第 18 卷　第 14 期　2018 年 5 月
1671—1815(2018)014-0186-04

科　学　技　术　与　工　程
Science Technology and Engineering

Vol. 18　No. 14　May 2018
ⓒ　2018　Sci. Tech. Engrg.

引用格式:杨立洪, 白肇强. 基于二次组合的特征工程与 XGBoost 模型的用户行为预测[J]. 科学技术与工程, 2018, 18(14): 186—189
Yang Lihong, Bai Zhaoqiang. User behavior prediction based on feature engineering of quadratic combination and XGBoost model[J]. Science Technology and Engineering, 2018, 18(14): 186—189

基于二次组合的特征工程与 XGBoost 模型的用户行为预测

杨立洪　白肇强
(华南理工大学数学学院,广州 510640)

摘　要　特征构造的难题在数据挖掘过程中一直存在,传统固化的特征工程对于业务场景千变万化的数据挖掘任务所带来的效益十分有限,因此解决特征工程的特征构造问题已经成为数据挖掘的瓶颈之一;尤其在机器学习算法快速发展的情况下,特征逐渐成为模型中急需重视的部分。基于电商平台的用户行为数据,在原有特征群的基础上提出了二次组合统计特征的构建方法。利用二次交叉衍生出丰富而又切合业务场景的特征群,同时结合两种滑动窗口的方法,分别是定长滑动窗口获取更多的训练样本,变长滑动窗口获取具有时间权重的训练特征,以此来最大限度地还原出用户真实的行为习惯。最后,使用不同的特征组合结合降维的方法建立对照检验模型;并利用线性的逻辑回归模型、线性支持向量机以及树模型极端随机森林与 XGBoost 对模型进行交叉验证。结果表明,组合特征在树模型的算法中得到了非常好的表达效果;而且无论在线性模型还是树模型中衍生特征群模型的 F1 值都优于基础特征群。

关键词　特征工程　　二次组合特征　　用户行为预测　　XGBoost
中图法分类号　TP391.41;　　**文献标志码**　A

特征工程作为数据挖掘的重要组成部分,在近年来的数据挖掘比赛甚至是算法比赛中都扮演着越来越重要的角色,在算法瓶颈难以突破的时候,好的特征组合可以往往可以事半功倍。

特征工程涵盖多个部分[1],包括特征提取、特征构造、特征缩放、特征编码、以及特征降维,每一个部分都有不同的方法,不同于机器学习,传统的机器学习研究主要偏重于算法的改进与突然,而且模型所用的数据集往往是特定的,并且有明确的目的,因此机器学习对特征的关注度并不是特别的高,而是把重心更多地放在模型与算法方面。而数据挖掘恰恰相反,不仅业务流程千差万别,甚至有时候连明确的目的都没有,因此数据的重要性不言而喻,而特征作为数据能获取到的最重要的信息,则更是重中之重,正是由于这些原因,基于数据挖掘的特征工程应该有一套完整的流程与理论,而且在不同的业务场景下应该有不同的构建方法,以使数据所变现出来的特征最切合当前的业务实际,否则数据挖掘无从谈起。

2017 年 11 月 8 日收到　　广东省产学研协同创新成果转化项目(2016B090918041)和广州市产学研协同创新重大专项(201504302222568)资助
第一作者简介:杨立洪(1961—),博士,教授。研究方向:概率统计与大数据分析。E-mail:malhyang@scut.edu.cn。

本文通过利用用户的个人特征、产品的特征以及用户与产品的交互行为数据,在常见特征工程的构建基础上,提出一种新的从业务应用场景[2—4]出发的方法来构建特征工程,同时利用两种滑动窗口法获取更多的样本和特征,对不同时间段的用户行为构建加和特征,使特征具有了时间的属性,最后利用线性模型与树模型对不同的特征组合进行交叉验证比较其 F1 值,检验组合特征的有效性。

1　特征工程理论

1.1　特征提取

原始数据难免会有缺失值以及冗余的情况,甚至还有一些与模型决策明显没有关系的特征,因此需要进行初步的筛选与提取。对缺失值的处理有多种选择,可以使用插补法,如均值插补、众数插补等,或者基于其他特征利用模型对缺失值进行预测填补。而对于冗余或者与挖掘任务无关的特征则采用直接删除的方法,以减少其对模型带来的干扰,同时降低特征维度。

1.2　特征构造

对于有效特征维度过低的数据集,需要进行二次组合构造新的特征;因此需要对不同的特征进行交叉组合,交叉使得特征之间可以相互联系相互作用,从而表达出单一特征所不具有的非线性性。如

上网时间和上网时长这种相关性较高的特征,组合之后的特征往往可以挖掘出用户的上网习惯,从而推断出用户的年龄段甚至职业等;或者把性别与用户浏览的产品进行组合,从而更好地找出用户的标签,因为男性用户和女性用户的对产品的偏好肯定会有差异。

1.3　特征缩放

对于连续型特征,其单位往往是千差万别的,而且很有可能值的大小也不是同一个数量级,如果不消除这种影响,很容易导致某些特征无法在模型得到应有的表达效果,造成特征失效。因此对原始特征往往需要进行无量纲化缩放,常用的方法有极差化与标准化。极差化利用极大值与极小值把特征缩放到区间[-1,1]或者[0,1],但极差化很容易受到离群点的干扰,使得缩放结果不够鲁棒。而标准化在样本量足够大时,其均值和方差越发接近真实值,所以缩放结果更为鲁棒,并且缩放后的特征关于 0 点中心对称,这对很多模型的训练收敛有很大的帮助。

1.4　特征编码

对于类别型特征,原始特征通常不适合直接作为模型进行输入,比如性别、颜色以及年龄这种数据,应当采用适当的编码方法转变成适合机器学习模型的输入。对于性别这种二值取值的特征,可以采用二值化,如 0 代表女性,1 代表男性。但对于颜色这种类型的特征却不能这样编码,若 0 代表红色,1 代表绿色,2 代表蓝色,那么机器学习模型会认为红色和绿色更接近,而红色和蓝色则差别较大,这和特征的本意相违背,颜色之间并不应该进行直接比较。这种情况可以使用哑编码方法(one-hot lable),即利用三个占位符,用(1,0,0)代表红色,(0,1,0)代表绿色,以此类推,有多少个类别,就用多少个占位符表示。而年龄这种特征,主要用来进行人群划分,所以一般采用离散化划分为区间,如 18 ~ 29 为青少年、30 ~ 39 为中年等。

1.5　特征降维

特征维度过低,会导致无法很好的挖掘出数据集里的有用信息,可是特征维度过高,也会造成维度爆炸,影响模型的计算速度,也使得重要的特征难以在模型得到应有的表达,甚至造成模型无法收敛,影响整个模型的效果。因此,综合模型的精度和速度要求,常对维度过高的特征进行降维处理,常用的降维方法主要分为特征抽取与特征选择两种。

特征抽取,主要使用映射的方法把原来的特征做一个映射变换,常用的线性降维方法有主成分分析与线性判别分析,非线性降维方法主要有基于流行学习的局部线性嵌入。

特征选择,主要是选取原始特征集的一个子集作为目标特征集,常用的方法有三种,分别是过滤式(filter)、包裹式(wrapper)以及嵌入式(embedding)。

2　二次组合统计特征

针对用户在网上购物的行为特性,可以对特征进行特征群划分,其中一级特征群分别为用户特征群 U(见表 1)、商品特征群 I(见表 2)以及商品类别特征群 C。然后,在一级特征群的基础上,重新设计适合于购买行为预测这种业务场景的二次组合统计特征。

表 1　用户基本特征

Table 1　Basic features of users

特征名称	特征含义
user_data	用户数据
user_id	用户 ID
age	年龄段
sex	性别
user_lv_cd	用户等级
user_reg_tm	用户注册日期

表 2　商品基本特征

Table 2　Basic features of items

特征名称	特征含义
sku_comment	商品评价数据
dt	截止到时间
sku_id	商品编号
commen_num	累计评论数分段
has_bad_comment	是否有差评
bad_comment_rate	差评率

针对购买行为预测所构造二次组合统计特征,其目的主要是找出用户的“动机”。在网上平台进行购物与线下购物一样,用户在购物前必然会不断对比同类别的商品,从用户的角度思考,购买其实就是商品在用户心中的一个排序行为,因为通过历史数据,尝试通过组合特征来捕捉这种潜在的行为信息,部分组合特征见表 3。

通过一级特征复合而成组合特征,很好地利用了特征之间的优势互补从而使特征的表达能力变强,

表 3　二次组合统计特征

Table 3　Statistical combination feature

UI 特征群	UC 特征群	IC 特征群
计数特征	计数特征	比例特征
加和特征	加和特征	排序特征
权值特征	权值特征	
比率特征	比率特征	
均值特征	均值特征	
交互时间特征	交互时间特征	
习惯偏差特征	习惯偏差特征	
规则特征	星期分布特征	

如商品特征(I)属于稀疏特征,所含的信息量比较少,而商品类别特征(C)的信息量相对较大,但在进行商品购买预测时其特征相关度较低,因此复合而成的 IC 特征群则互补了两类特征的优劣。

在二次组合统计特征的基础上,还可以再次生成衍生特征群,通过二次交互并基于业务场景的可解释性,可以得到 UI&UC 特征群、U&UI 特征群以及 U&UC 特征群,具体见表 4 和表 5。

表 4 衍生的二次组合统计特征

Table 4 Derivative statistical combination feature

UI&UC 特征群	U&UI 特征群	U&UC 特征群
竞争特征	基本比率特征	基本比率特征
排名特征	二次购买特征	二次购买特征
	交互时间比特征	竞争特征
	交互排名特征	交互时间比特征
		交互排名特征

表 5 部分组合特征含义与作用

Table 5 The meaning and function of partial combinatorial features

特征群	特征名	特征含义	优势及作用
UI&UC	uiuc_col_ln_weight	商品收藏加权值在同类别商品加权值中的排序	防止预测用户购买同类别下的大量商品
U&UI	uiu_col_ln_weight	商品收藏加权值在所有商品加权值中的排序	可以预测出用户最想购买的商品
U&UC	ucu_col_ln_weight	同类别商品收藏加权值在所有商品加权值中的排序	可以预测出用户最想购买的类别
UI	ui_click_apart_01	用户点击数量在第一时间段的累计值	可以预测出用户的购物习惯

除了从特征组合入手,还可以从特征与业务场景的关系进行特征分割。如对于商品点击量这个特征,可以进行分时段划分特征,若取 0:00 ~ 8:00、9:00 ~ 18:00 以及 19:00 ~ 23:00,则可以从中挖掘出用户的上网习惯,从而很好地区分出不同的用户群甚至推断出用户的职业等信息,与此相似的特征还有浏览时长(上网时长)、下单时间等。

3 两种滑动窗口法的使用

常用的滑动窗口法一般为定长滑动窗口法[5,6]。定长滑动窗口法首先固定特征与标记的区间长度,然后通过窗口的滑动不断提取训练样本。例如,在电商的购买预测中,把第 1 ~ 30 d 的用户的各种操作作为特征,以第 31 d 用户是否购买作为标记,便可以提取第一批样本;以第 2 ~ 31 d 为特征区间,第 32 d 为标记区间,即可提取第二批样本,以此类推,同时滑动的步长可以改变,对于具有星期分布的商品,滑动步长可以选为 7。定长滑动窗口法通过产生更多的样本,可以解决电商场景样本比例不均衡的问题,由于购买用户通常比例很小,直接训练导致模型无法拟合,有更多的样本之后可以通过欠采样解决样本平衡问题。

除此之外,提出了一种新的变长窗口滑动法。变长窗口滑动法首先固定标记区间的长度,然后通过改变特征区间的长度来获取不同时间段的加权特征。如用户的行为对购买的影响会随着时间的迁移逐渐减少,同样以第 31 d 作为标记区间,前 3 d(第 28 ~ 30 d)、前 7 d(第 24 ~ 30 d)、前 14 d(第 17 ~ 30 d)的用户行为对购买的影响肯定会随着时间往前推移而对后续购买的影响逐渐减少,因此可以对不同区间的行为赋予权值,以获得能表达时间属性的特征。特征区间的长度选取也是可变的,如同样对于具有星期分布的商品,可以选前 7 d、前 14 d、前 21 d 等。

4 特征组合效果验证

利用二次组合统计特征以及变长窗口滑动法构造的特征,可以构造出不同的特征组合,并以基础特征群作为对比,进行模型训练。构造好的特征采用正态分布标准化进行缩放,并且利用基于 L1 正则化的嵌入式特征选择[7]方法对二次组合统计特征进行降维处理。L1 正则化通过产生稀疏矩阵对特征进行选择,实验中采用的 L1 模型也使用了交叉验证进行模型选择,目的是为了消除数据随机抽样带来的干扰。

对所有的特征组合分别使用四种机器学习模型进行建模,并进行交叉验证,对特征组合进行实验并评价其效果。选取的机器学习模型分别是线性的逻辑回归模型、线性支持向量机、树模型极端随机森林以及树模型 Xgboost。其中线性支持向量机核函数固定为线性核,同时分类器构建模式使用一对多模式(OVR)[8]代替原有的一对一(OVO)模式,使得其在大样本量时也能有很好的训练性能。极端随机森林与普通随机森林的不同之处主要为每一颗决策树在训练时不再应用样本重抽样[9,10],而是使用所有的训练样本,并且分裂节点的选取不再进行子集内的遍历,而是随机选取,因此极端随机森林的随机性更强。XGBoost 模型是梯度提升树的一个改进模型,通过目标函数优化并且利用 C + + 实现底层算法,极大地提升了模型的速度与精度。

数据集采用某网上购物平台 2016 年 2 月 1 日 ~ 2016 年 3 月 31 日为期两个月的脱敏数据,包括用户数据、商品数据、评价数据以及行为数据,以此为基础构建用户购买商品的预测模型,输出高潜用户和目标商品的匹配结果,预测时间为 4 月 1 日 ~4 月 5 日,结果评价标准采用 F1 值,结果见表 6。

表 6　不同特征组合建模的 F1 值

Table 6　The F1 value of different feature combinations

模型	基础特征群	二次组合特征群	组合特征(降维后)
逻辑回归	0.757 2	0.767 3	0.759 3
线性支持向量机	0.760 5	0.766 2	0.760 4
极端随机森林	0.794 2	0.897 3	0.808 2
XGBoost	0.797 3	0.897 3	0.817 3

5　结论

从表 6 可以看出,在特征组合相同的情况下,逻辑回归与线性支持向量机的 F1 值相差不大,而两个树模型的 F1 值远高于逻辑回归与线性支持向量机,在使用的决策树数量一样的情况下,XGBoost 模型的训练时间仅仅约为极端随机森林的 80%。与此同时,二次组合特征群对模型带来的提升十分显著,即使在经过 l1 正则化降维以后,F1 值有所下降,但仍然显著优于使用基础特征群的模型。最后,二次组合统计特征在树模型下得到的提升更大,这与用户行为的高规则化不无关系,因此二次组合统计特征结合树模型可以最大限度地还原出网上用户的行为习惯,对网络用户行为挖掘的业务场景应用具有很好的指导作用。

参 考 文 献

1　周志华. 机器学习. 北京: 清华大学出版社, 2016
Zhou Zhihua. Machine learning. Beijing: Tsinghua University Press, 2016

2　Zhang C J, Zeng A. Behavior patterns of online users and the effect on information filtering. Physica A: Statistical Mechanics and Its Applications, 2012; 391(4): 1822—1830

3　邹　润. 基于模型组合算法的用户个性化推荐研究. 南京: 南京大学, 2014
Zou Run. A combined model of decision trees algorithm for predicting users' shopping behaviors. Nanjing: Nanjing University, 2014

4　冯　晨, 张旭翔. 数据挖掘技术及算法综述. 电脑知识与技术, 2009;(13):3331—3332
Feng Chen, Zhang Xuxiang. Review of data mining techniques and algorithms. Computer Knowledge and Technology, 2009; (13): 3331—3332

5　Yin Y F, Gong G H, Han L. Theory and techniques of data mining in CGF behavior modeling. Science China Information Sciences, 2011;54(4):717—731

6　王考杰, 郑雪峰, 宋一丁. 一种基于滑动窗口的数据流相似性查询算法. 计算机科学, 2010;(10):169—172,201
Wang Kaojie, Zhen Xuefeng, Song Yiding. Algorithm based on sliding window for similarity queries over data stream. Computer Science. 2010;(10):169—172,201

7　Quinlan J R. C4.5: Programs for machine learning. Elsevier, 2014

8　肖云鹏. 在线社会网络用户行为模型与应用算法研究。北京: 北京邮电大学,2013
Xiao Yunpeng. Research on modeling and algorithms of human dynamic in online social networks. Beijing: Beijing University of Posts and Telecommunications, 2013

9　Kuchta D. Use of fuzzy numbers in project risk (criticality) assessment. International Journal of Project Management, 2001; 19(5): 305—310

10　罗冬梅. 网络协议流不平衡环境下基于机器学习算法的在线流量分类方法研究. 科学技术与工程, 2017; 17(28): 103—107
Luo Dongmei. Online traffic classification method based on machine learning algorithm. Science Technology and Engineering, 2017; 17(28): 103—107

User Behavior Prediction Based on Feature Engineering of Quadratic Combination and XGBoost Model

YANG Li-hong, BAI Zhao-qiang

(Department of Mathematics, South China University of Technology, Guangzhou 510640, China)

[**Abstract**] Constructing feature has always been a problem in the process of data mining when conventional ways for feature engineering do not satisfy the need of various data mining mission any more. As machine learning is in a state of rapid development, feature engineering has been playing an important role gradually. The data of user behavior was used to construct statistical combination feature based on the original feature, which is particularly suitable for the business scene. At the same time two different window sliding method is used, in other words, fixed length window sliding to obtain more training samples, and variable length window sliding to get more feature from different time dimension, for the purpose of reproducing the real habit of user in daily life as much as possible. In the end of this paper, different combinations of features will be used for control experiment, while different models such as LR, SVM, ET and XGBoost are all used for experiment as well. The results show that no matter in the linear model or tree model, the F1 value of the combination feature group is better than the original feature group.

[**Key words**] feature engineering　feature combination　user behavior prediction　XGBoost