计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900018

电商平台用户再购物行为的预测研究

吕泽宇　李纪旋　陈如剑　陈东明
东北大学软件学院　沈阳 110167
(yuge0099@gmail.com)

摘　要　电商平台上用户的购物行为研究对于电商企业来说具有重要的商业应用价值。文中针对购物者在同一电商平台上的再次消费行为的预测问题进行了研究。首先,针对用户与商家的行为和交易记录,基于特征工程方法设计了多种不同的行为预测特征,基于可视化等方法对比分析了预测特征的重要性和特点,进行了属性筛选;然后,基于提出的预测特征设计使用了多种不同算法训练预测模型。实验研究表明,多 lightGBM 模型的融合方法能够达到很高的再购物行为预测准确度,其 *AUC* 值能够达到 0.7018,同时,基于这种方法实现的预测器只需要少数特征就能对预测结果产生很好的贡献。研究的数据来源是开源的真实大数据,研究成果具有应用和学术双重价值。

关键词:再次购物行为预测;特征工程;特征可视化;融合模型
中图法分类号　TP181

Research on Prediction of Re-shopping Behavior of E-commerce Customers

LV Ze-yu, LI Ji-xuan, CHEN Ru-jian and CHEN Dong-ming
Software College, Northeastern University, Shenyang 110167, China

Abstract　The study of customers' shopping behavior is a trending research topic and has great commercial value for e-commerce companies. This paper studies the prediction of customer's re-shopping behavior on the same e-commerce platform. Through the analysis of shopping related actions of customers and transaction records between customers and merchants, a variety of different behavior features are designed based on feature engineering principles, and the importance and characteristics of the prediction features are analyzed by using visualization approaches. Then, based on the proposed predictive features, a variety of different algorithms are used to train the prediction models. Experimental research shows that the multi-lightGBM model ensemble method can achieve high prediction accuracy, and the *AUC* value can reach 0.7018. Meanwhile, the predictor only needs a few features to obtain very good prediction results. The experimental data set studied in this paper is an open source big data collected in real environment, and the research conclusions have both application and academic value.

Keywords　Re-shopping behavior prediction, Feature engineering, Feature visualization, Model ensemble

1　概述

B2C 电子商务平台(本文简称电商平台)经常通过节日促销活动来吸引顾客,其中一部分受到促销活动吸引的消费者的购物行为只是一次性行为,也有一部分消费者可能成为潜在的"高忠诚度买家",即仍会再次消费。针对这种情况,电商企业希望通过消费者的行为挖掘来搜索潜在的"高忠诚度买家",从而通过"精准推销"来提升销售业绩并降低促销费用,提高投资回报率[1]。

电商平台上消费者的行为通常包括在虚拟店铺中点击链接、收藏商品和购买物品等,对这些记录信息的分析挖掘不但可以获得消费者的购物偏好,甚至可以反过来勾勒虚拟店铺的特征,而这些信息综合起来则可以用于预测消费者的消费行为。本文将针对消费者消费行为预测中的一个特例,即再次消费行为预测进行研究。再次消费行为预测是指消费者首次在某虚拟店铺购物之后,未来一定时间内再次在该店铺购物的可能性预测。2014 年 Kaggle 网站[2]发布开源消费者行为数据并提出"repeat buyer"预测问题,从而引起学术界的广泛关注。2015 年,阿里巴巴旗下的天池大数据平台也发布了类似的竞赛[3]。本文的研究就是基于天池大数据平台的开源数据进行的。本文把再次消费行为预测问题视为一个分类问题,从特征工程角度对消费者的购物行为做了相关性分析,提取了统计特征、比率特征和隐式特征等多种可用于再次消费行为预测的特征属性[4],对提取到的特征属性进行分类,基于缺失率和特征相关性方法进行属性筛选[5]。本文还通过训练好的模型输出特征的重要程度并且在保证模型预测准确度的同时显著减少了特征数量,提高了计算效率。

2　数据描述

本文使用的数据集是天池大数据平台在 2015 年发布的

基金项目:国家级大学生创新创业训练计划资助项目(201910145222);中央高校基本科研业务专项资金(N182410001)
This work was supported by the National Training Program of Innovation and Entrepreneurship for Undergraduates (201910145222) and Fundamental Research Funds for the Central Universities(N182410001).
通信作者:陈东明(chendm@mail.neu.edu.cn)

数据，该数据包含了 2014 年消费者在“双十一”活动及前 6 个月的消费者与商家的行为和交易记录[3]。数据集由 4 部分组成，分别是消费者行为记录、消费者信息、训练数据和测试数据。其中有标签的训练数据有 260864 条，用于在线评估的无标签测试数据有 261477 条，消费者行为记录有 54925330 条，用户数为 424170。具体数据如表 1－表 3 所列。

表 1　消费者行为记录

Table 1　Customer’s shopping logs

属性信息	定义
user_id	消费者的 id
item_id	商品的 id
cat_id	商品所属的类别 id
mechant_id	店铺的 id
brand_id	品牌 id
time_tamp	时间戳
action_type	操作类型

表 1 中消费者行为记录中的操作类型有 4 种，0 代表点击事件，1 代表加入购物车，2 代表购买，3 代表收藏。

表 2　消费者个人信息

Table 2　Customer’s information

属性信息	定义
user_id	消费者的 id
age_range	年龄范围
gender	性别

表 2 中，年龄属性分为 8 个范围，1 代表小于 18 岁，2 代表 18～24 岁，3 代表 25～29 岁，4 代表 30～34 岁，5 代表 35～39 岁，6 代表 40～49 岁，7 和 8 代表大于和等于 50 岁，0 代表未知。性别属性中，0 代表女性，1 代表男性，2 表示性别未知。

表 3　训练和测试数据格式

Table 3　Train and test data formats

属性信息	定义
user_id	消费者的 id
merchant_id	店铺的 id
label	标签

label 属性中，0 代表没有再次消费的消费者，1 代表是再次消费的消费者。训练数据集和测试数据集的格式一样，但是测试数据集中的 label 是需要预测的。训练数据集中，label 为－1 代表消费者在商家促销之前就有购买记录，虽然他们不在预测的范围内，但是其交互日志可以提供关于商家的信息。

对数据进行统计发现，约有 6％的买家在促销结束后再次在同一商家消费，如图 1 所示。平均每个消费者有 147 个交互日志，而最高日志数达到 2800 个，最少有 3 个，大多数消费者行为记录集中在 200 以下，如图 2 所示。

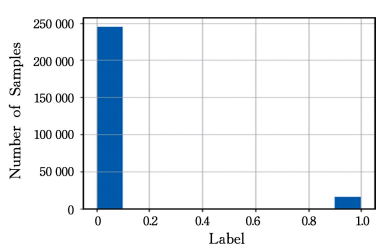


图 1　再次消费人群分布

Fig. 1　Distribution of repeat shopping users

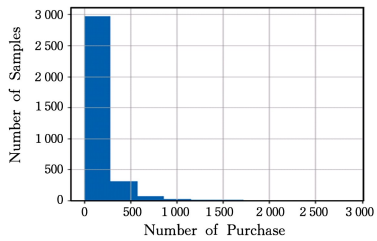


图 2　用户行为记录数分布

Fig. 2　Distribution of users’ shopping logs

3　特征工程

本文提取了一系列统计特征，利用 word2vec 词嵌入模型[6]提取了部分隐式特征，期望使用不同的特征类型覆盖不同的物理意义。这些预测属性从消费者历史点击或购买映射出消费者的购物兴趣，也可以从消费者类型映射出不同商家之间的相似程度。一个消费者可能对于某个商家来说是新客户，但是如果该消费者在促销活动之前对该商家的店铺或商品有过点击或添加收藏等行为，那么这些信息对消费行为预测也具有很重要的价值。

3.1　特征总览

本文依据提取特征的主体把特征归纳为 7 类：商家特征、用户特征、品牌特征、品种特征、商家处不同品牌的特征、用户在对应商家的特征以及商家处不同商品种类的特征。以上每类又分为统计特征和隐式特征。如图 3 所示，本文对这 7 类特征进行了总结。其中统计特征是指使用统计分析方法直接从数据集上计算获得的特征，而隐式特征则指使用一些复杂的算法学习获得的特征，如本文使用的 word2vec 算法、矩阵分解[8]方法等。其中取得最好结果的模型使用了 3273 维特征。

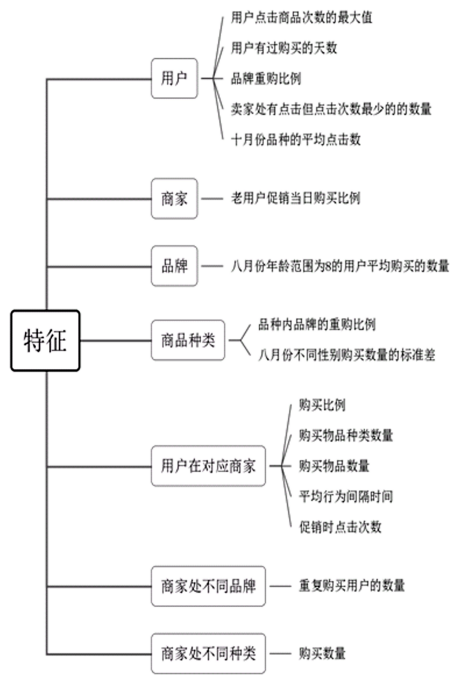


图 3　每一类特征示例

Fig. 3　Examples of each type of feature

3.2　统计特征

本文基于专家知识依据促销前与促销时两种情形，针对消费者 8 个年龄段与 3 种性别选项（男、女以及未知）分别进

行的点击、购买、收藏和加入购物车 4 种行为方式生成了 148 种不同的特征[9]。与非促销期间相比，促销期间购物次数波动可能反映出消费者是否为价格敏感型，销售量随季节波动较大的商家可能被归类为提供季节性商品的商家，再如不同类型的消费者、商家的购买和销售商品的类型也会有很大的不同，不同年龄段的消费者消费习惯不同，选购商品的类型也不同。依据这些假设，本文使用统计分析方法针对前述 7 类特征主体依照时间跨度提取了包括点击数、购买数等六大类统计特征，如表 4 所列。

表 4 统计特征

Table 4 Statistical features

Feature type	Content
Target entity	User, Seller, Catalog, Brand, User_seller, Seller_catalog, Seller_brand
Feature entity	User, Seller, Catalog, Brand, Item, Day, Month
Aggregation type	Count, Unique, Mean, Std, Trend, Mode, CountRepeat, RepeatPercentage, Max, Min, Median
Duration	Month, AllTime, OneWeekBeforePromotion, PromotionDay
Interesting value	Gender, ActionType, AgeRange
Transformation type	TimeSincePrevious

Target Entity 是需要输出的对应特征的主体，Feature Entity 是用来计算特征的主体，Aggregation type 是指用来计算特征的方法，Duration 是用来计算特征日志的时间跨度，Interesting value 是关注的值，Transformation 是对特征做的转换。

由于比率特征能够反映某种行为出现的可能性，进而可以用来对消费者的行为做预测，本文计算了重购率和购买率等比率特征。重购率是指日志范围内在对应的某个商家，在不同日期有多次购买的消费者数量占总的有过购买的消费者数量的比例。购买率是指日志范围内在对应的某个商家，有过购买行为的用户数占所有有交互行为即有点击、收藏、购买等行为的用户数的比例。重购率能够从侧面反映出某个商家是否出售的是消费者可能不会在短时间内重复购买的贵重物品，而购买率则体现了商品本身的价值属性。

3.3 隐式特征

本文使用 word2vec 和矩阵分解的方法生成隐式特征。对于词向量技术，选用了 skip-gram 模型[6]，参数设置如表 5 所列。本文使用商家 id 或者种类 id 作为单词，将消费者的行为按照时间顺序生成语句，通过 word2vec 来获得每个 id 的嵌入。本文同时考虑了消费者和商家之间的交互信息，并且使用该交互信息作为隐式反馈矩阵。根据上面对数据的描述，本文将再次消费的标签为 −1 的消费者、商家在矩阵中相应的值设为 1，而将再次消费标签为 0 或 1 的消费者、商家在矩阵中相应的值设为 0。接着使用矩阵分解将这个矩阵分解成两个低维的矩阵，随后将消费者矩阵和商家矩阵的内积作为一个隐式特征。其中，矩阵分解使用 Scikit-learn[7] 中的 NMF 实现，采用坐标下降法，维数设置为 128，其他为默认参数。

表 5 参数设置

Table 5 Parameter setting

Embedding_size	Skip_window	Num_skips	Num_sampled
32	1	2	5

4 特征分析

本文从训练好的模型中输出了特征重要程度并且做了简单的特征分析，虽然设计了很多隐式特征，但是发现模型在没有隐式特征的情况下就能取得比较好的效果，隐式特征只在一定程度上提升了模型的性能。本文同时为每个 id 随机生成了特征向量并直接输入到模型中，通过对比随机生成的特征和隐式特征发现，两个模型性能相差不大。如何识别出关键的特征来减少模型复杂度十分重要[10]。本文仅选取了最好的 78 个特征，*AUC* 就达到了 0.695。其根据下面所述的特征重要性排序选取最重要的几维特征。

4.1 特征的分析与选择

本文在使用 lightGBM[11] 训练的过程中，根据特征使用的次数以及对模型预测准确率的提升程度为依据来对特征重要性进行排序，重要性从高到低排序，其部分结果如图 4 所示。特征重要度的具体数值由 lightGBM[11] 直接输出。由图 4 可知，可以优先选择的特征首先是促销当天再次消费的消费者的比例，其次是消费者在对应商家的购买物品数量，接着是消费者在对应商家的点击、购买或收藏的商品数等。

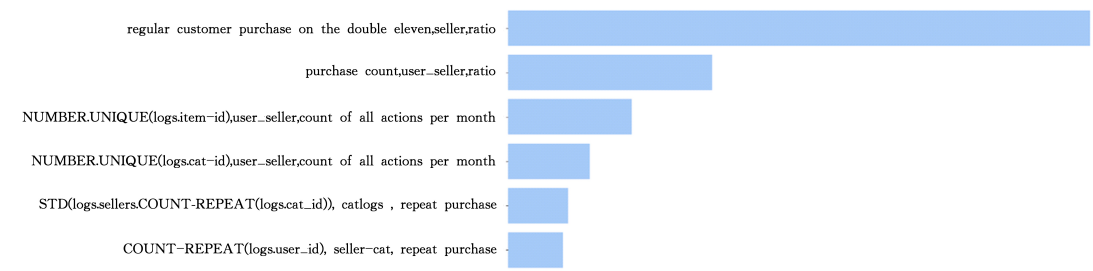


图 4 特征重要性排序

Fig. 4 Feature importance ranking

4.2 特征分布和相关性

本文对 3.1 节分析获得的几个重要的特征进行了相关性分析，并基于可视化方法对分析结果做了研究[12]，如图 5 的特征相关性矩阵所示。可以看出，本文所设计的特征之间的相关性都很低。促销当天，在目标商家虚拟店铺中再次消费的消费者所占的比例是指在促销之前就在对应商家有过消费并且在促销当天再次消费的消费者的数量除以促销当天在对应商家有过购买的所有消费者的数量。

图 6 是这个变量对应再次消费与非再次消费两种标签的特征分布情况并用核密度估计将概率密度函数可视化的结果。图 6 的横坐标是该属性特征的取值，纵坐标是概率密度函数的值。可以看出二者的概率密度函数基本一致，很难把两者区分开。从图 5 中也可以看出单个特征和标签之间的相关性很低。

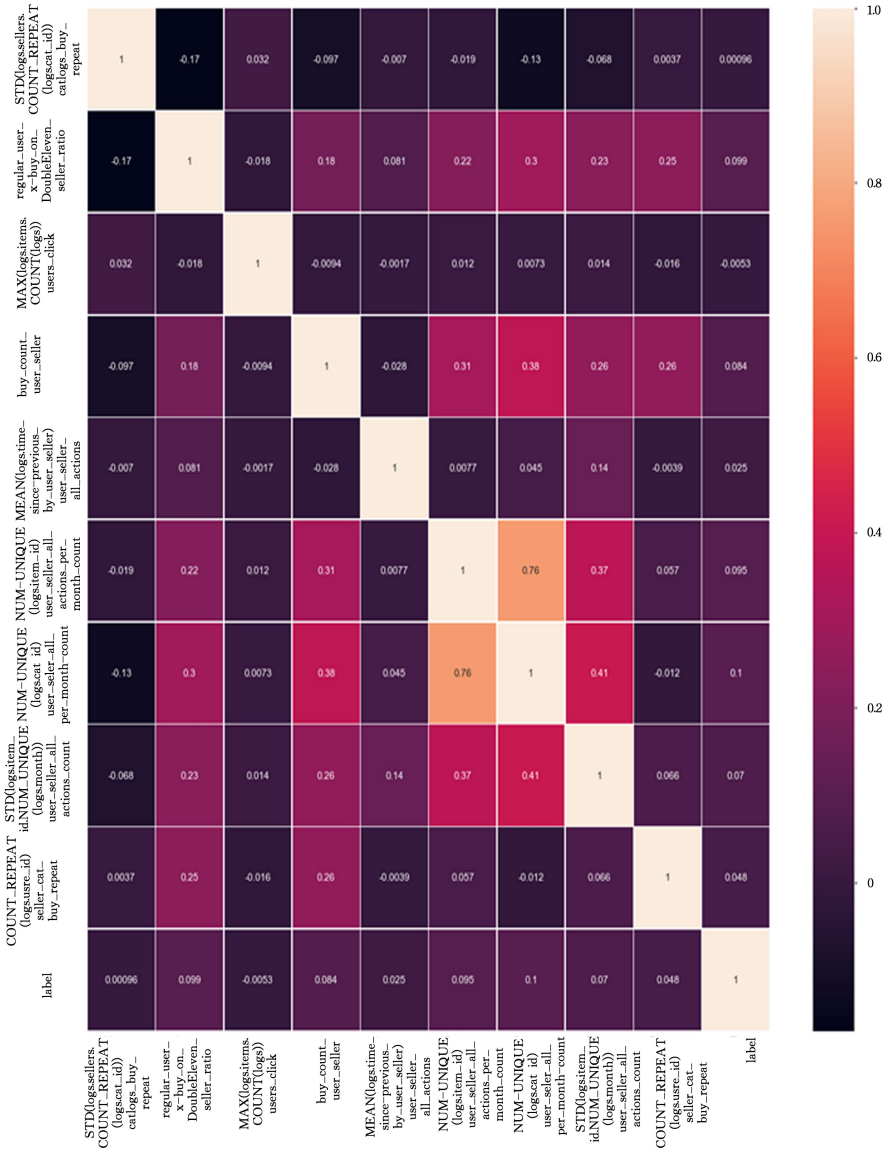


图 5 相关性矩阵

Fig. 5 Correlation matrix

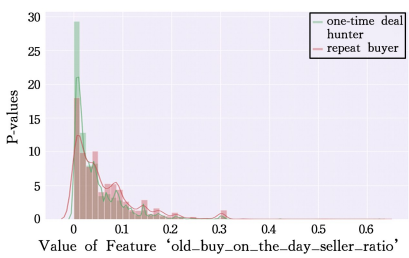


图 6 消费者于对应商家特征的分布曲线

Fig. 6 Distribution of feature between customers and sellers

5 实验

本文对比了多个不同的算法,其中性能最好的方法是多 lightGBM 模型融合,其 *AUC* 达到了 0.7018。

5.1 模型对比

实验对比了 lightGBM[11] 和 XGBoost[15] 两种方法,本文使用 Hyperopt[16] 分别对两个模型进行参数搜索。最终 lightGBM 参数设置如表 6 所列,xgboost 参数如表 7 所列。实验结果表明,lightGBM 的 *AUC* 值达到 0.6997,而 XGBoost 的 *AUC* 值为 0.6916。通过对比可以看出,lightGBM 的性能优于 XGBoost。

表 6 LightGBM 参数设置

Table 6 Parameter setting of lightGBM

Objective	Binary
Lambda_l2	1.4618
Learning_rate	0.0164
Max_bin	31
Min_data_in_leaf	3231
Num_leaves	39
feature_fraction	0.3892
Reg_alpha	4.9016

表 7 XGBoost 参数设置

Table 7 Parameter setting of XGBoost

Objective	binary
gamma	8.8003
eta	0.0892
max_depth	6
min_child_weight	772
subsample	0.8106
colsample_bytree	0.7421
reg_alpha	0.0674

5.2 模型融合策略

本文尝试划分不同的数据，选取不同的随机种子训练多个模型，并采取多种不同的模型融合策略[13-14]，其中包括对模型的预测概率取平均、非线性函数以及根据验证集上的性能赋予不同模型不同的权重等方法。通过实验可知，单模型的 *AUC* 达到了 0.696043，使用多个模型取平均其 *AUC* 值达到了 0.69959。使用公式 $Sigmoid(\frac{1}{x}\sum_{i=1}^{x}Logit(p(i)))$ 对多个模型进行融合的 *AUC* 值达到了 0.69997，其中 x 为模型融合个数，$p(i)$ 为单个模型输出的再次消费概率。而根据模型在验证集上的 AUC 设置权重，进行加权平均的 *AUC* 值为 0.69573。并且根据表 8 可以看出实验中模型融合数量越多，其性能也会越来越好。

表 8 模型数量与性能关系

Table 8 Relationship between number of models and performance

模型数量	1	27	54
AUC	0.69604303	0.69953254	0.699970

结束语 本文详细介绍了如何设计特征以及特征的表现，针对再次消费预测问题，本文提出了一种预测消费者是否会再次消费的模型。通过大量的实验对模型进行了改进，实验结果表明，模型取得了较好的预测准确率，为以后的相关工作提供了很好的借鉴方案。未来将尝试在特征工程的相关问题上做进一步研究，尤其是从再次消费人群的角度设计更好的特征。

参 考 文 献

[1] JOO J. An Empirical Study on the Relationship between Customer Value and Repurchase Intention in Korean Internet Shopping Malls[J]. Journal of Computer Information Systems, 2007, 48: 53-62.

[2] Kaggle. Acquire-valued-shoppers-challenge[OL]. www.kaggle.com/c/acquire-valued-shoppers-challenge/.

[3] AliCloud. Repeat Buyers Prediction-Challenge the Baseline [OL]. tianchi.aliyun.com/competition/entrance/231576/introduction? spm=5176.12281949.1003.8.708d2448oQdTSf.

[4] HEATON J. An empirical analysis of feature engineering for predictive modeling[J]. SoutheastCon, 2016, 2016: 1-6.

[5] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. The Journal of Machine Learning Research, 2003, 3: 1157-1182.

[6] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and their Compositionality [J]. NIPS, 2013, 2: 3111-3119.

[7] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python[J]. Journal of Machine Learning Research, 2011, 12: 2825-2830.

[8] KOREN Y, BELL R M, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42: 30-37.

[9] KANTER J M, VEERAMACHANENI K. Deep feature synthesis: Towards automating data science endeavors[C]// 2015 IEEE International Conference on Data Science and Advanced Analytics(DSAA). Paris: IEEE, 2015: 1-10.

[10] MOLINA L C, BELANCHE L, NEBOT A. Feature selection algorithms: A survey and experimental evaluation[C]// 2002 IEEE International Conference on Data Mining, 2002. Maebashi City: IEEE, 2002: 306-313.

[11] KE G L, MENG Q, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS'17). Long Beach: Curran Associates Inc, 2017: 3149-3157.

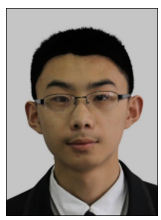
[12] HUNTER J D. Matplotlib: A 2D graphics environment [J]. Computing in Science & Engineering, 2007, 9(3): 90-95.

[13] ZHOU Z H. Ensemble Methods: Foundations and Algorithms [M]. Chapman and Hall: CRC, 2012.

[14] WOLPERT D H. Stacked generalization[J]. Neural Networks, 1992, 5: 241-259.

[15] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. California: ACM, 2016: 785-794.

[16] BERGSTRA J, YAMINS D, COX D D. Hyperopt: A Python library for optimizing the hyperparameters of machine learning algorithms[J]. Computational Science & Discovery, 2013, 8(1): 014008.



LV Ze-yu, born in 1998, postgraduate. His main research interests include machine learning and so on.

CHEN Dong-ming, born in 1968, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include complex networks, social network analysis, machine learning and information security.