文章编号：2095-2163(2022)07-0181-04 中图分类号：TP181 文献标志码：A

样本均衡与特征选择在员工离职倾向预测上的应用

吴学亮，娄 莉
（西安石油大学 计算机学院，西安 710000）

摘 要：本文采用 SMOTE、ADASYN、SMOTETOMEK、SMOTEENN 和 Borderline-SMOTE 5 种样本均衡方法，对数据进行了样本均衡。使用遗传算法对样本均衡前后的数据进行特征选择，并将 LightGBM 集成学习算法应用于员工数据，进行了离职倾向预测。结果表明，使用 SMOTEENN 方法对标准化后的数据进行样本均衡处理后，再使用遗传算法对其进行特征选择，LightGBM 预测效果最佳，验证了其优越性。

关键词：LightGBM；样本均衡；特征选择；员工离职；机器学习

Application of sample equilibrium and feature selection in predicting employee turnover intention

WU Xueliang, LOU Li
(School of Computer Science, Xi′an University of Petroleum, Xi′an 710000, China)

【Abstract】 This paper uses five sample equalization methods, SMOTE, ADASYN, SMOTETOMEK, SMOTEENN and Borderline-SMOTE, to perform sample equalization on the data. The data before and after sample balance are selected with genetic algorithm, and the LightGBM ensemble learning algorithm is applied to employee data to predict turnover intention. The results show that after using the SMOTEENN method to perform sample equalization processing on the standardized data and the genetic algorithm for feature selection, LightGBM has the best prediction effect, which verifies its superiority.

【Key words】 LightGBM; sample balance; feature selection; employee turnover; machine learning

0 引 言

近年来，随着经济社会的发展，员工流失问题是追求持续增长企业面临的重大挑战。这是一个在研究和实践中都受到广泛关注的问题。为了留住员工，并利用员工的知识促进公司的成长，人力资源部门利用机器学习算法预测员工是否有离职倾向解决此问题。

在现实生活中，数据普遍呈不平衡分布特征，其带来的问题也越加明显。随着分类问题研究的发展，越来越多的研究者开始研究不平衡数据集的极端不平衡分布特征，不平衡数据集的分类算法也越来越全面[1]。针对上述问题，本文对 SMOTE、SMOTETOMEK、ADASYN、SMOTEENN 和 Borderline-SMOTE 5 种样本均衡方法进行了研究与分析。

在应用机器学习的过程中，样本数据的特征通常差异很大，其中可能包含不相关的特征或存在紧密依赖的特征。综上所述，本文的贡献如下：

（1）提出了基于 LightGBM（Light Gradient Boosting Machine）的员工离职倾向预测模型，可根据给出的信息，评估员工是否有离职倾向并给出建议。

（2）实验过程中，对样本数据进行了详细的特征工程，包括：数据标准化、样本均衡和特征选择。

（3）利用 Data Castle 提供的数据集，评估了 LightGBM 方法。实验表明，使用样本均衡和特征选择后再使用 LightGBM 方法，优于直接使用 LightGBM 方法。

1 特征工程

1.1 数据标准化

数据采用不同的度量单位，可能导致不同的数据分析结果。通常，用较小度量单位表示的属性值，将导致该属性具有较大的值域，该属性往往具有较大的影响或“权重”。为了避免数据分析结果对度量单位选择的依赖性，需要对样本数据进行标准化或规范化，使之落入较小的共同区间（如：[0,1]或[-1,1]）。

对数据进行标准化不仅可以规避数据分析结果对度量单位选择的依赖性，有效提高结果精度；也可以简化计算，提升模型的训练和收敛速度。常用数

基金项目：陕西省重点研发计划项目（2021GY-138）；陕西省教育厅科研计划项目资助（21JK0847）。

作者简介：吴学亮（1983-），男，硕士研究生，主要研究方向：机器学习、深度学习；娄 莉（1970-）女，博士，副教授，主要研究方向：通信工程与图像处理。

收稿日期：2022-03-07

据标准化(Data Normalization,DN)方法有:最小-最大值标准化、z 分数标准化和小数定标标准化。

本文采用 z 分数标准化,经过处理后的数据符合标准正态分布,即均值为 0,标准差为 1。转化函数定义如式(1):

$$v_i^{'} = \frac{v_i - \bar{A}}{\sigma_A} \tag{1}$$

其中,A 表示数值属性,具有 n 个观察值为 v_1,$v_2,v_3,\cdots,v_n$;$\bar{A} = (v_1 + v_2 + \ldots + v_n)/n$ 为原始数据的均值;σ_A 为原始数据的标准差。

1.2 样本均衡

在现实生活中,为了更好地理解数据集类不平衡问题,本文从二分类问题的角度进行分析。设:br_χ、$\chi_{\min}$ χ_{maj} 分别表示样本的失衡率、少数类和多数类。一般情况下,如果关注的是少数类的样本数据且 $br_\chi \leqslant 0.2$(本文数据集 $br_\chi < 0.2$),就需要考虑对样本进行均衡处理,如式(2):

$$br_\chi = \frac{|\chi_{\min}|}{|\chi_{maj}|} \tag{2}$$

目前,已有多种方法用来克服类不平衡问题。其中,最常用的技术是采样方法[2],用于实现从数据集类的不平衡分布到平衡分布。采样方法可分为两种:欠采样和过采样技术。欠采样技术是指去除多数类中的少数数据点,而过采样方法是生成属于少数类的合成数据点,以获得所需的平衡比率。本文重点介绍过采样技术,主要包括:SMOTE、ADASYN、SMOTETOMEK、SMOTEENN、Borderline-SMOTE。

1.3 特征选择

特征选择可以消除不相关或冗余的特征,从而减少特征数量,提高模型的准确性,或减少运行时间[3]。此外,选择具有真实相关特征的简化模型,可以使研究人员更容易理解数据生成的过程。常见的特征选择方法可以分为 3 类:过滤、包装和嵌入方法[4-5]。本文在 LightGBM 算法的基础上,考虑特征的互补性,对特征进行选择和剔除。

对于包装方法,其主要组成部分是搜索策略和学习算法。包装模型中的搜索策略可以分为全搜索、启发式搜索和随机搜索。由于计算成本,完全搜索会耗尽所有可能的子集并找到最佳子集。与完全搜索不同,启发式搜索策略将会权衡搜索效率的最优性。顺序后向选择(Sequential backward selection,SBS)和顺序前向选择(sequential forward selection,SFS)是两种最常用的启发式搜索打包方法。但是,这两种方法都有一个单调的假设,即添加的特征不能被删除,并且被删除的特征不能再次添加,这使其易陷入局部最小值。随机搜索总是使用进化方法作为其众所周知的全局搜索能力。与确定性算法相比,进化搜索方法不仅能有效捕捉特征冗余和交互作用,而且不受单调假设条件的限制。进化搜索方法,可以避免陷入局部最优,并且可以找到小部分特征。然而,基于随机搜索的打包方法存在计算量大的缺点[6]。

遗传算法(Genetic Algorithm,GA)[7]是受自然进化过程启发而开发的一种启发式优化技术,其种群的成员以基因序列的染色体形式表示。在特征选择问题中,每个基因用 0 或 1 来表示,对应问题空间的一个属性或参数。本文选择基于 LightGBM 算法进行员工离职倾向预测,其结果的准确率作为适应度函数评估指标。遗传算法的基本思想是适者生存理论。每个新种群生成的算法,可通过选择、交叉和变异等 3 个主要步骤达到更高的适应度水平。

2 LightGBM 算法

2.1 算法原理[8]

LightGBM 是在传统的梯度提升树(GBDT)上使用直方图算法(histogram-based algorithm),在一个待分裂的结点上,为每一个特征构建直方图。具体实现过程是:先对特征值进行分箱处理,然后根据分箱值构造一个直方图;遍历结点中的每一个样本,在直方图中累积每个 *bin* 的样本数和样本梯度之和;当一次数据遍历完成后,直方图就累积了需要的统计量。

对于每个特征,根据构建的直方图,遍历每一个 *bin* 值从而寻找最优分裂特征及 *bin* 值。同时使用带深度限制的 Leaf-wise 叶子生长策略,经过一次数据可以同时分裂同一层的叶子,具有易进行多线程优化、易控制模型复杂度、不易过拟合的特点。

2.2 算法优势

为了更准确的残值建模和预测,LightGBM 算法在基于直方图的 GBDT 算法中引入了基于梯度的单边采样(Gradient-based One-Side Sampling,GOSS)和独占功能捆绑(Exclusive Feature Bundling,EFB)两种技术[9]。其中,GOSS 方法可在小样本情况下实现高精度预测,可减少计算成本,性能优于随机抽样方法且不会损失太多的训练精度。而 EFB 可将互斥的特征捆绑在一起解决高维特征的降维问题。

在 GBDT 算法中，信息增益由方差增益计算获得。而 LightGBM 算法采用的是 GOSS 算法，根据训练实例的梯度绝对值降序，对训练实例进行排序，并且生成 3 个特征子集：A、A^c 和 B。其中，特征子集 A 由前 $a \times 100\%$ 的实例与较大的梯度得到，特征子集 A^c 由 $(1-a)\times 100\%$ 组成的实例与较小的梯度得到；特征子集 B 是进一步随机采样 $b \times |A^c|$ 得到。估计方差增益 $V_j(d)$ 定义如式(3)：

$$V_j(d)=\frac{1}{n}\left(\frac{\left(\sum_{x_i\in A_l} g_i+\frac{1-a}{b}\sum_{x_i\in B_l} g_i\right)^2}{n_l^j(d)}+\frac{\left(\sum_{x_i\in A_r} g_i+\frac{1-a}{b}\sum_{x_i\in B_r} g_i\right)^2}{n_r^j(d)}\right)$$

$$A_l=\{x_i\in A: x_{ij}\leqslant d\},A_r=\{x_i\in A: x_{ij}>d\}$$

$$B_l=\{x_i\in B: x_{ij}\leqslant d\},B_r=\{x_i\in B: x_{ij}>d\} \quad (3)$$

3　实验结果与分析

为了验证 5 种样本均衡方法和遗传算法对数据进行处理的有效性，在配置为 Intel Corei7、SSD128 G、HDD 1TB、RAM 24 GB、Windows 操作系统的环境中进行了相关实验。实现代码工具利用 Conda 4.11.0 完成；GA 种群规模为 100，迭代次数是 50，交叉率是 0.5，变异率是 0.4；LightGBM 算法参数为默认值。本文实验使用 scikit-learn 版本为 0.24.1、LightGBM 版本为 3.3.0。

3.1　数据集描述

本文数据取自 Data Castle 平台发布的数据集，从中选取 1 100 条数据用于实验。其中，在职记录 922 条，离职记录 178 条。样本的失衡率即离职率为：0.161 8。原始数据中有 31 个条件属性，1 个决策属性。通过业务选择过滤了 3 个条件属性，利用已有的条件属性构造出了 6 个新的条件属性。

3.2　评价指标

本实验采用准确率、精确率、召回率和 F_1 值作为评价指标。准确率(*Accuracy*) 是指对于给定测试数据集，分类器正确分类的样本数与总样本数之比；精确率(*Precision*) 是预测的正例结果中，确实是正例的比例；召回率(*Recall*) 是所有正例的样本中，被找出的比例；*F*1 值是综合评价指标，*F*1 值越接近 1，表明模型预测越准确。准确率、精确率、召回率和 *F*1 值是由混淆矩阵计算得到。分类结果混淆矩阵见表2。准确率、精度率、召回率和 F_1 值的计算方法如公式(4) ~ 公式(7) 所示。

表 1　分类结果混淆矩阵

Tab. 1　Confusion matrix of classification results

真实结果	离职(预测结果正例)	未离职(预测结果反例)
离职(正例)	*TP*	*FN*
未离职(反例)	*FP*	*TN*

$$Accuracy=(TP+TN)/(TP+FN+FP+TN) \quad (4)$$

$$Precision=TP/(TP+FP) \quad (5)$$

$$Recall=TP/(TP+FN) \quad (6)$$

$$F_1=(2\times Precision\times Recall)/(Precision+Recall) \quad (7)$$

3.3　模型评估

为了达到验证的目的，在验证数据集时使用了分层 $k(k=10)$ 折交叉验证。每个数据集被随机分成 k 折，其中 $k-1$ 折为训练集，剩余的为测试集。分层 k 折交叉验证是评估建模结果最有效和广泛使用的验证和能力评估技术之一。通过分层 k 折交叉验证获得了不同样本均衡算法和是否使用遗传算法进行特征选择的最佳评价指标。实验结果见表 2 与图 1 所示。

由表 2 可知，样本处理方法为“SMOTEENN + GA”时，效果最好，其准确率达到 95.82%、精确率达到 97.42%、召回率达到 96.28%、F_1 值达到 96.66%。实践证明，采用样本均衡和遗传算法的特征选择，可以有效提高模型的性能。

表 2　样本采用不同处理方法性能对比结果

Tab. 2　The performance comparison of different processing methods　%

样本处理方法	准确率	精确率	召回率	F_1 值
DN	86.09	64.28	35.49	44.77
DN+GA	88.27	81.10	38.30	50.30
DN+SMOTE	92.10	94.77	89.02	90.07
DN+SMOTE+GA	93.40	95.99	90.65	92.04
DN+ADASYN	92.09	94.79	88.30	89.76
DN+ADASYN+GA	93.16	96.25	89.47	91.25
DN+Borderline-SMOTE	92.69	95.48	89.67	90.96
DN+Borderline-SMOTE + GA	93.51	96.38	90.43	92.03
DN+SMOTETOMEK	92.10	94.77	89.02	90.07
DN+SMOTETOMEK + GA	93.56	95.79	91.09	92.33
DN+SMOTEENN	94.45	96.73	94.86	95.46
DN+SMOTEENN+GA	**95.82**	**97.42**	**96.28**	**96.66**

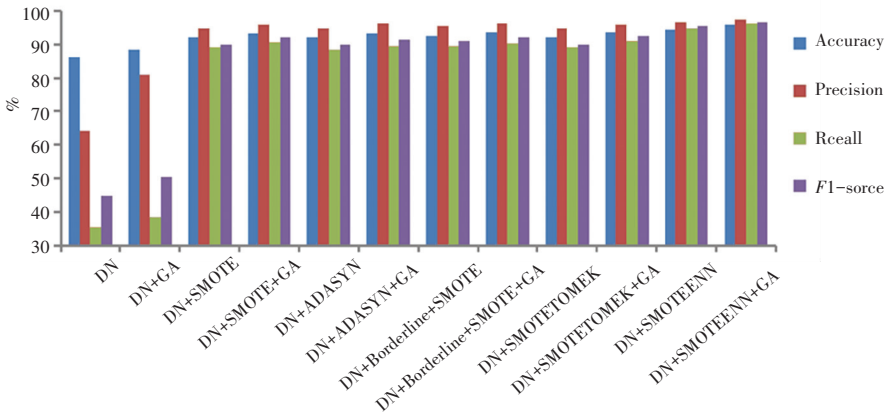

图 1 实验运行结果对比

Fig. 1 Comparison of experimental results

4 结束语

本文描述了研究预测员工离职的必要性,并在构建模型时使用了样本平衡、特征选择和机器学习算法,强调样本均衡和特征选择算法的重要性。模型选用 SMOTEENN、遗传算法和 LightGBM 的组合,与单独的 LightGBM 分类器给出的结果相比,该模型提供了更优越的性能。

参考文献

[1] WANG L. Review of Classification Methods on Unbalanced Data Sets [J]. IEEE Access, 2021, 9: 64606-64628.

[2] BATISTA G, PRATI R, MONARD M. A study of the behavior of several methods for balancing machine learning training data [J]. SIGKDD explorations, 2004,6(1): 20-29.

[3] LI C. A new feature selection algorithm based on relevance, redundancy and complementarity [J]. Comput Biol Med, 2020, 119: 103667.

[4] ZENG Z. A novel feature selection method considering feature interaction [J]. Pattern Recognition, 2015, 48(8): 2656-2666.

[5] ZHOU Y, KANG J, GUO H. Many-objective optimization of feature selection based on two-level particle cooperation [J]. Information Sciences, 2020, 532:91-109.

[6] XUE X, YAO M, WU Z. A novel ensemble-based wrapper method for feature selection using extreme learning machine and genetic algorithm [J]. Knowledge and Information Systems, 2018, 57(2): 389-412.

[7] 尚荣华, 焦李成,刘芳,等. 计算智能导论[M]. 西安: 西安电子科学大学出版社,2019:37-38.

[8] 宋海龙,黎明,苟江,等. 基于 LightGBM 的航空发动机剩余使用寿命预测[J]. 现代计算机,2021,27(35):47-52.

[9] KE G, MENG Q, FINLEY T, et al. Lightgbm:A Highly Efficient Gradient Boosting Decision Tree [C]// Advances in Neural Information Processing Systems, 2017: 3146-3154.

(上接第 180 页)

实体信息的代理运行与游戏信息的输出可供智能模型利用,为游戏 AI 的研究提供一定的帮助。

4 结束语

即时战略类游戏历来广受玩家喜爱,本文开发的海盗对战游戏在即时战略类游戏的基础上,具有真实性好,可拓展性高的特色,经过测试,该游戏运行流畅,体验良好,未来也有不错的提升空间。

参考文献

[1] 王蔚,史建婷. 电子游戏的分类与发展[J]. 北京视察, 2003(4):47.

[2] 刘宝谦. 一款即时战略游戏的关键技术研究与实现[D]. 成都:电子科技大学, 2011.

[3] 范腾,邵坤,唐振韬,等. 基于无监督学习的星际争霸 2 宏观决策[C]//2018 中国自动化大会(CAC2018),2018:361-366.

[4] 樊东东. 基于人工智能的星际争霸Ⅱ智能体研究与实现[D]. 四川:西南交通大学, 2019.

[5] 赵红亮. 探析游戏动画中的场景设计[J]. 吉林动画学院, 2020(1):108.