基于XGBoost特征选择的幕课翘课指数建立及应用

宋国琴[1]，刘 斌[2]

(1. 西华师范大学教育信息技术中心 四川 南充 637000；2. 电子科技大学计算机科学与工程学院 成都 611731)

【摘要】翘课行为反应了幕课的质量问题，也是在线教育的核心问题之一。该文通过对真实的在线教育数据进行分析，结合在线教育领域的先验知识，针对数据中的丰富海量的特征问题，提出了基于XGBoost特征重要度计算和分类的翘课特征选择方法，并建立了在线教育的翘课指数(DOI)。基于学堂在线数据集提取的海量特征的实证分析表明，基于XGBoost的特征选择方法比其他经典特征选择方法具有更好的效果。在数据集的不同时间点上使用翘课指数模型作翘课预测，验证了翘课指数的有效性。

关 键 词 翘课指数； 特征选择； 幕课； XGBoost

中图分类号 TP391 **文献标志码** A doi:10.3969/j.issn.1001-0548.2018.06.019

The Establishment and Application of Drop-Out-Index of MOOCs Based on XGBoost Feature Selection

SONG Guo-qin[1] and LIU Bin[2]

(1. Education and Information Technology Center, China West Normal University Nanchong Sichuan 637000;

2. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731)

Abstract Dropout of classes reflects the quality of MOOCs, which is the key issue of online education. In order to predict the dropout rate in advance, this paper presents an efficient prediction framework based on the analysis on real online education data and the prior knowledge of online education. This presented framework combines the feature importance learning and the selection by the classification algorithm of XGBoost, and establishes a Drop-Out-Index (DOI) for online courses. Experiments analysis on massive features extracted from the online-data of XueTang website shows that the feature selection method based on XGBoost achieves better results than other feature selection methods. The validity of DOI has also been verified by testing on different time points in the data set.

Key words drop out index; feature selection; MOOCs; XGBoost

大规模开放在线课程(massive open online courses, MOOCs)因为免费且学习环境宽松等特性，近年来得到了爆发式的增长。同时，MOOCs的宽松学习环境引发了辍学率居高不下的情况，辍学使在线教育教学有效性大大降低[1-2]。利用在线教育的大数据，结合机器学习技术对在线教育相关情况进行分析，是目前数据研究的热门方向之一。

另一方面，目前虽然有基于机器学习的在线教育的翘课预测工作[3]，但没有关于翘课指标量化的研究工作。翘课指数是预测翘课概率的指标，可以反映在线课堂的学习情况和变动程度、监测学习者的发展动态及预测翘课的发展趋势。目前国内外鲜有提供在线教育的公共大数据，KddCup2015 [4]提供了宝贵的应用实例。对基于KddCup2015的翘课指数的紧密追踪和预测，将为在线教育中的各种变化和改革提供指标参考，具有重要的现实意义。

在指数挖掘过程中，需要通过特征选择[5-8]实现高维特征的物理降维。特征选择往往伴随着机器学习过程。XGBoost[9]的特征选择基于初始特征集建立分类模型，考察特征在模型中的表现，得到特征的重要性，依据重要度进行特征子集搜索和评价，最终产生最优子集，是一种兼具嵌入式和过滤式特征选择方法。

本文基于数据建立丰富的特征并预处理，使用基于XGBoost算法做特征选择，从与翘课相关的海量特征中找到最佳特征子集，形成在线教育的翘课

收稿日期：2017－05－24；修回日期：2018－03－26

基金项目：中央高校基本科研业务费基础研究项目(ZYGX2014J058)；四川省教育厅项目(16ZA0171)

作者简介：宋国琴(1979–)，女，副教授，主要从事数据挖掘、机器学习、模式识别在教育方面的应用等方面的研究.

指数DOI(dropout index)，为监测学习者学习动态、反映整体课程的学习情况以及预测学习者学习趋势提供参考和依据。

1 翘课特征建立及预处理

1.1 基于KDDCup2015数据的特征建立

KDDCup2015提供了学堂在线网站39门课程一个月的学习日志数据，记录数逾百万，数据规模达600 MB。这些数据通过关系数据库连接并提取特征的空间复杂度和时间复杂度都非常高，其主要的关系数据表及实体如图1所示。

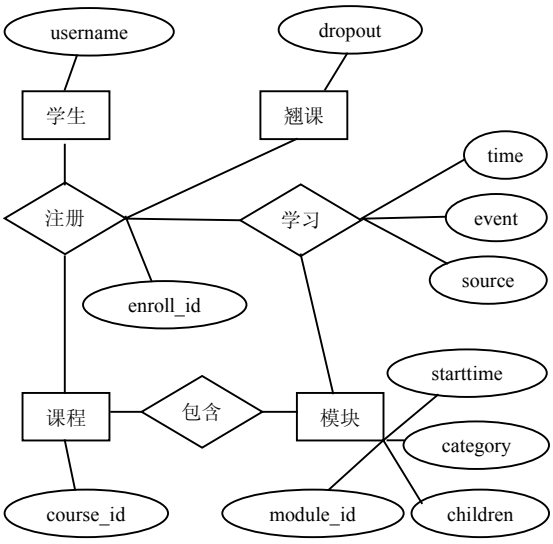


图1 翘课定义KDDCUP2015数据E-R图

翘课和辍学都是学习中的逃课行为，将辍学延伸到翘课，能准确和全面地反映幕课教学中的负面问题。翘课问题定义如图2所示。

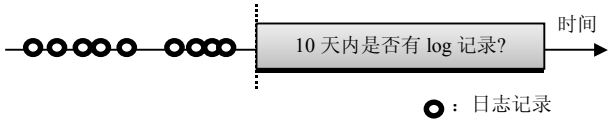


图2 翘课定义

根据KDDCup2015比赛任务规定，翘课界定标准为某注册号在某个时间点之后10天内没有日志记录。根据对学堂在线网站的调查统计，学堂在线的课程绝大多数是一周更新一次课程，学习者可以在每周课程内容上线后到下周课程推出前的任意时间登陆并学习。一门课程的更新时间有可能变化，最坏情况下，上次课周一更新而下次课周五更新，中间有10天间隔。因此，10天内没有学习记录，足够说明一次翘课行为。

本文按照行业经验，根据对翘课行为的理解和对数据集的挖掘，依据尽可能丰富和完备的原则，建立了翘课特征。这些特征分别从课程统计情况、用户课程行为、用户网站行为、不同时段行为、特殊时段如最后几次课行为等5个方面提取，特征总数为1 658，样本总数为120 542。

1.2 特征预处理

特征预处理的目的是在特征选择前做一次特征粗选，以简化后续特征选择过程。皮尔逊相关性系数是衡量变量间相关性的常用指数，设$(x_1,x_2,\cdots,x_n)$是一个n维随机变量，对于任意x_i与x_j的相关系数$\rho_{ij}(i,j=1,2,\cdots,n)$存在，则有：

$$\rho_{ij}=\frac{\sum(f_i-\overline{f_i})(f_j-\overline{f_j})}{\sqrt{\sum_{k=1}^{m}(f_{i,k}-\overline{f_i})^2}\sqrt{\sum_{k=1}^{m}(f_{j,k}-\overline{f_j})^2}} \tag{1}$$

已知ρ_{ij}为x_i和x_j两个特征的皮尔逊相关系数，设置0.9为高相关阈值，对于$\rho_{ij}\geqslant 0.9$的两个高相关度特征，再按照特征的总相关度进行筛选。特征的总相关度是为了考察某个特征x_i与其他所有特征的相关性，总相关度越高，特征的独立性越弱，可以相对地剔除。对于n个特征中的某个特征x_i设定义为：对于n维特征，为了考察某个特征x_i与其他所有$n-1$个特征的相关性，定义特征x_i总相关度为$r_i=\sum_{j=1,j\neq i}^{n}\rho_{ij}$。对于两个高相关特征$x_i$和$x_j$，比较$r_i$和$r_j$值，如果$r_i>r_j$则保留$x_i$，否则保留$x_j$。

根据此方法，特征数量缩减为原来的80%，实现了初步有效的预处理。

2 特征选择

2.1 特征选择流程

XGBoost在训练过程中为了提高生成新树的效率，会在每轮迭代中给出各个特征的重要度评分，从而表明每个特征对模型训练的重要性，为下一次迭代建立梯度方向的新树提供依据。这种统计出的特征重要性，可以直接作为特征选择的依据。

基于XGBoost的特征选择的步骤为：

1) 基于所有特征进行XGBoost分类；

2) 基于生成的模型过程中的信息，得到特征变量的重要性(FI)并按降序排序；

3) 按照前向搜索原则，即选择FI值最高的若干特征，生成特征子集；

4) 在特征子集上进行分类实验。特征子集的评价主要考察子集的分类能力，分类能力从分类结果中获得；

5) 重复步骤3)和4)，直到所有特征都被选择；

6) 考察所有子集的分类情况，选择AUC值相对较高同时特征数量较少的子集作为最优特征子集。

2.2 XGBoost分类方法

传统的适合二分类机器学习方法有逻辑回归、SVM、随机森林及Boosting算法[10]等等。XGBoost (extreme gradient boosting)也称为极端梯度提升，是一种基于梯度提升机制GBM(gradient boosting machine)的改进算法，在KDDCUP2015竞赛中的前3名均采用了此算法。它既是一个算法工程，也是已有算法的更新。其基本思想是选择部分样本和特征生成一个简单模型(如决策树)作为基本分类器，在生成新模型时，学习以前模型的残差，最小化目标函数并生成新的模型，此过程重复执行，最终产生由成百上千个线性或树模型，组合为准确率很高的综合模型。其核心在于，新的模型在相应损失函数梯度方向建立，修正“残差”的同时控制复杂度。XGBoost的目标函数包含两部分：

$$\text{obj}(\theta)=L(\theta)+\Omega(\theta) \tag{2}$$

1) 损失函数

$$L(\theta)=\sum_{i=1}^{n} l(y_i,\hat{y}_i^{(t)}) \tag{3}$$

式中，y_i为第i个样本标签；$\hat{y}_i^{(t)}$为i样本第t次迭代的预测值。$L(\theta)$部分使用LogLoss作为损失函数，采用泰勒展开式来逼近，其中同时使用了一阶导数和二阶导数。设g_i和h_i分别为一阶和二阶导数：

$$g_i=\partial_{\hat{y}_i^{(t-1)}} l(y_i,\hat{y}_i^{(t-1)}) \tag{4}$$

$$h_i=\partial^2_{\hat{y}_i^{(t-1)}} l(y_i,\hat{y}_i^{(t-1)}) \tag{5}$$

则有：

$$L(\theta)\approx\sum_{i=1}^{n}\left[l(Y_i,\hat{Y}_i^{(t-1)})+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right] \tag{6}$$

2) 正则项

本文基本模型为回归树，树的复杂度由两方面决定，一是叶子的个数，二是树的结构部分权重。最终的树的复杂度公式定义为：

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \tag{7}$$

式中，T为叶子结点个数；w为每个叶子的权重，叶子的权重值表达了在这个节点上的翘课的可能性。第二部分使用了w的L2范数，可以更好地避免过拟合。

综合以上两部分，略去常数项部分$l(y_i,\hat{y}_i^{(t-1)})$，目标函数可变化为：

$$\text{obj}(\theta)\approx\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \tag{8}$$

$$\text{obj}(\theta)=\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j+\frac{1}{2}\left(\sum_{i\in I_j} h_i+\lambda\right)w_j^2\right]+\gamma T \tag{9}$$

定义：

$$G_i=\sum_{i\in I_j} g_i, H_i=\sum_{i\in I_j} h_i$$

目标函数变为：

$$L(\theta)=\sum_{j=1}^{T}\left[G_j w_j+\frac{1}{2}(H_j+\lambda)w_j^2\right]+\gamma T \tag{10}$$

这时目标函数变成关于w的二次函数，目标函数最小时的最优权重为：

$$w_j^*=-\frac{G_j}{H_j+\lambda} \tag{11}$$

代入目标函数变为：

$$\text{obj}(\theta)=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{12}$$

2.3 特征重要度(FI)计算

基于XGBoost分类的特征选择时，特征重要度计算融入了分类过程。在每一轮的迭代中建立一棵新树，树中的分支节点即为一个特征变量，计算这些节点的重要度。特征重要度是基于一个特征被选择为此树的分裂节点的平方改进。在每次选择一个特征加入树中作为分裂节点前，会用贪心法枚举所有可能的分割点，从中选择增益最好的分裂点。

最好的分裂点对应最大的增益，增益计算公式为：

$$G_{\text{ain}}=\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}-\gamma \tag{13}$$

好的特征及分裂点能改进单棵树上的平方差，改进得越多，这个分裂点越好，这个特征越重要。当所有树建立完成后，将计算到的结点重要性在森林中求平均。特征被选择作为分裂点的次数越多，重要度也会越高。

对于有J个分支结点的树T，如果j被选择为此树上的分裂变量，则计算所有分支结点t上的平方误差和，即为特征j在这棵树上的的重要度为：

$$\hat{I}_j^2(T)=\sum_{t=1}^{J-1}\hat{I}_j^2 P(\upsilon_t=j) \tag{14}$$

式中，$\hat{I}_j^2$是某结点t的平方误差改进。设$\bar{y}_l,\bar{y}_r$分别是左右子树的预测均值，w_l,w_r分别是节点左右子树

节点的权重和，有：

$$I^2(R_l,R_r)=\frac{w_l w_r}{w_l+w_r}(\bar{y}_l+\bar{y}_r)^2 \tag{15}$$

对于有M棵树的森林，将每棵树上特征t的重要度相加再平均，得到最终的重要度：

$$I_t^2=\frac{1}{M}\sum_{m=1}^{M}I_t^2(T_m) \tag{16}$$

3 实验结果分析及DOI特征系统确立

目前，出现了很多种特征权重计算方式[11-17]。这些文献中的数据集来自不同的领域，特征数或样本数大多小于本文数据的规模，因此本文选择了几种经典的、具有普适性和稳定性的FI方法作对比实验，包括Relief、Fisher、信息增益、皮尔逊相关系数。

另一方面，使用逻辑回归(LR)与支持向量机(SVM)作分类方法的对比。支持向量机在超过10万的大样本集上工作非常困难，因此在分类前使用降采样[18]减少样本数至2万左右。本特征集数据的正负样本比例为：1:3.8，考虑到样本的不平衡性，实验中使用了采用受试者工作特征曲线(ROC)下围面积(AUC)作为分类评价指标。

3.1 对比实验

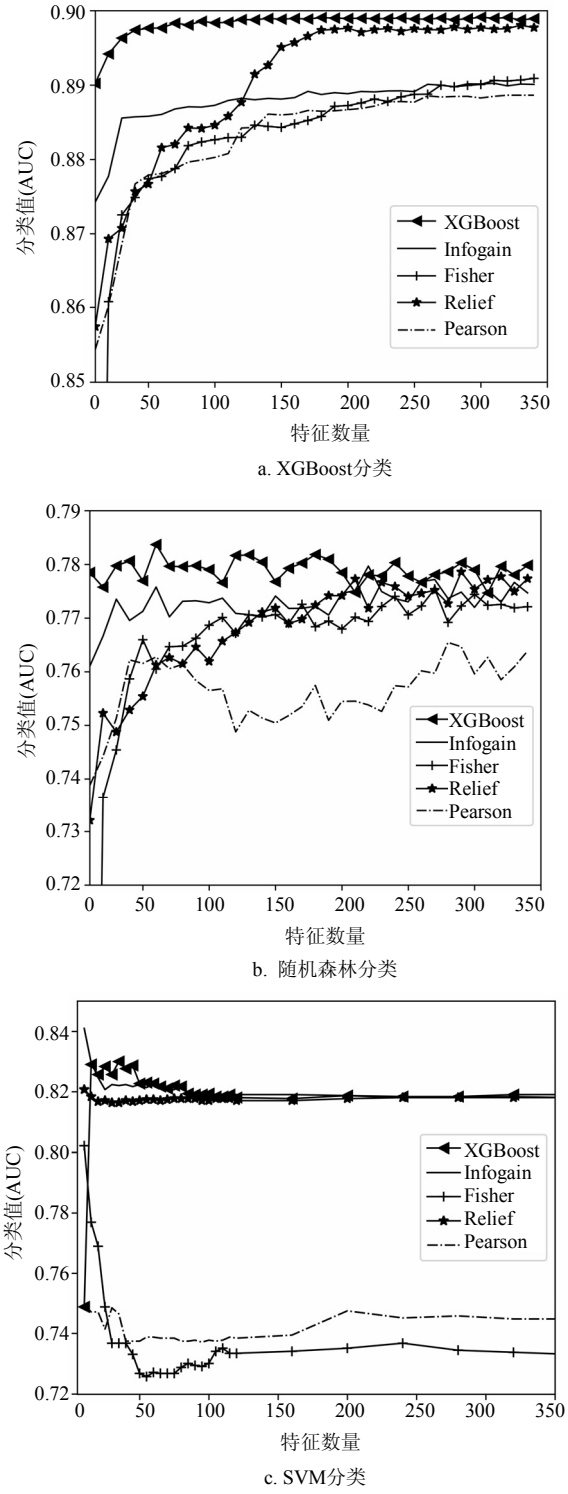
实验中，不同的特征重要度计算结合不同的分类方法，均呈现一定的共性：随着子集的特征数量增加，AUC值快速上升，最多增加到约400以后，AUC值趋于平稳或震荡下降。因此，不同实验结果的子集数量比较范围缩小至[10,400]。特征子集选择过程中的分类实验信息如表1、表2及图3所示。

表1 不同打分方法的耗时

打分方法	计算时间/min
皮尔逊系数	0.16
XGBoost	3.7
Fisher	79.56
信息增益	651
Relief	676

通过表1可以看出，基于XGBoost的特征计算耗时非常小，这在大维度数据集上非常重要。

图3显示了5种不同的特征重要度计算方法下产生的子集，使用3种分类方法的对比实验显示，基于XGBoost的特征重要度在3种分类下均有明显的优势，其中，在XGBoost分类下的分类值和收敛效果是最好的。



a. XGBoost分类

b. 随机森林分类

c. SVM分类

图3 不同FI方法配合不同分类器的性能比较

3.2 最优特征子集

图4单独显示了XGBoost特征选择下子集的分类信息。综合AUC全局和局部的最优值，以及特征数量最小化，选择重要度最高的前135个特征组成DOI特征。最优特征子集的信息如表2和表3所示。

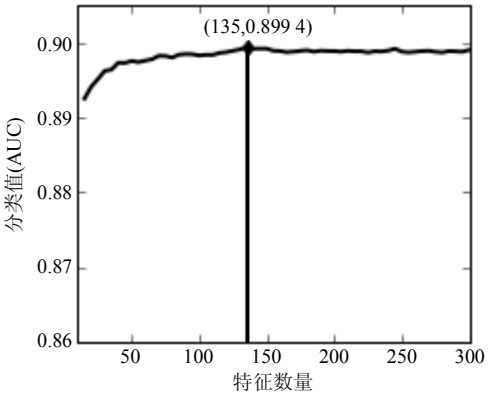


图4　基于XGBoost特征重要度的最优子集

表2　最优特征子集内容

特征	数量
用户在注册课程中的学习	54
最后一天行为	40
用户在整个网站的行为	35
统计信息	4
课程注册信息	1
时间信息	1

表3　最优特征子集表现

特征集	数量	提取时间/min	分类值
所有特征	1 339	40	0.899 8
最优子集	135	15	0.899 4

其中，重要度最高的特征包含最后一天访问其他课程对象的用时，最后一天关闭网页的用时等是非常有指示意义的特征。从表2中可以推断出，某用户翘一门课时，在其他同期课程也可能翘课；同时，用户在课程操作、网站操作上会有不同表现，而且愈临近翘课越明显。重点关注以上方向的趋势，可在很大程度上主导对翘课的预测。

如表3所示，最优特征子集只有135个特征，数量不到原来的1/10，而KDDCup2015前10名队伍的特征数量大多在1 000以上[5]。因为特征子集数量小，也缩短了特征提取的时间，而分类性能却下降极少。

4　DOI指数的建立及应用

在最优特征子集上使用XGBoost算法构造决策森林，森林中各棵树的预测值加性求和，再将结果进行逻辑回归，得到DOI指数的值。

$$\hat{y}_i = \frac{1}{1+e^{-\sum_{k=1}^{K} f_k(x_i)}} \quad f_k \in \mathcal{F} \tag{17}$$

式中，$\mathcal{F}$ 为所有树的函数空间；f_k 为单棵树，其中包含了特征到分值的映射。取0.5为DOI指标的基准线，大于0.5表示翘课概率增加，小于0.5表示翘课概率减少。DOI值在0~1范围内变化，偏离0.5的大小表示翘课可能与否的程度。

为了在更多时间点上验证DOI指数，实验在原数据集上以3天为间隔，产生前3,6,9…天的9个数据集，在这些新的数据集上提取DOI需要的特征和翘课标签，将特征值代入DOI指标体系模型作回归预测，实验结果如图5所示。

从图5可以看出，在缩减大部分特征之后，在不同的时间点上，DOI指标拥有不错的指示性。在开课初期，由于用户刚开始学习，相关的统计信息不够丰富，导致预测准确率相对偏低。

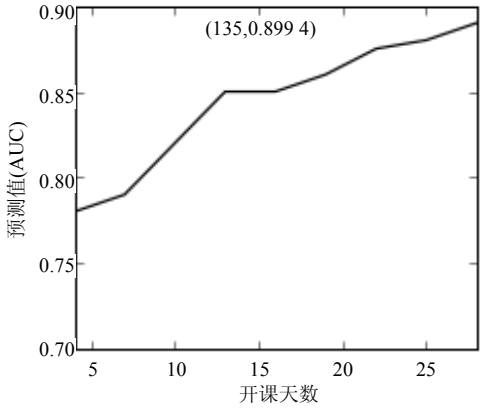


图5　DOI在多个时间点的预测结果

5　结 束 语

从实验结果看，本文使用的基于XGBoost分类相结合的特征选择及预测方法具有合理性、有效性、可行性、先行性，能够较好地在不同时间点上预测学生的翘课，为在线教育中建立翘课或辍学指标提供了有效的量化方法。DOI指数具备一定的先行性，有很大的应用价值。

最近几年，深度学习使图像、文本等数据的特征工程自动化了，但是人类行为数据的特征工程智能化程度较低，仍然很具有挑战性。文献[19]提出并开发了一种深度特征合成算法，一定程度上实现了特征工程的自动化并取得不错的分类结果，这将是人类行为数据特征提取和降维的重要研究方向。

KDDCup2015数据集还有诸多限制，本文得到的最优特征子集放在同类其他数据集上的有效性有待进一步考察。实际应用中，如能加入用户信息或更多学习行为数据如视频观看的进度等，结合人类行为观测到的阵发性、记忆性和非马尔可夫特性[20]等特征，将进一步提升DOI指数的精确性和泛化性。

本文研究工作还得到南充市研发基金(17YFZJ0020)的支持，在此表示感谢。同时，感谢电子科技大学Smile实验室及徐增林教授为本文提供的技术、资料、信息、物质上的帮助。

参 考 文 献

[1] JOSEP G, JULIA M. Rethinking dropout in online higher education: The case of the Universitat Oberta de Catalunya[J]. The International Review of Research in Open and Distributed Learning, 2014, 15(1): 293-308.

[2] DANIELl F O O, JANE S, RUSSELL B. Dropout rates of massive open online courses: Behavioral patterns,6th International Conference on Education and New Learning Technologies[C]//EDULEARN14 Proceedings. [S.l.]: IATED, 2014: 5825-5834.

[3] TAN M, SHAO P. Prediction of student dropout in e-learning program through the use of machine learning method[J]. International Journal of Emerging Technologies in Learning, 2015, 10(1): 11-17.

[4] SIGKDD, KDD Cup 2015-Predicting dropouts in MOOC [EB/OL].(2015-8-4). http://www.kaldcup2015.com/information.html.

[5] ISABELLE G, ANDR´E E. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2002, 3(6):1157-1182.

[6] ZUHAIR H, SELAMAT A, SALLEH M. Feature selection for phishing detection: a review of research[J]. International Journal of Intelligent Systems Technologies & Applications, 2016, 15(2): 147.

[7] VERGARA J R, ESTÉVEZ P A. A review of feature selection methods based on mutual information[J]. Neural Computing and Applications, 2014, 24(1):175-186.

[8] CHIN A, MIRZAL A, HARON H, et al. Supervised, unsupervised and semi-supervised feature selection: a review on gene selection[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2016, 13(5): 971-989.

[9] CHEN T, CARLOS G. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD '16). [S.l.]: ACM, 2016: 785-794.

[10] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Zhi-hua. Machine learning[M]. Beijing: Tsinghua university press, 2016.

[11] 郑义, 姚建铨, 吴峰, 等. 用修正的RELIEF方法测量高速空气流瞬时速度的理论研究[J]. 光学学报, 1996, 16(8): 126-129.
ZHENG Yi, YAO Jian-quan, WU Feng, et al. Theoretical study on the measurement of instantaneous velocity of high speed air flow with modified RELIEF method[J]. Acta Optica Sinica, 1996, 16(8): 126-129.

[12] 金理钻, 屠珺, 刘成良. 基于迭代式RELIEF和相关向量机的黄瓜图像识别方法[J]. 上海交通大学学报, 2013, 47(4): 602-606, 612.
JIN Li-zuan, TU Jun, LIU Cheng-liang. Cucumber image recognition method based on iterative RELIEF and relevance vector machine[J]. Journal of Shanghai Jiaotong University, 2013, 47(4): 602-606, 612.

[13] REN Y, ZHANG G, YU G, et al. Local and global structure preserving based feature selection[J]. Neurocomputing, 2012, 89: 147-157.

[14] WANG S, LI D. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification[J]. Expert Systems with Applications, 2011, 38(7): 8696-8702.

[15] HU Q, CHEN X, ZHANG L, et al. Feature evaluation and selection based on neighborhood soft margin[J], Neurocomputing, 2010, 73(10-12): 2114-2124.

[16] HE X, DENG C, PARTHA N. Laplacian score for feature selection[C]//Proceedings of the International Conference on Advances in Neural Information Processing Systems 18 (NIPS 2005). Vancouver, Canada: [s.n.], 2005: 505-512.

[17] ZHANG D, CHEN S. Constraint score: a new filter method for feature selection with pairwise constraints[J]. Pattern Recognition, 2008, 41(5): 1440-1451.

[18] 林舒杨, 李翠华, 江弋, 等. 不平衡数据的降采样方法研究[J]. 计算机研究与发展, 2011, 48(s3): 47-53.
LIN Shu-yang, LI cui-hua, JIANG Yi, et al. Under-sampling method research in class-imbalanced data[J]. Journal of Computer Research and Development, 2011, 48(s3): 47-53

[19] JAMES M K, KALYAN V. Deep feature synthesis: Towards automating data science[C]//Proceedings of the The 3rd IEEE International Conference on Data Science and Advanced Analytics(DSAA). [S.l.]: IEEE, 2015: 1-10.

[20] 周涛, 韩筱璞, 闫小勇, 等. 人类行为时空特性的统计力学[J]. 电子科技大学学报, 2013, 42(4): 481-540.
ZHOU Tao, HAN Xiao-pu, YAN Xiao-yong, et al. Statistical mechanics on temporal and spatial activities of human[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(4): 481-540.

编 辑 蒋 晓