计算机科学
COMPUTER SCIENCE

基于DE-lightGBM模型的上市公司高送转预测实证研究

岑健铭, 封全喜, 张丽丽, 佟锐超

引用本文

岑健铭, 封全喜, 张丽丽, 佟锐超.基于DE-lightGBM模型的上市公司高送转预测实证研究[J]. 计算机科学, 2022, 49(11A): 211000017-7.

CEN Jian-ming, FENG Quan-xi, ZHANG Li-li, TONG Rui-chao. Empirical Study on the Forecast of Large Stock Dividends of Listed Companies Based on DE-lightGBM[J]. Computer Science, 2022, 49(11A): 211000017-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000017

基于 DE-lightGBM 模型的上市公司高送转预测实证研究

岑健铭[1,2]　封全喜[1,2]　张丽丽[1]　佟锐超[1]
1 桂林理工大学理学院　广西 桂林 541004
2 广西高校应用统计重点实验室　广西 桂林 541004
(cenjianming0819@foxmail.com)

摘　要　“高送转”现象指上市公司转增较大比例的股票。针对上市公司实施“高送转”现象的预测问题,文中提出了一种基于差分进化算法超参数优化的 lightGBM 模型(简记为 DE-lightGBM)。该模型主要包括两个方面:首先,利用差分进化算法调整 lightGBM 模型的损失函数中少数类别的权重以及正则项系数,以处理数据类别不平衡的问题;其次,以 F1 和 AUC 作为评价指标,再次利用差分进化算法优化 li-ghtGBM 模型的重要超参数变量,找到一组预测效果最优的参数组合。数值结果显示,DE-lightGBM 模型取得了较好的效果,F1 和 AUC 值分别为 0.5368 和 0.8734。提出的 DE-lightGBM 模型能够有效识别下一年将会实施“高送转”的上市公司。

关键词:高送转;差分进化算法;lightGBM;不平衡数据处理;机器学习
中图法分类号　TP181

Empirical Study on the Forecast of Large Stock Dividends of Listed Companies Based on DE-lightGBM

CEN Jian-ming[1,2],FENG Quan-xi[1,2],ZHANG Li-li[1] and TONG Rui-chao[1]
1 College of Science,Guilin University of Technology,Guilin,Guangxi 541004,China
2 Guangxi Colleges and Universities Key Laboratory of Applied Statistics,Guilin,Guangxi 541004,China

Abstract　Large stock dividends refers to the transfer of a large proportion of shares by listed companies. Aiming at the prediction problem of large stock dividends phenomenon implemented by listed companies,this paper proposes alightGBM based on Differential Evolution algorithm hyperparametric optimization(Named as DE-lightGBM). The model mainly includes two aspects: Firstly,Differential Evolution algorithm is used to adjust the weight of a few categories and the coefficient of regular term in the loss function of lightGBM to deal with the problem of data category imbalance. Secondly,taking F1 and AUC as evaluation indexes,Differential Evolution algorithm is used to optimize the important hyperparametric variables of lightGBM model again to find a group of parameter combinations with the best prediction effect. The numerical results show that the DE-lightGBM has achieved good results,and the F1 and AUC are 0.5368 and 0.8734 respectively. DE-lightGBM proposed in this paper can effectively identify the listed companies that will implement stock dividends next year.

Keywords　Large Stock Dividends,Differential Evolution,LightGBM,Unbalance treatment,Machine learning

1　引言

随着资本市场结构的优化,我国证券市场出现了各种题材股。近几年“高送转”题材成为带有炒作性质的“数字游戏”,引起不少上市公司的关注[1]。“高送转”是指高比例的送股和转增股票,通常将每 10 股送 5 股以上或者转增 5 股以上(含 5 股)的股票定义为“高送转”股票。

在实际交易中,股民普遍认为公司发布“高送转”预案是利好消息,然而现实的情况往往并非完全是利好的。主要有两方面原因,一方面是因为证监会的监管不及时,某些公司趁机进行股票送转,其目的并不单纯。这些公司正是看中了普通投资者对“高送转”股票的青睐,进行非正常的股票送转行为,导致市场上出现了大股东借助炒作,出现“减持套现”的情况[2]。另一方面,投资者认为在送转方案公布后,公司的股价会被拉升,持有“高送转”股票很有可能获得超额收益。但事实上,“高送转”预案后并不意味着股票一定会上涨,也可能存在股价暴跌的情况,所以股民进行“高送转”投资时需认清其本质以及送转背后隐藏的真实目的,理性投资,切勿“跟风”买入[3]。

从投资者的角度出发,投资者如果能够成功预测并预先持有“高送转”股票,则从中受益的可能性较大。这是因为公司在公布“高送转”消息之后,就会有大批股民选择买入该股票,从而促进股价的上涨,使投资者获利。因此,对于普通投资者来说,能提早预测出该股票,并且找到合适的时间节点

基金项目:国家自然科学基金(62166015,61763008,62166013);防城港市科学技术攻关项目(防财教[2014]42 号)
This work was supported by the National Natural Science Foundation of China(62166015,61763008,62166013) and Key Science and Technology Project of Fanghenggang city(Fangcaijiao[2014] No. 42).
通信作者:封全喜(fqx9904@163.com)

买入股票尤为重要。从监管机构的角度出发，因为公司的“高送转”行为受到市场的追捧，如果不及时监管，公司的大股东可能借此“减持套现”，甚至可能踩着监管“红线”实施“高送转”，导致大批股民严重亏损，扰乱市场秩序，所以通过预测能帮助监管机构及时监管下一年可能实施“高送转”的公司，谨防其送转动机不纯。本文根据上市公司的财务数据进行“高送转”预测实证研究，预测未来实施“高送转”股票的公司，并分析其重要的影响因素，不仅能够降低投资风险，使投资者获益。而且监管层可以重点关注未来可能实施“高送转”的公司，能够起到及时监管的作用，严查有不良动机的“高送转”的公司。

国内外学者对送转模型的构建最开始是通过 probit 模型[4]、逻辑回归[5]、主成分分析方法等单一统计学方法来进行研究[6]。由于机器学习算法在实际问题中有较好的预测效果，因此近几年有学者开始应用机器学习建立“高送转”预测模型。Wang 等[7]在 2016 年采用由 KNN、决策树以及 Logistic 算法组成的组合模型构建预测模型，该模型的准确率达到 70%。Dong 等[8]在 2018 年采用 BP 神经网络对“高送转”建模，其预测准确率接近 92%，证实了该模型比逻辑回归模型能更好地拟合“高送转”现象。Chen 等[9]在 2020 年选取过滤法进行特征筛选，采用 SVM 构建预测模型，模型准确率接近 86%。Li 等[10]在 2020 年利用机器学习筛选上市公司年报的多个因子，通过构建多元逻辑回归模型预测次年发行“高送转”股票的公司，最终模型准确率稳定在 80%。Zhang 等[11]集成了多种方法进行组合，从中选择最优的模型组合，其在对制造业上市公司的“高送转”预测上，准确率达到 84.96%。虽然上述方法在扩展到神经网络、集成学习等机器学习算法上，预测效果有显著的提升，但是很少有学者考虑到“高送转”分类问题属于不平衡问题，并且未采用合适的评价指标去衡量模型效果。同时大部分学者是选择之前学者总结的“高送转”影响因子建立预测模型，很少有学者基于数据挖掘技术挖掘重要因子。

本文构建的“高送转”模型主要包括以下几个步骤：

(1)结合前人的研究成果，对一些可能影响公司实施“高送转”的潜在因素进行特征创造。

(2)利用方差过滤＋互信息＋基于 lightGBM 模型的嵌入法组合的方式筛选出影响股票“高送转”的特征。

(3)利用差分进化算法(Differential Evolution，DE)与 lightGBM 模型相结合，构建 DE-lightGBM 模型。该模型主要分成两部分。1)不平衡处理：引入类别权重，从算法模型入手，对 lightGBM 模型的损失函数进行修正，将实施“高送转”的样本设置更大的权重；同时，引入正则项来控制模型的复杂度，以降低过拟合风险，最后利用差分进化算法寻优得到适合的权重与正则项系数。2)超参数优化：挑选合适的评价指标，利用差分进化算法搜索性能最优的参数组合。

2 相关算法介绍

2.1 特征选择算法

特征选择算法通过某种评价标准和搜索策略过滤数据中的冗余特征，以达到优化预测模型的目的。特征选择算法一般包括：过滤式(Filter)、包裹式(Wrapper)和嵌入式(Embedding) 3 类[12]。与包裹式算法相比，过滤式算法的复杂度较低，在大规模数据集上也能表现良好，具有较强的通用性。嵌入式算法直接在训练分类器的过程中自动进行特征选择，不仅能使所训练的分类器具有较高的准确率，还能减少计算量。因此，本文结合过滤式和嵌入式算法进行特征筛选。下文对这两种方法的主要思想进行简单介绍。

(1)方差过滤：方差过滤通过特征本身的方差来筛选特征的类。首先计算出每个特征的方差，再通过指定阈值，删除方差小于该阈值的特征。这个步骤通常是用作特征筛选的预处理，删除方差变化小的特征。

(2)互信息法：信息度量法具有无参、非线性的优势，且不需要预先知道样本数据的分布，在特征选择算法中得到广泛应用。本文利用最大信息系数法来度量特征与标签之间的信息量。最大信息系数(Maximal Information Coefficient，MIC)由 Rashef[13]于 2011 年首次提出，它利用互信息定义了两个变量之间的最大信息系数，并利用该系数衡量两个变量之间的相关性。计算互信息的具体步骤是：

给定某一自变量 $\boldsymbol{X}=\{x_1,x_2,x_3,\cdots,x_n\}$ 和因变量 $\boldsymbol{Y}=\{y_1,y_2,y_3,\cdots,y_n\}$，$n$ 为样本数量，则 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 两个变量之间的互信息定义为：

$$\boldsymbol{I}(\boldsymbol{X},\boldsymbol{Y})=\sum_{x\in\boldsymbol{X}}\sum_{y\in\boldsymbol{Y}}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

其中，$p(x,y)$代表 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的联合概率密度函数；$p(x)$和 $p(y)$分别为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的边缘概率密度函数。

将两种变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 组成一个按顺序排列的集合 $Q=\{x_i,y_i\}$，$i=1,2,3,\cdots,n$，定义一个将自变量 $\boldsymbol{X}$ 划分为 a 段，将因变量 $\boldsymbol{Y}$ 划分为 b 段的网格 S，通过统计落在网络中散点的频率，计算出 $p(x,y)$，$p(x)$和 $p(y)$，在网格内部计算得到互信息 $\boldsymbol{I}(\boldsymbol{X},\boldsymbol{Y})$。由于网格划分的方式不唯一，因此需要选取互信息值最大的网格($a\times b$)划分方式。划分网络 S 下 Q 的最大互信息值为：

$$\boldsymbol{MI}(Q,a,b)=\max\boldsymbol{I}(\boldsymbol{X},\boldsymbol{Y}) \tag{2}$$

将互信息值进行归一化组成特征矩阵 $\boldsymbol{M}(Q)_{a,b}$：

$$\boldsymbol{M}(Q)_{a,b}=\frac{\boldsymbol{MI}(Q,a,b)}{\log_2\min(a,b)} \tag{3}$$

最大的互信息系数的计算式为：

$$\boldsymbol{MIC}(Q)=\max_{ab<B(n)}\{\boldsymbol{M}(Q)_{a,b}\} \tag{4}$$

其中，$B(n)$为网格划分($a\times b$)的上限值，一般给出 $B(n)=n^{0.6}$。

(3)基于 lightGBM 模型的嵌入法：结合特征选择和 lightGBM 模型训练，在训练的过程中，可以得到每个特征的重要性程度，因此能用来去除那些相关性小的特征。

2.2 差分进化算法

差分进化算法是一种基于群体差异的随机搜索优化算法[14]，其基本思想是：从当前种群中提取出搜索步长以及方向信息，同时对种群进行交叉和变异以获取新的个体，然后在原个体和新个体之间进行选择，将更优的个体保存到下一代。其主要过程包括初始化、变异、交叉和选择等。

(1)初始化

在差分进化算法迭代过程中，初始种群应尽可能均匀分布并且覆盖整个搜索空间。设最优化问题决策变量的维数为 D，变量的上界和下界分别为 U_j 和 L_j，$j=1,2,\cdots,D$。令种群的规模为 NP，常用的种群初始化方法是均匀随机初始化，生成由 NP 个 D 维向量个体组成的种群。第 i 个个体是：

$$X_{i,j}=L_j+rand*(U_j-L_j) \tag{5}$$

其中，$j=1,2,\cdots,D$，$i=1,2,\cdots,NP$，$rand$ 表示区间[0,1]上

的均匀分布随机数。

(2)变异

差分进化算法通过变异策略实现个体变异，本文从种群中随机选取 4 个不同的个体生成差分向量对每代的最优个体进行变异。变异操作方法如下：

$$V_i(g+1)=X_{r_1}(g)+F(X_{r_2}(g)-X_{r_3}(g))+F(X_{r_4}(g)-X_{r_5}(g)) \tag{6}$$

其中，$V_i(g+1)$为经过变异后得到的变异个体；r_1，r_2，r_3，r_4，r_5 是随机选择且互不相等的整数，F 是缩放因子。

(3)交叉

交叉的目的是增加变异个体的多样性，交叉操作的方法如下：

$$U_{i,j}(g+1)\begin{cases}V_{i,j}(g+1), & rand(0,1)<CR\\ X_{i,j}(g), & \text{otherwise}\end{cases} \tag{7}$$

(4)选择

差分进化算法的选择算子有二进制选择算子和指数选择算子。其中二进制选择算子采用一对一的“贪婪”选择方式，在子代和父代之间选择最优的一个个体进入下一代。

2.3 lightGBM

lightGBM 算法是在 XGBoost[15] 基础上改进的一种集成算法。该算法被成功应用于处理海量数据，不用通过全部样本计算增益，节约了内存，缩短了运行时间。在 XGBoost 算法的基础上，lightGBM 算法主要的改进部分有：

(1)采用直方图算法。该算法不用提前对特征进行排序，也不需要单独储存特征预排序的结果，能够直接保存特征离散化后的值。该算法的思想是首先将全部样本在每个特征上的取值划分到某一段(bin)中，目的是把连续的浮点数特征离散化为 k 个整数，然后构造一个宽度为 k 的直方图，最后使用该直方图遍历数据，通过计算每个 bin 的梯度和样本数量等来寻找最优的分割点[16]。该过程在一定程度上减小了内存的消耗，其内存所占用的比例缩小到了 XGBoost 的 1/6。

(2)选取带有深度限制的按叶子生长的决策树生成策略(见图 1)，不再使用 XGBoost 的按层生长的策略(见图 2)，该策略每次从全部叶子中找到分裂增益最大的叶子进行分裂，能够区别对待每一层的叶子，减少了不必要的开销[17-18]。

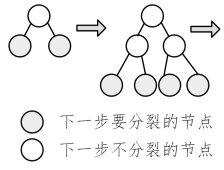

图 1　按层生长策略

Fig. 1　Layer by layer growth strategy

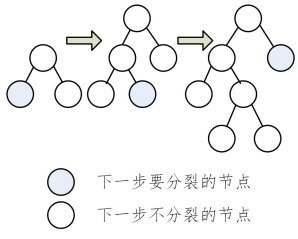

图 2　按叶生长策略

Fig. 2　Growth strategy by leaf

(3)采用 GOSS(单边梯度采样)算法选取样本。该算法先按降序排列全部分裂的特征，舍弃掉部分小梯度样本。由于直接舍弃了这部分样本，难免会造成信息损失。于是 lightGBM 保留下大梯度的样本，随机留下部分小梯度样本，并且乘上一个系数，达到扩大小梯度样本信息增益的目的，有效地防止改变数据分布。

3　DE-lightGBM 模型的构建

实际证券市场中，实施股票“高送转”方案的上市公司只占少数，其数据具有类别不平衡的特点，即未实施“高送转”的上市公司样本数量占多数，导致预测模型对实施“高送转”公司的查全率较低，且误报率较高。为增强机器学习模型对实施股票“高送转”的上市公司的预测能力，本文基于 lightGBM 模型，从不平衡处理和超参数优化两个方面进行改进。1)不平衡处理：在模型训练时，在损失函数中引入类别权重和 L1，L2 正则化项，修正模型损失函数，并通过差分进化算法估计，获得“高送转”样本权重和正则项参数的最优取值，从而加强对“高送转”样本的学习；2)超参数优化：在模型预测时，为得到更高预测精度的股票“高送转”预测模型，继续利用差分进化算法对 lightGBM 模型的主要参数进行优化，寻找到一组最优的参数组合。

(1)不平衡处理：损失函数反映了模型训练过程中模型的预测值与真实值之间的差异。损失函数的度量关乎模型的训练结果，模型训练的过程，即损失函数最小化的过程。对于“高送转”数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$共有 m 个样本，其中第 i 个样本 x_i 的类别为 y_i，$y_i=0$ 表示未实施“高送转”的股票样本，$y_i=1$ 表示实施“高送转”的股票样本，标准 lightGBM 模型的损失函数为：

$$\Omega(\varphi)=\sum_i l(\hat{y}_i,y_i) \tag{8}$$

其中，$\hat{y}_i$ 为单棵决策树对第 i 个样本的预测类别；$l(\hat{y}_i,y_i)$为单棵决策树的损失函数。

由于样本数量的差异，“高送转”样本所累积的损失小于未“高送转”样本所累积的损失，使未“高送转”样本所累积的损失在 lightGBM 模型的损失函数中占据主导地位，导致模型对“高送转”样本的学习不够充分，不能准确地预测出未“高送转”样本。这偏离了使用机器学习技术预测“高送转”样本的初衷。此外，还可通过正则化来控制模型复杂度，防止模型对收集到的训练集进行过度的学习，降低过拟合风险。因此，本文对 lightGBM 模型的损失函数进行了修正，引入类别权重，为“高送转”样本设置更大的权重，从而在损失函数中放大“高送转”样本的损失。同时引入 L1，L2 正则项控制模型复杂度，降低过拟合风险。修正后的损失函数为：

$$\Omega(\varphi)=\sum_i a_i l(\hat{y}_i,y_i)+\delta_1|\omega|_1+\delta_2|\omega|_2 \tag{9}$$

其中，$a_i=\begin{cases}1, & y_i=0\\ \gamma, & y_i=1\end{cases}$；$\delta_1|\omega|_1$ 和 $\delta_2|\omega|_2$ 分别是 L1 和 L2 正则项，ω 为该树的参数，可由决策树自动获得；δ_1 和 δ_2 分别为 L1 和 L2 正则化系数；a_i 为类别权重系数；γ 为少数类(“高送转”样本)的权重系数。

最后，对于少数类的权重系数 γ 和 L1 和 L2 正则化系数 δ_1 和 δ_2，通过差分进化算法进行估计。其中目标函数选择 5 折交叉验证的损失函数得分。

(2)其他超参数的优化

在确定了少数类的权重系数 γ 和 L1 和 L2 正则化系数 δ_1 和 δ_2 后(即完成不平衡处理之后)，需要进一步提升 light-

GBM 模型的精度，本文主要通过对 lightGBM 模型的一些主要参数进行超参数学习，从而得到一个具有良好预测能力的“高送转”模型。与估计权重系数 γ 和 L1 和 L2 的正则化系数 δ_1 和 δ_2 类似，选择合适的评价指标并继续使用差分进化算法对选定的主要参数进行优化，具体步骤与不平衡处理中相似，这里不再赘述。本文选定的 lightGBM 模型的主要参数如表 1 所列。

表 1 lightGBM 模型确定的主要参数

Table 1 Main parameters determined bylightGBM

参数	含义	作用
n_estimators	迭代次数	提高模型精度
max_depth	树的最大生长深度	控制训练时间、模型的性能
num_leaves	叶节点的数量	控制模型复杂性
min_data_in_leaf	叶子的最小样本数量	处理过拟合
feature_fraction	每次迭代使用的特征比例	加速训练、处理过拟合
bagging_fraction	每次迭代所用的数据比例	提高泛化能力、训练速度
learning_rate	学习率	提高模型精度

(3)DE-lightGBM 模型实验过程

DE-lightGBM 模型的流程图如图 3 所示。由图 3 可以看出，本文提出的基于差分进化算法的超参数优化的 DE-lightGBM 模型主要包含两部分：不平衡处理和模型超参数优化。不平衡处理部分通过引入正则项和损失函数权重，并利用差分进化算法获取最优参数。超参数优化部分根据本文分析的数据特征，选取 F1 和 AUC 作为评价目标，利用差分进化算法优化 lightGBM 模型的几个重要超参数，并对相应数据进行识别分类。

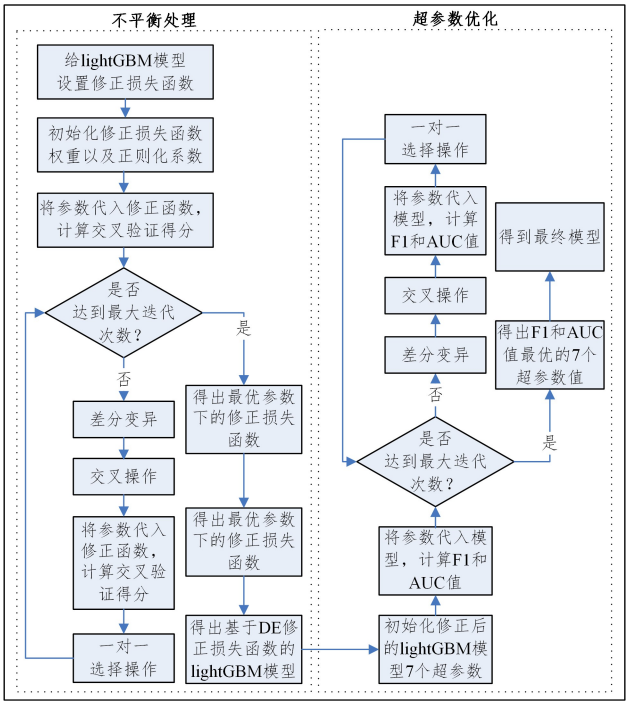

图 3 DE-lightGBM 模型的构建流程图

Fig. 3 De-lightGBM model construction flow chart

4 实践研究与分析

4.1 数据来源

本文数据来源于第八届“泰迪杯”全国数据挖掘挑战赛 A 题赛方提供的数据[19]。数据包含近 7 年 3466 家上市公司的基础数据、年数据、日数据。其中，基础数据包含 3466 家上市公司的编号、上市年限、所属概念板块、所属行业 4 个指标；年数据包含上市公司连续 7 年的财务指标数据及是否进行高送转等信息；日数据表包含上市公司开盘价、收盘价、换手率等市场的行情数据，以及一些财务类和情绪类的指标。另外，未“高送转”样本有 16377 个，“高送转”样本有 2709 个，属于不平衡数据集。

4.2 数据预处理

由于在日数据表和年数据表中存在很多缺失值、重复值以及无效值，而这些值的存在会使模型的预测不佳。因此，为了使建模数据更完整并且充分保留有用信息，在构造“高送转”模型前对原始数据进行预处理。主要处理方式如下：

(1)首先，剔除数据缺失值占比大于 40%的特征，保留缺失值占比 40%以下的特征。然后，为防止极端数据对数据产生影响，对缺失值占比 15%到 40%的特征数据进行中位数填补。最后，由于随机森林回归填补是利用多棵决策树填补缺失值，确保了预测值的随机性，能够表现数据的真实性，同时该方法自身的分类精度较高，因此保证了其填补值的准确度和可靠性[18]。故本文将占比 15%以下的特征数据用随机森林填补。

(2)由于上市公司的财务数据中很多特征的单位不同且数量级相差很大，导致特征之间不能进行比较，影响模型预测的准确度，因此在建模前必须对数据标准化。本文选择 Z 标准化，消除特征之间的量纲差异。

4.3 分类评价指标

由于上市公司只有两个选择，实施“高送转”和不实施“高送转”，属于二分类问题，因此本文的评价指标是针对二分类问题。首先介绍二分类混淆矩阵，如表 2 所列，TP 表示将“高送转”类别正确地预测为“高送转”类别；FP 表示将未“高送转”类别错误地预测为“高送转”类别；TN 表示将未“高送转”类别正确地预测为未“高送转”类别；FN 表示将“高送转”类别错误地预测为未“高送转”类别。

表 2 混淆矩阵

Table 2 Confusion matrix

		预测值	
		0(多数类)	1(少数类)
真实值	0	TN	FP
	1	FN	TP

基于上述矩阵，可得到以下评估指标值：

Recall(召回率)：

$$Recall=\frac{TP}{TP+FN} \tag{10}$$

Accuracy(准确率)：

$$Accuracy=\frac{TP+TN}{TP+FP+FN+TN} \tag{11}$$

Precision(查准率)：

$$Precision=\frac{TP}{TP+FP} \tag{12}$$

FPR(假正率)：

$$FPR=\frac{FP}{FP+TN} \tag{13}$$

F1：

$$F1=\frac{2\times precision\times Recall}{Precision+Recall} \tag{14}$$

AUC：在很多二分类实际问题中，还可以用 ROC 曲线和

AUC 值来衡量模型的优劣。ROC 曲线综合考虑了查准率和假正率两个值，该曲线以 FPR 为横坐标，以 Recall 为纵坐标。在得到的预测结果中，FPR 越小越好，Recall 越大越好，即 ROC 曲线越靠近左上角，分类效果越好。AUC 值是 ROC 曲线下的面积，该指标可以表示模型的整体分类效果以及模型的泛化能力，其值越大表示模型的整体分类效果越好。

本文要解决的问题是对两个类别进行预测分类，由于该问题具有不平衡性，需要选择合适的评估指标对模型的效果给予评估。之前的学者大多选取准确率作为“高送转”预测模型的指标，但是该指标并不能很好地解释不平衡问题。

为了解决“高送转”股票预测问题，本文选择 F1 值为第一个评价指标。由于投资者会根据预测的名单买入“高送转”股票，所以在选择评价指标时，一方面要保证“高送转”类别预测的精度，即召回率；另一方面由于很多公司发行股票“高送转”方案的目的不单纯，所以监管机构需要尽可能地将下一年会实施“高送转”的公司全部预测出来，即查准率。而 F1 能找到一个平衡点，它既考虑了真实的“高送转”样本被正确预测的概率，又考虑了预测未“高送转”样本中真实为“高送转”类的概率。该指标作为一个综合评价指标能较好地平衡召回率和查准率，F1 值越大，说明模型对“高送转”类别的识别性能越好。除此之外，本文选择 AUC 值为第二个评价指标。AUC 值可以反映模型的整体分类性能，值越高，模型的整体分类性能越好，泛化能力越强。

最终本文选择基于混淆矩阵的 F1 值和 AUC 值作为“高送转”模型的评估指标，综合考虑这 2 个指标评判模型的优劣。

4.4 特征创造与筛选

从以往学者的研究可知，“是否是次新股”“是否是国企”“是否是小盘成长股”都会影响上市公司实施“高送转”，因此将这 3 个特征单独从数据表中提取出来[19]。同时，由于很少有文献研究上市公司“所属行业”和“所属概念板块个数”对实施“高送转”的影响，因此本文考虑将“所属行业”这一特征的 18 个行业单独提取出来，转换为 18 个新的特征以及将“所属概念板块”转换为“所属概念板块个数”。再者，本文最终的研究目的是预测次年将会发行“高送转”股票的企业，所以创造了一个新的特征，即“下一年是否高送转”，该特征为本文所研究的因变量。

将数据经过前几节的处理后，并且将基础数据、日数据、年数据按股票编号、年份进行数据合并，最终还有 329 个特征。进一步通过方差过滤＋互信息＋基于 lightGBM 模型的嵌入法相结合的方法进行特征筛选，最终剩余 151 个特征。

4.5 特征创造与筛选

4.5.1 不平衡处理

本文通过第 2 节构建的 DE-lightGBM 模型进行不平衡处理，用差分进化算法估计出最优的少数类别的权重与正则项系数。实验中差分进化算法的参数设置为：种群规模 $NP=10$、最大迭代次数 $MAX\_num=300$、缩放因子 $F=0.5$、交叉概率 $CR=0.5$；目标函数设置为损失函数的 5 折交叉验证得分，旨在通过差分进化算法找到损失函数最小的参数组合。实验结果显示：通过差分进化算法调参后，少数类的权重系数 γ 和 L1 和 L2 正则化系数 δ_1 和 δ_2 分别为 19.746，0.389，0.933。图 4 为损失函数交叉验证得分的收敛图。

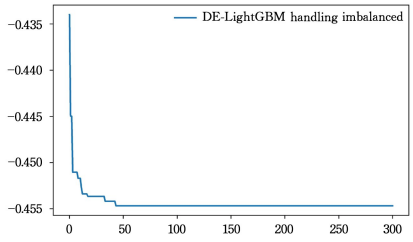

图 4 DE-LightGBM 下损失函数交叉验证得分收敛图

Fig. 4 Convergence diagram of cross validation score of loss function in DE-lightGBM

通常从数据和算法两个角度来处理不平衡数据，本文从算法角度，通过代价敏感学习[20]的方式进行不平衡处理。而从数据角度解决不平衡问题也被广泛应用，数据角度的采样方法通常有过采样、欠采样和综合采样[21]。

为了验证 DE-LightGBM 处理不平衡数据的有效性。本文将 DE-lightGBM 模型分别与随机过采样（ROS）、SMOTE 过采样、随机欠采样（RUS）、ENN 欠采样、SMOTEENN 综合采样这 5 种采样方法进行比较。其中 ROS 和 SMOTE 属于过采样，RUS 和 ENN 属于欠采样，SMOTEENN 属于综合采样。所有的不平衡处理方法均是建立在训练集的数据上，测试集的数据则使用原始数据。除此之外，本文还选取决策树模型[22]、随机森林模型[23]、XGBoost 模型和这些不平衡处理方法结合，再与 DE-lightGBM 模型进行多方面比较（各模型均只做不平衡处理，其余参数均选择默认参数）。各种方法所得的 F1 值和 AUC 值如表 3 所列。

表 3 DE-lightGBM 处理不平衡数据的有效性比较结果

Table 3 Comparison results of the effectiveness of DE-lightGBM in processing unbalanced data

算法	评价指标	未采样	ROS	RUS	SMOTE	ENN	SMOTEENN
DT	F1	0.0423	0.3554	0.3754	0.3053	0.3476	0.3955
	AUC	0.6076	0.7111	0.7452	0.6600	0.7460	0.7418
RF	F1	0.1541	0.4847	0.4253	0.4426	0.4377	0.4573
	AUC	0.8284	0.8447	0.8225	0.8101	0.8308	0.8253
XGBoost	F1	0.4054	0.4878	0.4530	0.4717	0.4740	0.4837
	AUC	0.8298	0.8516	0.8252	0.8274	0.8364	0.8365
lightGBM	F1	0.4116	0.4915	**0.4923**	0.4885	0.4746	0.4862
	AUC	0.8323	0.8526	**0.8598**	0.8387	0.8416	0.8390
DE-lightGBM	F1	**0.4973**	—	—	—	—	—
	AUC	**0.8636**	—	—	—	—	—

由表 3 可知，XGBoost 和 lightGBM 两个集成算法在各种采样方法下的整体效果优于决策树和随机森林算法。其中 RUS-lightGBM 模型在 24 个模型中效果最好，F1 和 AUC 值分别为 0.4923 和 0.8598。但这些模型的 F1 和 AUC 值均

小于 DE-lightGBM 模型的 F1 和 AUC 值。而本文提出的 DE-lightGBM 模型的 F1 和 AUC 值分别为 0.4973 和 0.8636，与 RUS-lightGBM 模型相比效果又有所提升。综上所述，无论是在模型选择还是不平衡数据处理的方式上，DE-lightGBM 模型均具有明显的优势。由此可知，本文基于算法模型角度引入损失函数权重和正则项及其参数优化能有效处理不平衡数据。

本文通过调整少数类别的权重来处理不平衡问题属于代价敏感学习，以往的研究中通常将少数类别的权重设置成多数类与少数类的比值。而本文提出的 DE-lightGBM 模型是通过差分进化算法来搜索最优的少数类别权重值。为了验证其有效性，将两组结果进行对比，结果显示：当将“高送转”样本的权重设置成未“高送转”样本与“高送转”样本的比值时，得出模型的 AUC 值为 0.8603，F1 值为 0.4843。从结果上看，DE-lightGBM 模型的效果较好，说明利用差分进化算法进行少数类别权重的设置是有效的。

4.5.2 超参数优化

为了进一步验证 DE-lightGBM 是一个精度更高的“高送转”预测模型，需要对 lightGBM 模型的主要参数进行超参数优化。实验中，差分进化算法的参数设置为：种群规模 $NP=10$、最大迭代次数 $MAX\_num=300$、缩放因子 $F=0.5$、交叉概率 $CR=0.5$、目标函数设置为 F1 和 AUC 的 5 折交叉验证得分，旨在通过差分进化算法找到 F1 和 AUC 最大的参数组合。lightGBM 模型主要参数的搜索范围和最终调参结果如表 4 所列。图 5 为 F1 和 AUC 的交叉验证得分收敛图。

表 4 DE-lightGBM 模型调参结果

Table 4 DE-lightGBM parameter optimization results

参数	搜索范围	参数最优值
n_estimators	(0,1000]	390
max_depth	(0,100]	86
num_leaves	(0,100]	64
min_data_in_leaf	(0,100]	62
feature_fraction	(0,1]	0.610
bagging_fraction	(0,1]	0.636
learning_rate	(0,1]	0.030

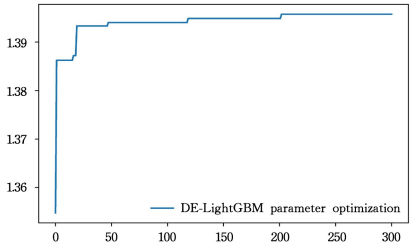

图 5 DE-lightGBM 下损失函数交叉验证得分收敛图

Fig. 5 Convergence diagram of cross validation score of loss function in DE-lightGBM

对模型参数进行优化的方法有许多种，其中较常用的有网格搜索、随机搜索、贝叶斯优化和 TPE 算法[24]。为了验证本文提出的 DE-lightGBM 模型中利用差分进化算法进行参数优化的有效性，将其与上述 4 种常用的调参技术进行对比。具体结果如表 5 所列。综合来看，DE-lightGBM 模型在对“高送转”的预测效果上均优于其他几种调参结果，$F1$ 值为 0.5368，AUC 值为 0.8734。

表 5 不同参数优化方法结果

Table 5 Results of different parameter optimization methods

算法	评价指标	值
网格搜索	F1	0.4923
	AUC	0.8598
随机搜索	F1	0.5231
	AUC	0.8672
贝叶斯优化	F1	0.5219
	AUC	0.8624
TPE 算法	F1	0.5290
	AUC	0.8734

通过上文中的研究结果可知，DE-lightGBM 模型在对“高送转”现象的识别上具有较好的分类效果。为了进一步挖掘影响公司发行“高送转”股票的因素，本文依据最优模型提取最重要的 20 个因子，因为在构造 DE-lightGBM 模型时，算法能按照特征对模型的贡献性进行排序，此排序能够充分说明所有特征的重要性。于是本文按照特征对模型的重要性排序的结果，提取影响最优模型的最关键的前 20 个特征，将这些特征作为影响公司“高送转”行为的关键因素。本文将靠前的 20 个特征按重要性由大到小排序，其结果如图 6 所示。

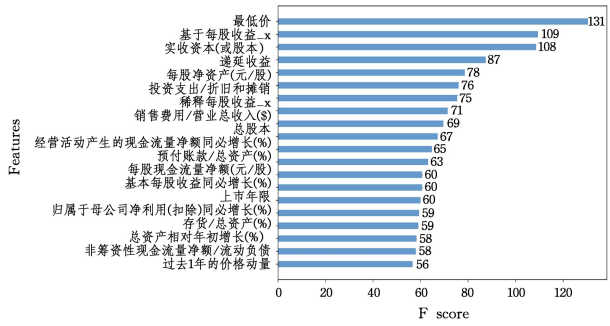

图 6 DE-lightGBM 模型特征重要性排序

Fig. 6 DE-lightGBM feature importance ranking

结束语 本文以 3466 家上市公司的相关年报数据和财务数据为例，首先对数据做缺失值、重复值、无效值处理等预处理操作；接着对特征进行转换和创造，在此基础上将方差过滤＋互信息＋基于 lightGBM 模型的嵌入法相结合进行特征筛选，以此减小模型的复杂度。针对“高送转”数据类别不平衡的特点以及追求构建更高精度的“高送转”预测模型，构建了一种 DE-lightGBM 模型。

首先，利用差分进化算法调整 lightGBM 模型损失函数中少数类别的权重以及正则项达到平衡数据的效果，并和 5 种常用的数据即采样方式随机过采样、随机欠采样、SMOTE 过采样、ENN 欠采样、SMOTEENN 综合采样与决策树模型、随机森林模型、XGBoost 模型和 lightGBM 模型相组合的 24 个模型做比较，结果显示 DE-lightGBM 模型具有明显的优势；其次，以 F1 和 AUC 作为评价指标，再次利用差分进化算法对 lightGBM 模型的主要参数进行调优，找到一组预测效果最优的参数组合，并与 4 种较常用的参数优化方法网格搜索、随机搜索、贝叶斯优化和 TPE 算法相比较。结果显示，DE-lightGBM 模型效果最优，F1 和 AUC 值分别为 0.5368 和 0.8734。最后通过 DE-lightGBM 模型中每个特征对模型的重要性排序，得到影响公司发行“高送转”股票的 20 个因素。其中排名最靠前的 5 个影响因素为最低价、基本每股收益、收盘价、每股实收资本(或股本)、递延收益。

由于中国股市尚处于发展阶段，历史数据相较于成熟的

欧美股市还有一定的差距。且行业间存在轮动规律,较不稳定。虽然本文也将行业作为影响的特征,但所得到的 F1 值仍有较大提升空间,后续会分行业对高送转影响因素进行挖掘,然后针对分行业所筛选出的高送转影响因子进行建模。

参 考 文 献

[1] CHE Z C,ZHAO Y X,GUAN S. Analysis on the Trend and Characteristics of "High Delivery" Policy of Listed Companies [J]. Friends of Accounting,2013,17:26-31.

[2] LIU Y,YE D L. High Transfer,Corporate Performance and Executive Reduction Scale[J]. Collected Essays on Finance and Economics,2019,9:62-72.

[3] LI C,HU Z Y,SHI S R. Research on Irrational Speculative Bubble Model Based on Stock Market Investor Sentiment [J]. The Theory and Practice of Finance and Economics,2018,39(5):51-57.

[4] KRIEGER K,PETERSON D R. Predicting Stock Splits with the Help of Firm-specific Experiences[J]. Journal of Economics and Finance,2009,33(4):410-421.

[5] XIONG Y M,CHEN X. Research on the Motivation of Turnover Behavior of Chinese Listed Companies——Based on the Test of High Turn-Over Samples[J]. Research on Economics and Management,2012,5:81-88.

[6] SHI H,TING X Y. Prediction Model of "High Delivery and Turn" Based on Pattern Recognition [J]. Times Finance,2016,12:289-290.

[7] WANG K,LONG W J. Research on High Stock Transfer Based on Integrated Learning [J]. Times Finance,2016,36:163-164,167.

[8] DONG K M,ZHAO S S. Research on the Motivation of "High Turnover" of Chinese Listed Companies--Based on BP Neural Network Model Method analysis [J]. Review of Investment Studies,2018,1:139-153.

[9] CHEN J W,CHEN Y X,FAN W H. Research on the Influence Factors of High Turnover of Listed Companies Based on Data Mining [J]. China Computer & Communication,2020,14:162-164.

[10] LI Y,FANG Z Q. Research on Enterprise High Transfer Method Based on Machine Learning [J]. Digital Space,2020,10:220-221.

[11] ZHANG T H,LUO K Y. An Empirical Study on High Turnover Forecasting of Listed Companies Based on Integrated Learning [J]. Computer Engineering and Applications,2021,57(4):1-7.

[12] YANG J,OLAFSSON S. Optimization-based Feature Selection with Adaptive Instance Sampling[J]. Computers & Operations Research,2006,33(11):3088-3106.

[13] RESHEF D N,RESHEF Y A,FINUCANE H K,et al. Detecting Novel Associations in Large Data Sets[J]. Science,2011,334(6062):1518.

[14] YANG Q W. Overview of Differential Evolution Algorithms [J]. Pattern Recognition and Artificial Intelligence,2008,4(21):506-513.

[15] SONG L L,WANG S H,YANG C,et al. Application Research of Improved XGBoost in Unbalanced Data Processing [J]. Computer Science,2020,47(6):98-103.

[16] YAN S X,ZHU P,LIU Z. Research on Vehicle Fault Prediction Method Based on Improved LightGBM Model [J]. Automotive Engineering,2020,42(6):815-819,825.

[17] TANG K,QIN M,ZHAO X,et al. Prediction of Gaseous Nitrite Based on Stacking Integrated Learning Model [J]. China Environmental Science,2020,40(2):582-590.

[18] Al DAOUD E. Comparison Between XGBoost,LightGBM and CatBoost Using a Home Credit Dataset[J]. International Journal of Computer and Information Engineering,2019,13(1):6-10.

[19] 第八届"泰迪杯"数据挖掘挑战赛赛题[EB/OL]. https://www.tipdm.org/ bdrace/ index.html.

[20] CHEN S L,SHEN S Q,LI D S. Integrated Learning Method for Unbalanced Data Based on Updating Sample Weights [J]. Computer Science,2018,45(7):31-37.

[21] KAUR H,PANNU H S,MALHI A K. A Systematic Review on Imbalanced Data Challenges in Machine Learning:Applications and Solutions[J]. ACM Computing Surveys(CSUR),2019,52(4):1-36.

[22] ZHOU Z H. Machine Learning[M]. Beijing:Tsinghua University Press,2016.

[23] LI G H,LI J Q,ZHANG L,et al. A Feature Selection Method Based on Ant Colony Algorithm and Random Forest [J]. Computer Science,2019,46(S2):212-215.

[24] BERGSTRA J,BARDENET R,BENGIO R,et al. Algorithms for hyper-parameter optimization[C]// Advances in Neural Information Processing Systems. 2011.

CEN Jian-ming, born in 1994, postgraduate. His main research interests include Intelligent algorithms and machine learning.

FENG Quan-xi, born in 1980, Ph.D, professor. His main research research interests include computational intelligence and machine learning in real world and so on.