数据挖掘期末项目

宋新喆 2016110252 黄舒文 2016110849

**一 摘要**

最后写

**二 引言**

为了吸引大量的新买家，商家有时会在特定的日子(如“618”、“黑色星期五”或“双十一”等)举办大型促销活动，通过提供折扣或优惠券的方式，提高下单量。可是这样做经常吸引到一些只会购买一次的用户，在这种情况下，这些促销活动对销售的持久影响可能微乎其微。所以对于商家，确定谁可以转化为重复购买者是很重要的。但在网络广告领域，客户定位尤其对于新买家的定位是有难度的。如果能确定哪些用户会成为长期购买者而为他们定向投放优惠券，那么就可以解决这一问题。

基于该问题的实用价值和个人兴趣，如果我们可以根据用户的行为和特征，去预测出它成为忠实顾客或再次购买的可能，发掘出潜在客户，那商家针对这些潜在的忠诚客户，可以大大降低推广成本，提高投资回报率(ROI)，在适当的情况下，也可以实现二级价格歧视，提高社会整体福利。

我们创新性的运用了……参考了……最终……A brief intro of 方法和发现

简单描述新颖性

**三 文献回顾**

对于用户行为预测的研究自2003年开始并持续被关注至今，前期主要集中在Web浏览行为（高卫华，2007；何丽，2008；党小超，2011；张玉成，2011；Awad Mamoun A，2012；）且多以Markov链或模糊加权优化后的马尔可夫链建立预测模型分析用户行为以及基于差别矩阵的粗糙集提取行为日志的关联规则并用于预测（李学俊，2008；）。

自2012年前后出现关于社交网络中用户行为的研究如信息转发、分享、关注、评论等并开始使用Machine Learning相关算法，主要以微信或微博转发概率或分析传播路径的信息过滤能力为导向的预测：基于贝叶斯模型的预测（谢婧，2013；）、基于逻辑回归模型和特征提取的预测（黄炜，2016；）、基于关联规则的分类算法分析（刘臣，2013；）、基于逻辑回归、SVM与BP神经网络的集成学习模型与优化（张效尉，2018；）。

自2016年随着电子商务平台的发展和数据挖掘和机器学习算法的完善，研究或预测非契约用户购买行为的文献增多，多为预测用户未来购买行为以提升平台体验或制定营销策略（杨光胜，2019；谭晶，2019；邱耀，2018；）、评估用户价值和流失损失（李美其，2016；），算法选择从简单的基分类器过渡到典型的Pareto/NBD模型、BG/NBD 模型以及集成学习如XGBoost算法和深度学习算法。这一阶段对于用户行为预测的算法探索已经基本完善，从基分类模型到集成学习在这一领域都取得了比较好的效果，在本文中可以考虑继续使用这些效率较高的预测模型并加以比较。

而针对与电商平台用户重复购买行为相同主题的研究在近年来也取得了一定的成果，其中有两个高度贴合主题和研究目的的文献。其一张李义等人（2018）以淘宝某店铺的后台数据结合订单与消费者信息为数据集，因数据量过小，采用了SMOTE算法扩充少数类，解决了重复购买消费者与单次购买消费者数量上存在较大差异的问题，并用随机森林算法达到了96%的预测准确率。其二王克利等人（2018）选用了和本文相同来源的数据集即阿里云“天池”大赛的海量天猫用户访问数据，采用对不同训练样本训练不同参数的SVM算法，以AUC作为预测精度的指标，我们借鉴了设置变量的方法。

在已有研究成果基础上，为提高预测的准确性，我们采取不同抽样方法，在数据清洗和类别不平衡的处理上提出创新，在特征分析部分加入特征之间内部关系的挖掘和特征抽取和降维的处理，并丰富集成算法的基算法种类以得到更全面的预测效果。

**四 问题定义（改了一下，但可能还是有点少）**

**先定义符号和概念，在此基础上给出研究问题**

**（写模型的时候把自己定义的值在这里写一下）**

在本文中，我们利用阿里巴巴平台提供的天猫后台数据包括用户个人信息和该用户购买某商品的行为日志构建特征变量，训练不同分类器模型使之能在保证一定准确率及其他评估指标的情况下，预测某用户会再次购买该商品的概率。我们希望能探索一个用户在使用优惠券消费后，是否或有多大概率可以成为长期客户，以给商户提供营销上的决策支持。

特征分析后用于建模的数据表：

|  |  |  |
| --- | --- | --- |
| 特征 | 特征类型 | 释义 |
|  |  |  |
| usertime | 用户特征 | 用户购买的天数 |
| agerange | 用户特征 | 用户的年龄范围 |
| gender | 用户特征 | 用户的性别 |
| merlike  sellitem  sellcat  sellbrand  selluser  sellreuser  label | 用户商户特征  商户特征  商户特征  商户特征  商户特征  商户特征  标签列 | 某用户对商户操作的平均值  商户商品总数  商户种类总数  商户品牌总数  商户买家总数量  商户重复买家总数量  用户是否重复购买该商家 |

**五 数据探索分析**

5.1 数据集和字段类型说明

用户行为日志（user\_log\_format1.csv）：

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 字段说明 |
| user\_id | Norminal | 用户的唯一标识（外键） |
| item\_id | Norminal | 商品的唯一标识 |
| cat\_id | Norminal | 商品所属种类的唯一标识 |
| seller\_id | Norminal | 商户的唯一标识 |
| brand\_id | Norminal | 商品品牌的唯一标识 |
| time\_stamp | Norminal | 用户行为的日期(格式:mmdd) |
| action\_type | Norminal | 用户行为{0,1,2,3}  0表示单击、1表示加入购物车,、2表示购买、3表示收藏 |

用户信息文件（user\_info\_format1.csv）：

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 字段说明 |
| user\_id | Norminal | 用户的唯一标识（外键） |
| age\_range | Ordinal | 用户的年龄范围  1表示(<18)、2 表示[18-24]、3表示[25-29]、4 表示[30-3]、5表示[35-39]、6表示[40-49]、7和8表示(>=50)、0和NULL表示未知 |
| gender | Norminal | 用户的性别  0表示女性、1表示男性、2和NULL表示未知 |

训练数据文件（train\_format1.csv）：

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 字段说明 |
| user\_id | Norminal | 用户的唯一标识（外键） |
| merchant\_id | Norminal | 商户的唯一标识 |
| label | Norminal | 是否重复购买的标签  0表示没有重复购买，1表示有重复购买 |

数据集分为三个，分别介绍了用户的信息和购买行为，以及提供预测的标签，多为唯一标识的Norminal数据即id，缺少统计和描述数据。

5.2 数据清洗与抽样

5.2.1 缺失值和异常值处理

通过分析，缺失值主要在用户信息表中存在于年龄和性别，在用户行为日志表中存在于商品品牌id。在上千万条的数据量下，出于对数据精度的考量，我们选择保留原始精确数据而非用均值填充空值。删除了商品品牌id、年龄和性别三个属性中为空的值，且根据数据源对年龄和性别的定义，同时删除了性别为2的数据行和年龄为0的值，得到用户信息表完整数据317840\*3条数据和用户行为日志表完整数据54834315\*7条数据。观测数据得到用户信息中女性用户为男性的两倍（图1），25-29岁购买者最多，50以后和18岁以下的购买者基本没有（图2）。

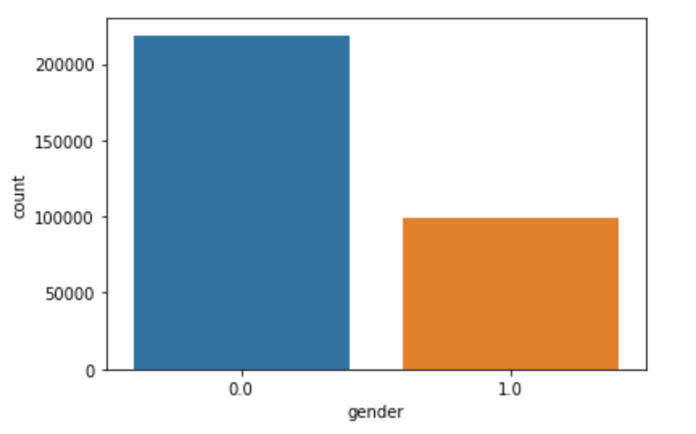


图1 样本男女比

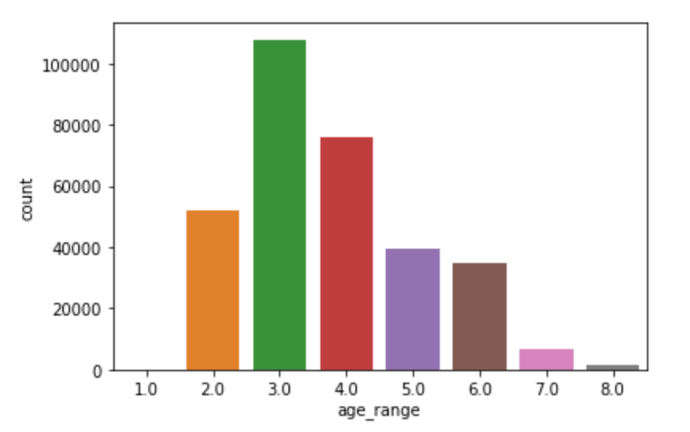


图2 样本年龄范围

5.2.2 数据抽样

在本文中，因用户行为日志表的数据量过大，出于降低运算时间的角度，我们选择了抽样分析的方法。选用的数据集为不平衡数据集，出于运算量的考量，我们选择选择简单随机抽样的方式先在用户信息表中抽取500个用户，然后将新生成的用户信息抽样表和用户行为日志表以及训练数据表进行左交，从而提取出所有含左表的用户id的用户行为日志数据和商品id与重复购买标签，生成的user\_merge\_sample表含有91594条数据，包括了所有用户的个人信息和购买行为信息。我们可以得到样本数据中的男女比仍然为1:2，年龄状况也为真实数据等比例缩小，实现了样本的分层，使各个类别的样本数量比例和原来一致。

5.3 特征分析

阿里巴巴竞赛网站上所给的数据除用户、商户、品牌、种类、商品的唯一标识外，只有用户的年龄、性别和操作行为与时间，在进行预测时是远不能描述用户和商品以及商户的特征的。在此基础上，本文自行统计了以用户特征和商户特征为2类的达12个新增特征。用户特征表示了用户的个人信息以及对商户的操作行为，得到用户的购物习惯，商户特征表示了商户的受欢迎程度，在一定程度上也能表现吸引忠实客户的能力。其中用户特征包括：点击次数、加入购物车次数、购买次数、收藏次数、购买天数，商户特征包括：被点击次数、被加入购物车次数、被购买次数、被收藏次数、商户商品总数、商户种类总数、商户品牌总数、商户买家总数量。经过筛选，删除其中的不足原始数据集0.1%的添加购物车行为即action\_type为1的特征，得到10个新增特征，删除作为唯一标识的id特征列，保留用户的年龄范围和性别，最终共得到12个特征。经过合并数据，共得到613条数据。

5.4 数据标准化

选取的特征在尺度上存在很大的差别，如果属性值的量级比其他属性值的方差还大，这个属性就会改变算法学习的方向且会忽略其他属性的影响。统计分析所选特征，数量级差别较大，为了使分类器的训练更加稳定有效，本文对除年龄和性别的数据进行标准化处理也即归一处理。

5.5 不平衡问题解决办法

在数据抽样时按照随机抽样的方式得到非重复购买的标签0占90%左右，而重复购买的标签1只占10%左右。可见我们的数据集样本类别极不均衡，本文在数据层面和算法层面分别进行处理。在数据层面，为避免再丢失重要信息，我们采取简单随机过采样的方式，增大少数类样本的数量。在算法层面，我们采取用SVM这种仅对少数类进行训练的单类分类器和集成学习的方法对数据进行拟合预测。

**六 模型设计**

**理论方法，模型如何构建**

模型算法介绍

6.1 单分类器：SVM算法

6.2 集成学习：随机森林

**七 数据实验**

**数据集介绍（如何拆分）、实验设置、预测性能、调参、解释分析八 结论与未来工作**

**九 参考文献**

[1]高卫华,谢康林.Web用户行为预测的一种新模型及算法[J].计算机应用与软件,2007(03):142-144+162.

[2]何丽.基于类Markov链的用户浏览行为预测方法[J].计算机工程,2008,34(22):32-33+36.

[3]党小超,郝占军,王筱娟.模糊加权Markov链的用户行为预测[J].兰州大学学报(自然科学),2011,47(01):110-115.

[4]张玉成,徐大纹,王筱娟.基于加权马尔可夫链的主动用户行为预测模型[J].计算机工程与设计,2011,32(10):3334-3337+3418.

[5]Awad Mamoun A,Khalil Issa. Prediction of User's Web-Browsing Behavior: Application of Markov Model.[J]. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society,2012.

[6]谢婧,刘功申,苏波,孟魁.社交网络中的用户转发行为预测[J].上海交通大学学报,2013,47(04):584-588.

[7]刘臣,田占伟,于晶,单伟.在线社会网络用户的信息分享行为预测研究[J].计算机应用研究,2013,30(04):1017-1020.

[8]黄炜,余辉,李岳峰.基于微信用户行为的分享预测模型研究[J].情报理论与实践,2016,39(11):89-94.

[9]李学俊,李龙澍,徐怡.基于粗糙集的Web用户行为预测研究[J].计算机工程与应用,2008(13):134-136.

[10]张效尉,王伟,秦东霞.基于集成学习的微博用户转发行为预测[J].河南师范大学学报(自然科学版),2018,46(02):111-116.

[11]李美其,齐佳音.基于购买行为及评论行为的用户购买预测研究[J].北京邮电大学学报(社会科学版),2016,18(04):18-25.

[12]邱耀,杨国为.基于XGBoost算法的用户行为预测与风险分析[J].工业控制计算机,2018,31(09):44-45.

[13]杨国胜,郭贝贝.机器学习增强的电子商务平台用户行为预测[J].科技与创新,2019(01):124-125.

[14]谭晶.O2O电商用户行为预测应用研究[J].信息通信,2019(01):276-277.

[15]张李义,李一然,文璇.新消费者重复购买意向预测研究[J].数据分析与知识发现,2018,2(11):10-18.

[16]王克利,邓飞其.基于阿里巴巴大数据重复购买预测的实证研究[J].时代金融,2018(03):237-239.

**十 代码：**

数据预处理部分：

（python）

import pandas as pd

import numpy as np

import seaborn as sns

data = pd.read\_table("user\_log\_format1.csv",sep=",")

print(data.head())

print('数据量行\*列',data.shape)

#有54925330行，7列

print('数据集列中是否存在缺失值：\n',data.isnull().any())

print('每一行的缺失值个数：',data.isnull().sum(axis=1))

print('每一列的缺失值个数：',data.isnull().sum(axis=0))

#F为不存在，T为存在

#只有brand\_id有缺失值，共91015条

data=data.dropna(axis=0)

data.shape

#删去有缺失值的行数，现有数据量为54834315

sns.countplot(x="action\_type",data=data)

#sns.countplot(x="time\_stamp",data=data)

user\_data = pd.read\_table("user\_info\_format1.csv",sep=",")

print(user\_data.head())

print(user\_data.shape)

#删除空值和gender列中为2的行

user\_data=data.dropna(axis=0)

user\_data=user\_data[~user\_data['gender'].isin([2])]

user\_data=user\_data[~user\_data['age\_range'].isin([0])]

user\_data.shape

sns.countplot(x="age\_range",data=user\_data)

#sns.countplot(x="gender",data=user\_data)

user\_data.to\_csv("userinfoprep1.csv")

data.to\_csv("userlogprep1.csv")

#随机抽样1000条数据

user\_info\_sample=user\_info.sample(n=1000,random\_state=123,axis=0)

user\_info\_sample.to\_csv("user\_info\_sample.csv")

（R）

install.packages("corrplot")

install.packages('DMwR')

library(ggplot2)

library(corrplot)

library(DMwR)

getwd()

setwd("Documents/学业作品/数据挖掘R/data/data\_format1")

getwd()

#将抽样后的用户信息表和行为日志表合并

#继续合并训练数据中的标签列

user\_sample <- merge(user\_log\_clean,user\_info\_sample,all.y =T, sort =T)

user\_sample1 <- merge(user\_sample,train)

write.csv(user\_sample)

user\_merge\_sample<-read.csv("user\_sample.csv")

head(user\_merge\_sample)

#加入用户特征

#点击次数userclick0、加入购物车次数userclick1、购买次数userclick2、收藏次数userclick3、购买天数usertime

##将用户id存入userid数组里并新建数据框user

#分别计算用户action\_type为0、1、2、3的数据量

userid<-unique(user\_merge\_sample$user\_id)

user<-data.frame(userid)

agerange<-vector("integer",length = nrow(user))

gender<-vector("integer",length = nrow(user))

click<-vector("integer",length = nrow(user))

clickcart<-vector("integer",length = nrow(user))

clickbuy<-vector("integer",length = nrow(user))

clicklike<-vector("integer",length = nrow(user))

usertime<-vector("integer",length = nrow(user))

for(i in 1:nrow(user)){

u<-userid[i]

agerange[i]<-unique(user\_merge\_sample[user\_merge\_sample$user\_id==u,]$age\_range)

gender[i]<-unique(user\_merge\_sample[user\_merge\_sample$user\_id==u,]$gender)

click[i]<-nrow(user\_merge\_sample[user\_merge\_sample$user\_id==u & user\_merge\_sample$action\_type==0,])

clickcart[i]<-nrow(user\_merge\_sample[user\_merge\_sample$user\_id==u & user\_merge\_sample$action\_type==1,])

clickbuy[i]<-nrow(user\_merge\_sample[user\_merge\_sample$user\_id==u & user\_merge\_sample$action\_type==2,])

clicklike[i]<-nrow(user\_merge\_sample[user\_merge\_sample$user\_id==u & user\_merge\_sample$action\_type==3,])

usertime[i]<-length(unique(user\_merge\_sample[user\_merge\_sample$user\_id==u,]$time\_stamp))

}

for(i in 1:nrow(user)){

user$agerange[i]<-agerange[i]

user$gender[i]<-gender[i]

user$click[i]<-click[i]

user$clickcart[i]<-clickcart[i]

user$clickbuy[i]<-clickbuy[i]

user$clicklike[i]<-clicklike[i]

user$usertime[i]<-usertime[i]

}

head(user)

nrow(user)

write.csv(user,file = "user.csv")

#加入商户特征

#被点击次数、被加入购物车次数、被购买次数、被收藏次数

#将商户id存到sellerid的数组里并新建数据框seller

sellerid<-unique(user\_merge\_sample$seller\_id)

seller<-data.frame(sellerid)

head(seller)

nrow(seller)

#分别计算action\_type为0、1、2、3的数据量

merclick<-vector("integer",length = nrow(seller))

merclickcart<-vector("integer",length = nrow(seller))

merclickbuy<-vector("integer",length = nrow(seller))

merclicklike<-vector("integer",length = nrow(seller))

for(i in 1:nrow(seller)){

s<-sellerid[i]

merclick[i]<-nrow(user\_merge\_sample[user\_merge\_sample$seller\_id==s & user\_merge\_sample$action\_type==0,])

merclickcart[i]<-nrow(user\_merge\_sample[user\_merge\_sample$seller\_id==s & user\_merge\_sample$action\_type==1,])

merclickbuy[i]<-nrow(user\_merge\_sample[user\_merge\_sample$seller\_id==s & user\_merge\_sample$action\_type==2,])

merclicklike[i]<-nrow(user\_merge\_sample[user\_merge\_sample$seller\_id==s & user\_merge\_sample$action\_type==3,])

}

#添加新列merclick、merclickcart、merclickbuy、merclicklike

for(i in 1:nrow(seller)){

seller$merclick[i]<-merclick[i]

seller$merclickcart[i]<-merclickcart[i]

seller$merclickbuy[i]<-merclickbuy[i]

seller$merclicklike[i]<-merclicklike[i]

}

head(seller)

#商户商品总数、商户种类总数、商户品牌总数、商户买家总数量

sellitem<-vector("integer",length = nrow(seller))

sellcat<-vector("integer",length = nrow(seller))

sellbrand<-vector("integer",length = nrow(seller))

selluser<-vector("integer",length = nrow(seller))

for(i in 1:nrow(seller)){

m<-sellerid[i]

sellitem[i]<-length(unique(user\_merge\_sample[user\_merge\_sample$seller\_id==m,]$item\_id))

sellcat[i]<-length(unique(user\_merge\_sample[user\_merge\_sample$seller\_id==m,]$cat\_id))

sellbrand[i]<-length(unique(user\_merge\_sample[user\_merge\_sample$seller\_id==m,]$brand\_id))

selluser[i]<-length(unique(user\_merge\_sample[user\_merge\_sample$seller\_id==m,]$user\_id))

}

#添加新列sellmerchant、sellcat、sellbrand、selluser

for(i in 1:nrow(seller)){

seller$sellitem[i]<-sellitem[i]

seller$sellcat[i]<-sellcat[i]

seller$sellbrand[i]<-sellbrand[i]

seller$selluser[i]<-selluser[i]

}

head(seller)

write.csv(seller,file = "seller.csv")

#按照train表里的user\_id和merchnat\_id合并user特征表和seller特征表与标签列

train<-read.csv("train\_format1.csv")

sample1<-merge(user,train,by.x = 'userid', by.y = 'user\_id')

sample2<-merge(seller,sample1,by.x = 'sellerid', by.y = 'merchant\_id')

head(sample2)

#添加商户重复买家数量

sellreuser<-vector("integer",length = nrow(sample2))

for(i in 1:nrow(sample2)){

p<-sample2$sellerid[i]

sellreuser[i]<-length(unique(sample2[sample2$sellerid==p & sample2$label=='1',]$userid))

}

for(i in 1:nrow(sample2)){

sample2$sellreuser[i]<-sellreuser[i]

}

write.csv(sample2,file = "final1.csv")

#样本年龄、性别可视化

nrow(sample2)

ggplot(sample2,aes(x=agerange)) + geom\_bar()

ggplot(sample2,aes(x=gender)) + geom\_bar()

nrow(sample2[sample2$label=='0',])#标签为0的行数

nrow(sample2[sample2$label=='1',])#标签为1的行数

#删除商户id和用户id

sample3<-sample2

head(sample3)

sample3<-sample3[,-11]

write.csv(sample3,file = "final3.csv")

#标准化

final3<-read.csv("final3.csv")

head(final3)

head(data\_scale)

data\_scale<-final3[,-15]

data\_scale <- scale(data\_scale[,-(9:10)],center = TRUE, scale = TRUE)

#加载到数据框中

data\_scale <- as.data.frame(data\_scale)

head(data\_scale)

data\_scale$agerange <-final3$agerange

data\_scale$gender<-final3$gender

data\_scale$label<-final3$label

write.csv(data\_scale,file = "final4.csv")