随机森林的特征重要性排序, 选择其中较为重要的特征, 融合医药领域的疾病标签, 特殊病类标签, 是否是医保用户特征, 融合用药周期的用户价值标签. 对于所有的数值型特征依据式 (5) 和式 (6) 进行标准化.

(2) 利用 FCM 聚类的软化分优势, 对用户进行 FCM 聚类. 隶属度特征往往作为权重应用在各类场景之中.

将用户的向量矩阵$X=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}$ (x_i为第 i 个用户的特征向量) 作为 FCM 模糊聚类的输入, 在满足式 (12) 隶属度需要的条件下, 调整最大迭代次数, 目标函数为式 (13). 其中, c_i为第i个聚类中心; u_{ij}表示c_j对应的隶属度; k为聚类数目; N为样本总数; x_i代表样本中第i个数据; m为模糊参数; 隶属度矩阵$U=\left[u_{ji}\right]$的大小为$N\times k$; $Dis_{(x_i,c_j)}=\left\|x_i-c_j\right\|$表示第$j$个数据与第$i$个聚类中心的距离, 数值型数据就是欧式距离.

$$\sum_{i=1}^{k}u_{ij}=1,\ j=1,2,\cdots,N \tag{12}$$

$$J(U,c_1,c_2,,\cdots,c_k)=\sum_{i=1}^{n}J_i=\sum_{i=1}^{k}\sum_{j=1}^{N}u_{ij}^{m}d_{(x_i,c_j)}^{2} \tag{13}$$

由于疾病标签不是数值型, 无法根据标签直接进行距离的计算, 鉴于此, 本文提出使用基于语义树路径的相似度计算方式.

假设树结构每层的病类相互独立, 并且子树属于父节点, 以两节点到公共父节点的带权路径长度为距离, 计算相似度, 当某一药品实体同时存在于两个及以上的父类时, 取另一实体与该实体所有组合的加权路径的平均值为最终距离结果. 疾病语义树的节点基于

© 中国科学院软件研究所 http://www.c-s-a.org.cn

实体提取技术[28] 提取病类实体, 由药师参与构建. 树结构的部分节点如图 3.

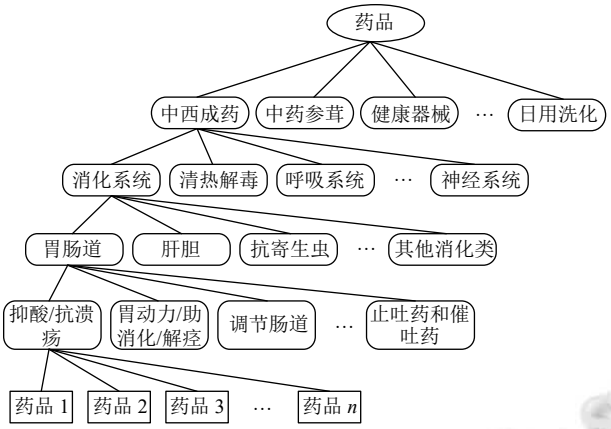

图 3 药品语义树

两实体的距离定义如式 (14) 和式 (15), 其中, $H(E)$ 为实体 E 在病类树的层高, $Father(E_i,E_j)$ 为实体 E_i 和 E_j 的最近公共父节点. 例如药品 1 和药品 2 的最近公共父节点为抗酸及抗溃疡药物, 则他们之间的距离为 2.

$$Dis\left(E_i,E_j\right)=Dis\left(E_i,Father\left(E_i,E_j\right)\right)+Dis\left(E_j,Father\left(E_i,E_j\right)\right) \tag{14}$$

$$Dis\left(E_i,Father\left(E_i,E_j\right)\right)=H\left(E_i\right)-H\left(Father\left(E_i,E_j\right)\right) \tag{15}$$

FCM 聚类过程如算法 2.

算法 2. FCM 模糊聚类算法

输入: 用户特征向量矩阵 X, 模糊参数 m, 聚类数 k, 最大迭代次数 $Iterator\_Max$.

输出: 聚类结果及隶属度矩阵 U.

1. $U^{(0)}=\{\}$//初始化隶属度矩阵 U
2. step = 0 //初始化
3. While ($\|U^{(t)}-U^{(t-1)}\|<\varepsilon$ and step < $Iterator\_Max$) do
3. update c_j //按照式 (16) 更新聚类中心:

$$c_j=\frac{\sum_{i=1}^{N}u_{ij}^{m}x_i}{\sum_{i=1}^{N}u_{ij}^{m}},\quad i=1,2,\cdots,k \tag{16}$$

4. update $U^{(t)}=[u_{ij}]$ //按照式 (17) 更新隶属度矩阵:

$$u_{ij}=\left(\sum_{t=1}^{k}\left(\frac{\|x_i-c_j\|}{\|x_i-c_t\|}\right)^{\frac{2}{m-1}}\right)^{-1},\quad i=1,2,\cdots,N \tag{17}$$

5. step++;
6. end while
7. end

(3) 针对每个群体的特征, 形成用户画像.

3 实验分析

实验运行环境: Windows 11 操作系统, 16 GB 内存, 3.5 GHz 四核心处理器, 实验软件为 Python 3.7.

3.1 数据预处理

本文实验医药数据来自武汉线下某大型连锁药店平台所提供的某一个地区连锁的历史消费数据, 主要包括药店会员注册信息约 139 801 条、会员 3 个月购买药品行为数据约 1 000 000 条. 按照每个人的唯一标识会员卡号作为一条记录的键值, 以某药品活动期间会员是否到店消费作为正负样本判断依据.

3.2 评价指标

将预测结果分类为真正类 (TP): 进行活动推送且到店消费的会员; 真负类 (TN): 未推送且未消费的睡眠会员; 假正类 (FP): 进行活动推送但实际上未消费的会员; 假负类 (FN): 未推送但到店消费的会员.

对于预测实验, 本文所讨论的单品活动到店预测场景是将会员在收到活动推送后是否到店消费的问题抽象为一个二分类预测问题, 为提高实验可信度, 本文采用准确率 (Accuracy)、召回率 (Recall) 和 $F1$ 三个值来评估模型.

3.3 预测效果对比实验

活动药品的目标人群筛选问题, 可以看作一个二分类问题. 在大量用户中挖掘出活动商品可能到店的目标用户, 需要选择一个分类精准度高的识别模型, 本文将基于用户画像的目标人群到店预测作为用户画像的一个应用场景.

为降低单个模型的泛化误差, 本文采取基于 Stacking 策略的集成学习模型, 利用多个学习器完成学习任务, 再通过元模型对一级结果进行融合. 预测模型为 Stacking 两层模型, 第 1 层为 5 种基分类器, 第 2 层为逻辑回归. 具体流程如图 4 所示.

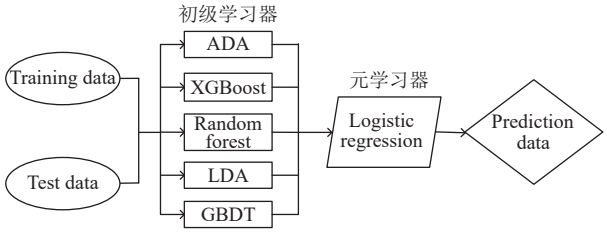

图 4 基于 Stacking 的预测模型

本文在单品活动到店预测场景下, 对比现有用户画像特征预测与本文提出 UPP 用户画像模型加入新

© 中国科学院软件研究所 http://www.c-s-a.org.cn

的特征预测目标群体, 观察活动商品的用户到店预测准确率与模型执行效率.

预测模型效果对比, 本文选择基于 Stacking 策略的集成学习模型, 采用 10 折交叉的方法, 将数据集按照 8:2 的比例进行 10 次随机划分, 取平均. 将本文选取集成学习模型与其包含的单个模型进行对比, 结果表明基于 Stacking 策略的集成学习模型效果要优于单个的机器学习模型, 实验结果如图 5.

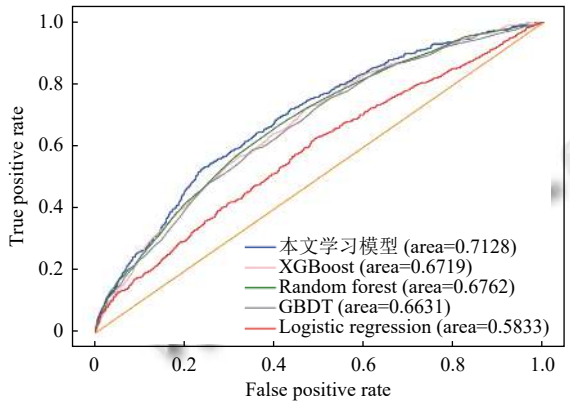

图 5 预测模型 ROC 曲线对比图

将现有画像标签拆分为特征, 并筛选重要特征用于预测实验. 经过信息增益选择 IG>0.033 的特征作为显著特征. 为验证画像扩充的新特征的有效性, 将新的画像特征集分组处理. 对于相同的会员数据集, 对比实验所使用特征集合如表 2 所示, 实验结果表示为表 3.

表 2 对比实验特征集解释

特征集合	解释
A	现有用户画像预测
B	现有用户画像+DISEASE标签转向量作为特征预测
C	现有用户画像+VALS分类扩展为特征预测
D	现有用户画像+IS_CHRONIC+SPAECIAL作为特征预测
E	现有用户画像+PROFIT+ACTIVE_SEN标签作为特征预测
F	UPP用户画像特征预测

表 3 对比实验预测结果

对照组	Accuracy	Recall	$F1$
A	0.754	0.526	0.577
B	0.810	0.585	0.591
C	0.790	0.542	0.580
D	0.783	0.530	0.565
E	0.780	0.557	0.580
F	0.887	0.619	0.608

B 组在 A 组的基础上加入了分类处理后疾病向量特征, 在预测结果上提升了 5.6% 的精确率, 并且召回率和调和平均值也有所提高, 证明疾病向量在预测中起到了较显著的作用. C 组特征加入融合了用药周期的用户近期购买趋势特征, 精准率较不加入提升了 3.6%. D 组加入了疾病的两个较粗粒度的标签, 从是否为慢病、老年病、儿童病、职业病方向扩充特征维度, 准确率较不加入提高了 2.9%. E 组特征加入是否为高价药品偏好特征和是否为活动高敏感度会员两个对消费属性扩充的特征, 因为粒度较粗, 精确率较不加入略有提升. 而合并 BCDE 四组特征后, F 组特征的精准率较现有用户画像特征有了显著的提高, 提高了 13.3%, 实验结果表明, 在活动药品到店预测这个典型场景中, 本文提出的 UPP 用户画像模型有着较好的应用效果.

4 结束语

本文针对医药大数据具有特殊的专业性的特点, 运用多种标签生成算法, 加入知识的嵌入, 从个人和群体视角对现有用户画像特征维度进行了扩充, 提出面向大型连锁药店会员的用户画像建模方法, 在此基础上结合集成模型预测算法对画像精确度进行验证, 为数据驱动情境下消费预测、精准营销等场景提供更精准的决策支持. 实验证明, 本文提出的 UPP 用户画像模型对用户刻画更精准, 基于画像的预测具有较高的准确率.

参考文献

1 宋美琦, 陈烨, 张瑞. 用户画像研究述评. 情报科学, 2019, 37(4): 171–177.
2 李锐. 用户画像研究述评. 科技与创新, 2021, (23): 4–9, 12.
3 汪倩, 徐勇, 张心蕊, 等. 用户画像研究进展综述. 现代计算机, 2020, (24): 60–63.
4 刘学太, 李阳, 巴志超, 等. 数据驱动环境下数据画像若干问题探析. 情报理论与实践, 2022, 45(4): 87–94.
5 王冬羽. 基于移动互联网行为分析的用户画像系统设计 [硕士学位论文]. 成都: 成都理工大学, 2017.
6 杨欧亚, 龚婕, 魏松杰. 面向互联网上网服务行业的用户画像系统设计. 计算机与数字工程, 2021, 49(9): 1782–1787.
7 李斯. 大数据背景下面向运营商精准营销的用户画像研究 [硕士学位论文]. 大连: 大连理工大学, 2019.
8 尤明辉, 殷亚凤, 谢磊, 等. 基于行为感知的用户画像技术. 浙江大学学报 (工学版), 2021, 55(4): 608–614, 638.
9 吴彦文, 刘雪纯, 杜嘉薇, 等. 融合微观行为特性的用户画像增强研究. 情报科学, 2021, 39(3): 19–24, 50.
10 Chikhaoui B, Wang SR, Pigot H. Causality-based model for

© 中国科学院软件研究所 http://www.c-s-a.org.cn

user profile construction from behavior sequences. Proceedings of the 27th IEEE International Conference on Advanced Information Networking and Applications (AINA). Barcelona: IEEE, 2013. 461–468.

11 Francisco M, Castro JL. A fuzzy model to enhance user profiles in *microblogging* sites using deep relations. Fuzzy Sets and Systems, 2020, 401: 133–149. [doi: 10.1016/j.fss.2020.05.006]

12 刘海鸥, 李凯, 姜波. 移动图书馆推荐系统中的用户画像与资源画像情境化融合研究. 图书馆, 2021, (6): 66–71, 93.

13 安璐, 胡俊阳, 李纲. 突发事件情境下社交媒体高影响力用户画像研究. 情报资料工作, 2020, 41(6): 5–16.

14 Zigoris P, Zhang Y. Bayesian adaptive user profiling with explicit & implicit feedback. Proceedings of the 15th ACM International Conference on Information and Knowledge Management. Arlington: ACM, 2006. 397–404.

15 赵建建. 基于数据驱动的图书馆用户画像模型构建方法研究. 新世纪图书馆, 2021, (10): 43–49.

16 巨星海, 周刚, 王婧, 等. 用户画像构建技术研究. 信息工程大学学报, 2020, 21(2): 242–250.

17 张秋平. 基于 BERT 的用户画像算法研究 [硕士学位论文]. 广州: 广东工业大学, 2020.

18 邹京甫. 面向线下医药零售的精准推荐技术研究 [硕士学位论文]. 长沙: 湖南大学, 2018.

19 Massanari AL. Designing for imaginary friends: Information architecture, personas and the politics of user-centered design. New Media & Society, 2010, 12(3): 401–416.

20 Iglesias JA, Angelov P, Ledezma A, *et al*. Creating evolving user behavior profiles automatically. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(5): 854–867. [doi: 10.1109/TKDE.2011.17]

21 Wang G, Zhang XY, Tang SL, *et al*. Clickstream user behavior models. ACM Transactions on the Web, 2017, 11(4): 1–37.

22 周林兴, 林腾虹. 用户画像视域下智能化档案信息服务: 现状、价值、运行逻辑与优化路径. 档案学研究, 2021, (1): 126–133.

23 刘宝, 车礼东, 黄红花, 等. 基于自然语言处理 (NLP) 技术建立化学品危险评估知识图谱的研究. 计算机与应用化学, 2018, 35(7): 605–610. [doi: 10.16866/j.com.app.chem20181807010]

24 李阳, 杜睿山, 张豪鹏. 面向医药信息的知识图谱构建. 计算机技术与发展, 2022, 32(10): 189–193. [doi: 10.3969/j.issn.1673-629X.2022.10.031]

25 韩姗姗, 代彦林, 丁樱, 等. 雷公藤毒性研究进展的 CiteSpace 知识图谱分析. 中国中药杂志, 2022, 47(4): 1085–1094. [doi: 10.19540/j.cnki.cjcmm.20211108.501]

26 Dahiwade D, Patle G, Meshram E. Designing disease prediction model using machine learning approach. Proceedings of the 3rd International Conference on Computing Methodologies and Communication (ICCMC). Erode: IEEE, 2019. 1211–1215.

27 贾志强, 李涛, 乐金祥. 基于集成学习算法的消费行为预测. 计算机技术与发展, 2022, 32(5): 141–146. [doi: 10.3969/j.issn.1673-629X.2022.05.024]

28 谢君臣, 李涛, 黄甫, 等. 面向药店会员用户画像的构建及应用研究. 计算机技术与发展, 2022, 32(3): 145–150.

(校对责编: 孙君艳)

© 中国科学院软件研究所 http://www.c-s-a.org.cn