基于特征选择与模型融合的睡眠会员唤醒算法

乐金祥,李　涛,贾志强,肖鉴涛
(武汉科技大学 计算机科学与技术学院,湖北 武汉 430065)

摘　要:针对药品销售行业传统低效营销方式的缺点,将药店睡眠会员是否容易被唤醒的问题抽象为二分类问题,提出了一种面向药店平台的预测睡眠会员唤醒算法,来解决现有睡眠会员唤醒模型应用于药店睡眠会员用户唤醒的局限性且预测用户到店消费精度不高的问题。从会员的行为、属性、动态三个维度提出多种传统营销特征属性,在多视角的基础上,设计出药品营销的独有特征属性构建出特征集合,将特征集合代入到支持向量机 SVM 以及 XGBoost 算法模型并使用 Soft Voting 方法进行模型融合。通过实验表明,相对于使用传统特征的单一模型,使用集成学习提取的特征集合所训练的融合模型的 precision 高出 4% 左右,recall 高出 5% 左右,AUC 值提升了 15% 左右,由此可知,基于特征选择与模型融合的睡眠会员唤醒算法具有更好的唤醒效果。

关键词:睡眠会员;行为特征;多视角;特征发现;集成学习
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2022)02-0125-05
doi:10.3969/j.issn.1673-629X.2022.02.020

Sleeping Members Wake-up Algorithm Based on Feature Selection and Model Fusion

YUE Jin-xiang, LI Tao, JIA Zhi-qiang, XIAO Jian-tao
(School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract: Aimed at the shortcoming of traditional inefficient marketing pharmaceutical sales industry, pharmacy sleep members would be easier to be awakened the abstraction for binary classification problems, and a predictive sleep member awakening algorithm for drugstore platform is proposed to solve the limitations of the existing sleep member awakening model applied to drugstore users and the low accuracy of predicting users' consumption in stores. A variety of traditional marketing characteristic attributes are proposed from the three dimensions of members' behavior, attributes and dynamics. On the basis of multiple perspectives, the unique characteristic attributes of drug marketing are designed to construct feature sets which are inserted into support vector machine SVM and XGBoost algorithm model, and Soft Voting method is used for model fusion. The experiment shows that compared with the single model using traditional features, the precision of the fusion model trained by the feature set extracted by ensemble learning is about 4% higher, the recall is about 5% higher, and the AUC value is about 15% higher. Therefore, the sleep member awakening algorithm based on feature selection and model fusion has better awakening effect

Key words: sleeping members; behavioral traits; multiple perspectives; feature discovery; ensemble learning

0　引　言

随着中国社会经济的快速增长,人民生活质量和水平不断改善与提高,同时健康和保健的重视程度也在不断加大与提高,而且在互联网与实体经济融合发展的大背景下,医药行业是现阶段增速最快的行业之一,国内药品零售业也在快速发展。作为药品零售的主要经营模式——连锁零售药店也进入快速扩张的阶段,越来越多的医药平台开始采用线上活动引流至线下营销的模式来提升营销能力。随着线上购物的普及,通过短信发送或是在 APP 上推送优惠券是刺激老用户以及吸引新客户进店消费的重要方式,然而直接对所有睡眠会员推送优惠信息来进行唤醒操作,不仅会导致营销成本过高,而且随机发放信息既无法达到预期效果,也会对用户的生活造成干扰。对商家而言,针对所有睡眠会员进行优惠信息大规模发送,不仅在增加药店营销成本的同时,也很大可能会损害药店的品牌声誉[1]。

对于当下的药店而言,药品的种类、价格都相差无

收稿日期:2021-03-22　　**修回日期**:2021-07-22
基金项目:国家自然科学基金资助项目(61702383);湖北省教育厅重大项目(17ZD014)
作者简介:乐金祥(1995-),男,硕士研究生,研究方向为机器学习;李　涛,博士,教授,研究方向为推荐系统人工智能、信息安全。

几。在这种现状下,消费者可能会考虑到地理位置更近的药店消费,各大药店几乎都有会员流失严重的现象。如果想要获得更好的发展,在吸引更多新会员的同时,要维护好现有的老会员,减少会员流失。但是发展新会员的成本远高于维护现有会员的成本,在发展新用户相对困难的现阶段,如何维护好老会员,唤醒睡眠会员是更好的选择。因此怎样更好地识别这些易唤醒的睡眠会员是该文的研究重点,将从两个方面出发:(1)特征集构建:从会员的行为、属性、动态三个维度提出多种具有代表性的特征属性,与传统消费特征结合构建为特征集。该特征集相比传统消费特征具有更好的模型训练效果。(2)唤醒模型构建:将会员是否会前来消费抽象为二分类问题,可用于二分类的算法比较多,该文选择了支持向量机以及 XGBoost 两种算法,使用 Soft Voting 方法将 SVM 唤醒模型和 XGBoost 唤醒模型进行融合得到唤醒模型,比单一的唤醒模型具有更好的唤醒效果。

1 相关研究

目前在电子优惠券营销方面已经有了较多研究成果,通过用户的消费数据来分析用户的行为特征、发现用户的偏好与兴趣以及了解用户的潜在需求,通过发送优惠信息来实现精准营销,增加营业额。药店营销正属于电子优惠券营销的一部分,有着传统营销模型的一般性,该文在特征分析与研究方法上,也借鉴了传统消费预测的方法。

文献[2]较为系统地总结了国内现有发展阶段的电子优惠券营销的各种类型及其发展趋势,深入分析了使用网络营销、数据库营销以及线下营销等具体的营销手段发放的不同种类电子优惠券信息发放方式的原因和优势,并进一步阐述了电子优惠券营销的未来发展趋势。文献[3-4]都是使用 XGBoost 算法模型,通过分析历史销售数据,来对未来的销售趋势做出预测。文献[5]为了探究对消费者使用优惠券有影响的特征因素,从个人属性和行为特征等角度来构建特征模型,其中行为视角包括经济效益、便利性以及娱乐性,对消费者优惠券的使用影响最大的是娱乐性,其次是便利性。而个人属性中影响优惠券使用的因素有创新性以及所得优惠券的力度等,这些因素对用户的优惠券使用也有一定的影响。文献[6]提出了一种基于随机森林算法的改进算法模型。利用随机森林的集成思想与训练数据集的随机分割重组的属性,将原始预测数据集进行随机分割重组,得到高维训练数据集,将其输入模型后得到的结果进行求和作为预测的最终值。文献[7]在 XGBoost 算法的基础上,提出一种对用户优惠券使用行为进行预测的算法模型,通过分析历史购买和历史消费券使用等特征来实现优惠券的精准推送。

文献[8]通过实验对比了不同二分类算法模型在用户对优惠券使用预测的 ROC 曲线,得出随机森林预测模型的准确率更高。文献[9]通过对优惠券的领取方式、历史访问记录以及使用优惠券的购买历史等特征进行分析,探究了这些特征与优惠券的使用存在的联系。文献[10]通过实验研究验证了线上优惠券的发放以及历史浏览行为等特征与用户是否购买商品存在正相关关系。

以上研究方法对实体产品营销有一定的适用性,但由于用户对药店的主动选择性和对药品的特殊依赖性,导致仅使用传统模型进行消费预测时准确率不高,这些方法应用于药品营销时存在不足。为了解决该问题,该文在已有传统营销模式的基础上进行相关数据预处理工作,提出新的特征模型,采用 soft Voting 方法融合随机森林和 XGBoost 来进行睡眠会员的唤醒。实验证明,该特征模型以及将 SVM 唤醒模型和 XGBoost 唤醒模型进行融合得到的唤醒模型,比单一的唤醒模型具有更好的唤醒效果。

2 基于特征选择和模型融合的睡眠会员唤醒模型介绍

2.1 特征选择

该文数据来源于某连锁药店近期营销活动所发放优惠券的睡眠会员的属性、该药店活动中所有睡眠会员在活动期间的消费数据,将这些数据进行组合后所形成的数据集比较庞大,共 380 026 条。

借鉴睡眠会员消费行为预测采用的特征,从睡眠会员自身属性、睡眠会员行为属性和药品属性三个角度进行计算,对于连锁药店和可能影响睡眠会员是否被唤醒前来消费的特征及描述如表 1 所示。

药品销售与普通产品销售在用户属性和产品属性上有很多相似点,但同时也存在一定差异性。根据药品的特殊性、药店的连锁性、会员病类特征与药品的关联性,该文在传统直观特征的基础上进行了新的特征设计。

支付倾向性(PT):在普通实体店消费中,用户并不会关心最后的支付方式,而在药店消费中,可使用医保卡支付的药店更受用户好评,那么他的潜在消费可能性也就越大。

支付倾向性即用户消费后使用医保卡支付的次数(TotalCard)与支付总次数(TotalNumber)的比值,定义如下:

$$PT = \frac{TotalCard}{TotalNumber}$$

表 1　睡眠会员自身属性、行为属性和药品属性

角度	特征	特征描述
属性	会员性别(sex)	会员性别男或者女
	会员年龄(age)	会员的年龄
	会员卡年份(create_date)	会员卡注册至今的年份
	距消费门店的平均距离(distance)	会员过去一年内到药店消费距住址的平均距离(总距离/消费次数)
	会员活跃度(level)	会员在店内的消费活跃度(根据近期消费次数决定)
	会员积分(point)	会员通过消费总共获取的积分
行为	近 6 个月平均消费(ma6)	会员在店内最近六个月的平均消费金额 (总金额/总次数)
	近 6 个月消费次数(mt6)	会员在店内最近六个月的消费次数
	近 6 个月消费种类(num6)	近六个月会员到药店所购买的药品总数量
	近 1 年平均消费(ma1)	会员在店内最近一年的平均消费金额 (总金额/总次数)
	近 1 年消费次数(mt1)	会员在店内最近一年的消费次数
	近 1 年消费种类(num1)	近一年会员到药店所购买的药品总数量
	近 3 个月平均消费(ma3)	会员在店内最近三个月的平均消费金额 (总金额/总次数)
	近 3 个月消费次数 mt3	会员在店内最近三个月的消费次数
	近 3 个月消费种类 num3	近三个月会员到药店所购买的药品总数量
	使用优惠券消费次数(total)	会员在一年内使用折扣参加活动的总次数
药品	是否购买慢病药品(is_suff)	是否购买慢性病药物

药品依赖度(DD):由于药店的特殊性,经常到药店进行购药的会员可能身患某种慢病、重症或者老年症,他们需要定期购药,对于这些用户,他们的潜在消费可能性通常会更高。药品依赖度与会员自身所患疾病和该类疾病的购药频率相关,定义如下:

$$DD = \sum \frac{DiseaseDrug_i}{TotalDrug}$$

药品关联性(DA):同类药品之间通常存在相同的作用标签,会员在购买该类药品的同时可能会考虑购买同类同效药品,在药店营销中,提供更多的同类药品能够提升对会员的吸引力。药品关联性即用户在消费中购买某类药品的同类同效药数量与该药店拥有该类型药品的总数量之比,定义如下:

$$DA = \frac{CTDrug}{Total}$$

购药频率(BR):通常而言,购买药品频率高的会员潜在消费可能性更大。药品购买频率即单位时间内会员购药的次数,由于采用的样本集是近一年数据,因此计算月平均购药频率,定义如下:

$$BR = \frac{TotalNumber}{12}$$

品牌依赖度(BD):根据连锁药店的性质,会员在该连锁药店下所消费的次数越高,表明会员对品牌的亲密度越高,那么对该品牌药店的潜在消费可能性也就越大。品牌依赖度即在某一连锁药店的消费次数与会员的总消费次数之比,定义如下:

$$BD = \frac{\sum DrugStore_i}{TotalNumber}$$

2.2　睡眠会员唤醒模型

针对连锁药店近一年的会员消费数据、活动推送数据以及会员基本信息,经过数据预处理、特征选择与构建得到最终数据集,按照 7∶3 的比例将最终数据集划分为训练集与测试集。训练集用来进行 XGBoost 与 SVM 算法模型的训练,并按照网格搜索算法进行参数调节对最佳模型调优,然后使用 Soft Voting 方法将两个模型融合得到最终的睡眠会员唤醒模型。再使用测试集验证模型,根据评估指标对模型进行唤醒效果评估。流程如图 1 所示。

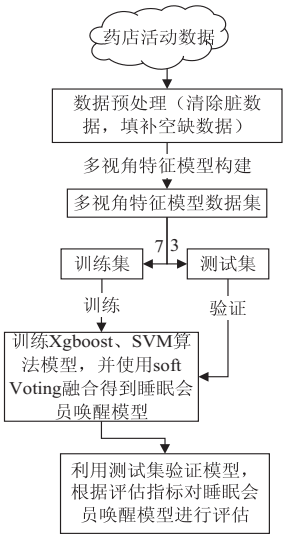

图 1　睡眠会员唤醒模型构建流程

机器学习就是在若干假设中找到一个最匹配当前场景的假设,但现实场景的复杂性决定了没有任何算法模型可以适用于所有学习场景,并且保有较高的学

习准确率[11]。算法融合是指将若干个单一且独立算法的学习结果通过集成方法组合成为一个新的算法模型,从而提高算法最终学习的准确率[12]。而融合算法模型准确率的提升取决于基算法的个数、基算法的选择以及融合方法的选择。

现有的研究证明,对于不同且独立的分类算法,当这些算法在同一应用场景上分类结果的准确率比随机猜测要高,即准确率大于 0.5 时,可通过多数投票的方式提升分类结果的准确率。设 d_i 为每个样本分类结果的后验概率,并且 d_i 是独立同分布的,$E(d_i)$ 为期望,$\mathrm{Var}(d_i)$ 为方差,如果将每个基算法的权重设为:

$$w_i = 1/T(i = 1,2,\cdots,T) \tag{1}$$

当使用简单平均法进行算法融合时,融合后算法的期望和平均值分别为:

$$E(\bar{d}_i) = E(\sum_{i=1}^{T} \frac{1}{T} d_i) = \frac{1}{T} TE(d_i) = E(d_i) \tag{2}$$

$$\mathrm{Var}(\bar{d}_i) = \mathrm{Var}(\sum_{i=1}^{T} \frac{1}{T} d_i) = \frac{1}{T^2} T\mathrm{Var}(d_i) = \frac{1}{T}\mathrm{Var}(d_i) \tag{3}$$

从公式 3 得知,融合后算法的 $E(d_i)$ 对比原来基算法并没有改变,但是 $\mathrm{Var}(d_i)$ 会随着基算法个数 T 的增加而降低,因此融合算法的准确率将高于单个算法。但是,在实际的模型训练中,模型准确率并不是与 T 的值成正相关,应当考虑实际模型的复杂度、训练时间等因素来为 T 取值。对于该文的实际应用场景以及连锁药店活动的用户购买数据,通过实验确定 T 的值取 2 或 3 时,融合之后的模型效果较好。

基算法本身具有较高的准确率,同时它们之间具有较大的差异是融合算法比构成其的单一算法具有更优效果的充要条件。XGBoost 本质仍是 GBDT,但它通过迭代的方式将弱学习器模型重组为强学习器模型,相较于 GBDT 具有更高的效率和泛化能力[13]。SVM 则是一个在特征空间中寻找间隔最大化的分离超平面的分类算法[14]。两个算法模型相互独立且准确率较高,因此将这两个算法进行融合所获得的分类模型可以有效降低分类结果的方差,从而提高分类结果的准确率。

加权平均法在 1993 年被 Perrone 和 Cooper 正式用在算法融合方面。算法融合的本质是通过对样本数据集进行学习,从而确定每个基算法的权重,因此可以说各类算法融合的方法都是加权平均法的特例或变体[15]。简单平均法就是公式 1 中的 T 值直接取基算法的个数的特例。但由于现实场景中的样本数据或多或少存在着噪声,这使得从样本数据中学习到的权重并不是完全可靠,不准确的权重甚至会让模型在训练过程中过拟合。因此,该文选择简单平均法来进行算法融合。

Soft Voting 方法就是一个常用的简单平均方法,它是将基算法模型对各个类别的预测结果的概率进行平均,再对类别预测的平均值进行大小比较来确定样本的最终类别[16]。假设使用 XGBoost 对一个样本的预测结果是属于易唤醒类别的概率为 0.4,属于难唤醒类别的概率为 0.6。使用 SVM 对该样本的预测结果是属于易唤醒类别的概率为 0.7,属于难唤醒类别的概率为 0.3。则该样本属于易唤醒类别的概率为(0.4+0.7)/2=0.55,属于难唤醒类别的概率为(0.6+0.3)/2=0.45。因为 0.55>0.45,所以该样本被划分为易唤醒类别。

3 实验与分析

3.1 评估指标

该文讨论的是将睡眠会员在收到优惠券之后是否到店消费的问题抽象为一个二分类问题。对于二分类问题,该文选择用混淆矩阵进行评价,将预测结果分类为真正类(TP):预测被唤醒且实际上到店消费的睡眠会员;真负类(TN):预测不被唤醒且实际上未到店消费的睡眠会员;假正类(FP):预测被唤醒但实际上未到店消费的睡眠会员;假负类(FN):预测不被唤醒但实际上到店消费的睡眠会员。由于睡眠会员的特殊性,即统计过后到店消费的睡眠会员数量远远低于未到店消费的睡眠会员数量,导致了正负样本数据不平衡,负样本数量远大于正样本数量。结合实际情况,该文采用精确率(precision)、召回率(recall)和 AUC 三个值进行模型评估。

精确率即正确预测为正的占全部预测为正的比例,计算公式为:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

召回率即正确预测为正的占全部实际为正的比例,计算公式为:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

AUC 的定义是 ROC 曲线下的面积,而 ROC 曲线一般是位于直线 $y = x$ 的上方,所以 AUC 值的取值范围一般在 0.5 到 1 之间,一个算法模型的效果好坏与其对应的 AUC 值成正相关。

3.2 对比实验

为了验证该文设计的特征以及融合模型在睡眠会员唤醒的有效性,进行如下对比实验:

(1)采用表 1 中的传统特征(TD)作为分类算法特征向量集合,分别使用 SVM、XGBoost 以及论文方

法进行实验;

(2)在 TD 基础上加入该文设计的新特征得到新的特征向量集合(ND),再分别使用 SVM、XGBoost 以及文中方法进行实验。

对比实验结果如表 2 所示,ROC 曲线如图 2 所示。

表 2　对比实验结果

特征集	模型	precision/%	recall/%	AUC
TD	SVM	73.91	72.62	0.61
TD	XGBoost	74.36	74.27	0.67
TD	论文方法	76.84	77.32	0.72
ND	SVM	75.24	74.18	0.68
ND	XGBoost	75.93	75.15	0.71
ND	论文方法	78.42	78.95	0.74

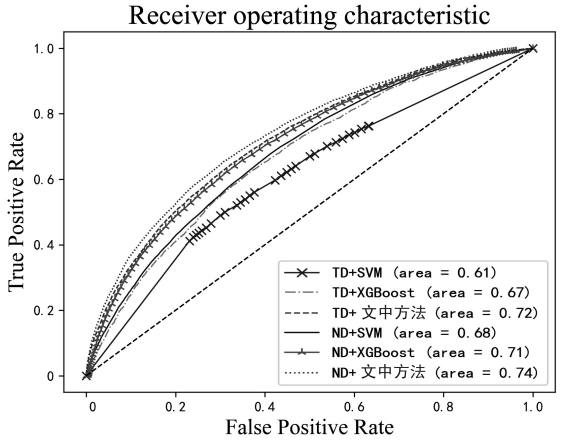

图 2　对比实验 ROC 曲线

由实验结果可见,对于使用相同特征向量的算法模型,该融合模型比单一模型的 precision 平均高出 3% 左右,recall 平均高出 4% 左右,而 AUC 值提升了约 12.5%;对于相同的算法模型,使用了加入新设计特征的特征集合比使用传统特征集合的 precision 平均高出 2% 左右,recall 平均高出 2% 左右,而 AUC 值提升了约 10%;同时使用新的特征集合和融合模型比使用传统特征和单一模型的 precision 平均高出 4% 左右、recall 平均高出 5% 左右、AUC 值提升了约 15%。

综上,该文构建的特征集合以及融合模型能更有效地对睡眠会员唤醒进行预测。

4　结束语

该文研究的主要是药店睡眠会员唤醒问题,研究目的是基于睡眠会员具有丰富的特征,建立训练集与测试集,挖掘出具有什么样特征的会员在收到药店发放的优惠活动信息之后容易被唤醒前来消费,什么样的二分类模型能够更有效地对睡眠会员唤醒进行预测。预测后将优惠券信息发放给最有可能被唤醒的睡眠会员,提高药店活跃会员数量的同时尽可能减少营销成本。以某连锁药店近期营销活动给睡眠会员所发放的优惠券为数据来源,通过对比实验证明了提出的特征集合与集成学习分类模型对药店睡眠会员具有更好的唤醒效果。

参考文献:

[1] LUO Jianing. An offline transferable and divisible mobile coupon based on NFC[C]//Proceedings of international conference on computing, mechanical and electronics engineering. Singapore: International Institute Engineers, 2015.

[2] 陆　平,陈笑天. 基于梯度提升树模型的网络优惠券使用预测[J]. 科学技术与工程,2019,19(18):234-238.

[3] 石纯一. 基于数据挖掘的实体零售业销售额预测研究[D]. 广州:广东工业大学,2019.

[4] 朱振峰,汤静远,常冬霞,等. 基于 GBDT 的商品分配层次化预测模型[J]. 北京交通大学学报,2018,42(2):9-13.

[5] 刘　芬,赵学锋,张金隆,等. 移动优惠券的消费者使用意愿研究:基于个人特征和动机的视角[J]. 管理评论,2016,28(2):93-102.

[6] 葛绍林,叶　剑,何明祥. 基于深度森林的用户购买行为预测模型[J]. 计算机科学,2019,46(9):190-194.

[7] 徐　宁,喇　磊. 基于 XGBoost 的新零售优惠券使用行为预测[J]. 西南师范大学学报:自然科学版,2019,44(3):101-105.

[8] WU J, ZHANG Y L, WANG J F. Research on usage prediction methods for O2O coupons[C]//Neural information processing. Siem Reap, Cambodia: Springer, 2018: 175-183.

[9] 管轶楠. 面向电商企业的电子优惠券投放决策研究[D]. 南京:南京理工大学,2018.

[10] 张建同,方陈承. 顾客历史行为和优惠券对其购买决定的影响——基于一项实验研究[J]. 软科学,2017,31(2):109-112.

[11] 祝　歆,刘潇蔓,陈树广,等. 基于机器学习融合算法的网络购买行为预测研究[J]. 统计与信息论坛,2017,32(12):94-100.

[12] 张建彬,霍佳震. 基于 Stacking 模型融合的用户购买行为预测研究[J]. 上海管理科学,2021,43(1):12-19.

[13] 张微薇,刘　盾,贾修一. 基于 XGBoost 的三分类优惠券预测方法[J]. 南京航空航天大学学报,2019,51(5):643-651.

[14] 蔡　男. 基于改进随机森林算法的电信客户流失预测及分析[D]. 南昌:南昌大学,2020.

[15] 张雷东,王　嵩,李冬梅,等. 多种算法融合的产品销售预测模型应用[J]. 计算机系统应用,2020,29(9):244-248.

[16] AGNIHOTRI D, VERMA K, TRIPATHI P, et al. Soft voting technique to improve the performance of global filter based feature selection in text corpus[J]. Applied Intelligence, 2019, 49(4): 1597-1619.