文章编号：2095-2163(2024)01-0147-05 中图分类号：TP399 文献标志码：A

基于 SMOTE_GA_XGBoost 的葡萄酒质量预测

丁海萌，郭小燕
（甘肃农业大学 理学院，兰州 730070）

摘 要：随着经济发展和消费升级，人们对高品质葡萄酒的需求不断增加，如何利用葡萄酒理化指标进行高效准确的质量评定显得尤为重要。本文基于 UCI 葡萄酒数据集，建立了 SMOTE_GA_XGBoost 模型来预测葡萄酒质量。结果表明，SMOTE_GA_XGBoost 模型得出的级别判别准确率为 89.36%，类别判别准确率为 96.46%，均高于其他对比模型，具有更高的预测精度。

关键词：葡萄酒质量预测；机器学习；SMOTE；GA_XGBoost

Wine quality prediction based on SMOTE_GA_XGBoost

DING Haimeng, GUO Xiaoyan
(College of Science, Gansu Agricultural University, Lanzhou 730070, China)

Abstract: With the economic development and consumption upgrading, the demand for high-quality wines is increasing, and how to utilize wine physicochemical indicators for efficient and accurate quality assessment is particularly important. In this paper, based on the UCI wine dataset, the SMOTE_GA_XGBoost model was established to predict wine quality. The results show that the SMOTE_GA_XGBoost model yields a level discrimination accuracy of 89.36% and a category discrimination accuracy of 96.46%, both of which are higher than those of other comparative models and have higher prediction accuracy.

Key words: wine quality prediction; machine learning; SMOTE; GA_XGBoost

0 引 言

随着当今经济社会的发展，人们的生活水平不断提高，消费能力也逐渐升级。与此同时，越来越多的人开始了解葡萄酒，葡萄酒消费在国内得到了广泛的普及，成为大众饮品消费中重要的一环，中国的葡萄酒行业在生产、进口和消费方面也有大幅增长。行业的发展离不开对产品质量的把控，保证生产质量，满足人们对产品质量、形色、口感的要求，才能促进消费，形成产业良性循环。

目前，对于葡萄酒质量的评定，国际上尚未制定一个统一标准，生产企业对葡萄酒质量评估把控较为松散[1]。通常情况下，对葡萄酒质量的评定是通过专业品酒师来进行评分得到分数，但由于能力水平和个人喜好难免会产生一定差异[2]，况且在实际生产过程中，大批量的葡萄酒显然无法采取这种评定方式。葡萄酒的品质好坏与其中的化学成分有关，因此除去人工品尝的方式，葡萄酒的理化数据指标也是衡量品质的重要标准。测定葡萄酒的各项理化指标，寻找理化指标和质量分数之间存在的某种关系，建立一个有效的葡萄酒质量预测模型，能够完成对葡萄酒质量的评价，且更为高效、客观、准确。

针对葡萄酒的质量评价，程再冉等人[3]运用主成分分析，对酿酒葡萄进行分级，并运用多元回归验证评价可行性。夏铭泽等[4]利用支持向量机模型，通过葡萄酒的理化指标来进行葡萄酒质量预测。赵峰等[5]对神经网络的学习率进行优化，建立基于自适应遗传算法优化的 BP 神经网络预测模型，用于葡萄酒质量预测。于尚琨[6]采用新型主动学习多分类 SVM 算法，选取代表数据集特征的少量样本作为训练集，提高分类的准确率。

基金项目：甘肃农业大学盛彤笙创新基金(GSAU-STS-2021-16)；甘肃农业大学青年导师基金(GAU-QDFC-2021-18)。

作者简介：丁海萌(1998-)，女，硕士研究生，主要研究方向：机器学习与数据挖掘。

通讯作者：郭小燕(1976-)，女，博士，副教授，主要研究方向：智能信息处理与农业信息化研究。Email：guoxy@ gsau.edu.cn

收稿日期：2023-02-03

与传统的多元回归、支持向量机、BP 神经网络相比,XGBoost 模型对于大量数据具有很好的处理速度和精度,鲁棒性强。因此,本文采用 XGBoost 模型作为主体,使用 SMOTE 扩充样本生成均衡数据集,针对 XGBoost 参数众多、调节不便的问题,利用遗传算法自适应全局优化搜索的优点来寻找最优参数组合,建立改进的 SMOTE_GA_XGBoost 模型。通过与 GA_XGBoost、SMOTE_XGBoost 以及单独的 XGBoost、随机森林、SVM 模型的预测结果对比,验证了本文提出的模型的预测精度。

1 基本理论

1.1 SMOTE 算法

针对数据集中样本分布不均衡的问题,本文采用 SMOTE 算法对少数类样本进行过采样,生成更多样本,均衡每种质量类别的样本数量,以提升模型对于不均衡数据集中每个类别的判别效果[7]。该算法在随机过采样的基础上,利用最近邻样本加入随机噪声,使生成样本与真实样本拥有相似的分布,以此可以防止采样样本过度重复,避免过拟合,增强模型的泛化能力。SMOTE 的采样过程如图 1 所示。

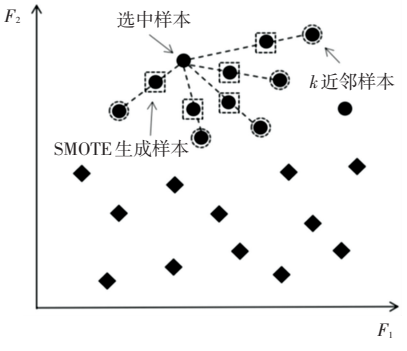


图 1 SMOTE 采样过程图

Fig. 1 Diagram of the SMOTE sampling process

SMOTE 采样的具体步骤如下:

(1)对于少数类中的每个样本 x_i,以欧氏距离作为指标,计算从 x_i 到同一类样本集中的所有样本的距离,得到相应的 k 个最近邻。

(2)根据数据集的不平衡比,设置采样比 M 来确定不同样本的新比例。对于 x_i,从其 k 最近邻中随机选择几个样本,假设所选择的最近邻为 x_m,对于每一个选中的 x_m,在式(1)的基础上从原始样本中生成一个新的样本。

$$x_{new} = x_i + rand(0,1) * | x_j - x_i | \quad (1)$$

其中,$j = (1,2,\cdots,M)$,$rand(0,1)$ 表示[0,1]区间的随机数。

(3)重复上述步骤进行 M 次采样,生成 M 个新样本。将新样本与少数原始样本合并后,生成一个新的平衡数据集。

1.2 遗传算法(GA)

遗传算法是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型[8],是一种通过模拟自然进化过程自适应地搜索最优解的方法。遗传算法的实现过程如下:

(1)种群初始化:种群数量对遗传算法优化性能有较大影响,种群越大优化时间越长,种群过小则不易找到最优解。

(2)评估种群中个体的适应度:适应度函数值增大的方向就是进化方向,个体的适应度越高,表示该个体越优,被选中作为父代的概率也就越大。XGBoost 模型预测得出的准确率越高代表期望个体适应度越高,预测效果也就越好。

(3)选择父母:依据轮盘赌选择法,适应度高的个体被选中的概率高,最优个体直接遗传复制到下一代。

(4)交叉:从群体中随机选取两个个体,再生成一个随机数,若小于交叉概率则进行交叉,选取一个或多个交叉点进行交换部分基因以生成新的子代个体。

(5)变异:随机选择一个个体,再生成一个随机数,若小于变异概率则进行变异,随机地改变个体某个或某些基因数,进而产生新的个体。

(6)由交叉和变异产生新一代种群,返回步骤(2)迭代寻优,不断进行循环操作得到新的种群,直到所需要求适应度值的个体或者运行到最大迭代次数时循环结束,产生最优个体。

1.3 XGBoost 模型

极致梯度提升树[9](eXtreme Gradient Boosting,XGBoost)是一种通过 Boosting 思想将多个基学习器集成构建强学习器的算法。XGBoost 有优良的学习效果以及高效的训练速度,广泛应用于数据挖掘等各个领域。该算法是 GBDT 算法的改进,传统的 GBDT 方法只利用了一阶导数,XGBoost 则是对损失函数做二阶泰勒展开,并在目标函数之外加入了正则项抑制模型的复杂程度,避免过拟合。树的集成模型如式(2)所示:

$$\hat{y}_i = \sum_{t=1}^{k} f_t(x_i) \quad (2)$$

其中,f_k 为第 k 个基模型,$\hat{y}_i$ 为第 i 个样本的预测值。

目标函数由模型的损失函数 L 与抑制模型复杂度的正则项 Ω 组成，因此有：

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{k} \Omega(f_t) \tag{3}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

式中：$\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ 为真实值与模型预测值的误差，式(4)为用于控制模型复杂度的正则项，其中 T 为叶子结点的数量，w 为叶子结点的取值，γ、λ 分别用来控制叶子结点的数量和取值，通过正则化可避免模型出现过拟合。

由于 boosting 模型是前向加法，目标函数可表示为

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \sum_{t=1}^{t} \Omega(f_i) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \sum_{t=1}^{t} \Omega(f_i) \tag{5}$$

找到最小化目标函数，相当于求解 $f_t(x_i)$，对于函数 $l(y_i, \hat{y}_i^{t-1} + f_t(x_i))$，在 $\hat{y}_i^{t-1}$ 处进行泰勒二阶展开，$f_t(x_i)$ 视为 Δx，则目标函数近似为

$$Obj^{(t)} \approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \sum_{t=1}^{t} \Omega(f_i) \tag{6}$$

由于 $l(y_i, \hat{y}_i^{t-1})$ 是一个常数，对目标变函数的优化没有影响，因此目标函数可以写为：

$$\begin{aligned} Obj^{(t)} &\approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{t=1}^{t} \Omega(f_i) = \\ &\sum_{i=1}^{n} \left[g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 = \\ &\sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T = \\ &\sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \end{aligned} \tag{7}$$

其中，$G_j = \sum_{i \in I_j} g_i$，$H_j = \sum_{i \in I_j} h_i$。

从上式中可以看出，目标函数被转化为一元二次函数，对其求解得到最优 w 及化简后的目标函数，如式(8)、式(9)所示：

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{8}$$

$$Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{9}$$

1.4　SMOTE_GA_XGBoost 模型

基于机器学习的葡萄酒质量预测模型，需要全面且充足的数据，然而葡萄酒质量数据集中高评分和低评分的样本较少，而这正是需要重点判别的部分。使用样本不均衡的数据集训练模型，会影响模型精度，得不到较为精确的预测结果。XGBoost 模型参数的值对模型的预测结果影响较大[10]，且参数众多，调节不便，需要一种可以自适应寻找最优参数的算法进行调参优化。基于此，本文提出 SMOTE_GA_XGBoost 模型，扩充训练样本，以质量预测准确率作为适应度值不断迭代，得到 XGBoost 最优参数组及预测结果。模型流程如图 2 所示。

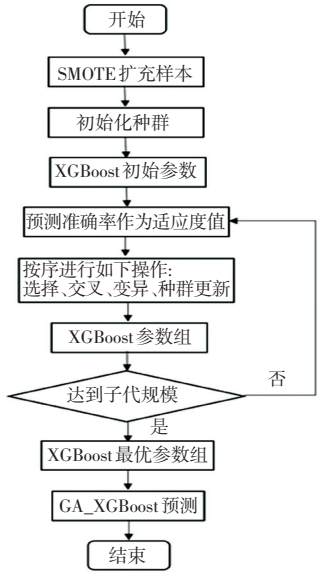


图 2　SMOTE_GA_XGBoost 流程图

Fig. 2　SMOTE_GA_XGBoost flow chart

实现步骤如下：

(1)使用 SMOTE 方法扩充样本；

(2)初始化种群，确定 XGBoost 模型初始参数；

(3)计算适应度值，依据轮盘赌选择法选择父母；

(4)通过交叉和变异产生新的个体并计算其适应度值；

(5)得到 XGBoost 模型参数组及当代最优个体；

(6)判断是否达到子代规模，是则退出并输出最优结果，否则返回(3)；

(7)得到 XGBoost 最优参数组及其预测结果。

2　数据分析与预处理

本文使用的葡萄酒数据来源于 UCI 机器学习库，选取红葡萄酒样本数据共计 1 599 组，涉及的变

量有 12 个,主要理化指标分别为非挥发性酸、挥发性酸、柠檬酸、残糖、氯化物、游离二氧化硫、二氧化硫总量、密度、pH 值、硫酸盐、酒精度共十一个,最后一项葡萄酒质量级别用 0~10 之间的分数表示。各变量概况见表 1。

表 1 变量概况表

Table 1 Variable overview table

变量(单位)	取值范围	均值	变量说明
非挥发性酸(g/dm^3)	4.6~15.9	8.32	有机酸,包含酒石酸、苹果酸或乳酸
挥发性酸(g/dm^3)	0.12~1.58	0.53	主要为醋酸
柠檬酸(g/dm^3)	0~1	0.27	葡萄本身带有的酸,含量较少
残糖(g/dm^3)	0.9~15.5	2.54	葡萄糖与果酸的混合物
氯化物(g/dm^3)	0.01~0.61	0.09	与葡萄酒咸度相关
游离二氧化硫(mg/dm^3)	1~72	15.87	抑菌防腐,质子状态的酸或酸式盐
总二氧化硫(mg/dm^3)	6~289	46.47	游离二氧化硫与结合二氧化硫之和
密度(g/cm^3)	0.99~1.004	0.997	单位体积内葡萄酒质量的多少
PH 值	2.74~4.01	3.31	酸碱性强弱程度
硫酸盐(g/dm^3)	0.33~2	0.66	葡萄酒发酵过程中产生的天然物质
酒精含量/%	8.4~14.9	10.42	葡萄酒中酒精所占百分比
质量级别	3~8	5.64	与专业品酒师鉴定结果对应

为减少数据不同量纲的影响,对其进行归一化处理,式(10):

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{10}$$

其中,x^* 为归一化后的值;$\max(x)$ 为数据最大值;$\min(x)$ 为数据最小值。

在模型判别过程中,可能存在将样本误判至相邻级别的情况,例如将级别 5 的样本误判至级别 6 中,可以认为该判别结果比较理想。因此,基于葡萄酒质量级别频率分布和 K-Means 聚类结果如图 3 和表 2 所示,除了级别判别准确率之外,增加类别判别准确率指标,将 3~4 分的酒认为是质量较差的酒,记为Ⅲ类,将 5~6 分的酒认为是质量合格的酒,记为Ⅱ类,将 7~8 分的酒认为是质量较好的酒,记为Ⅰ类。

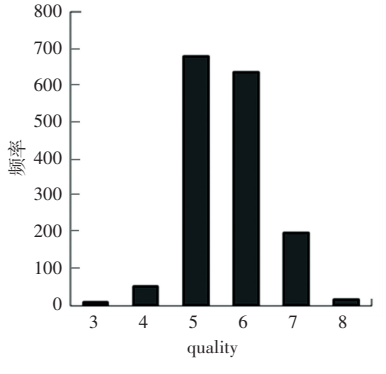


图 3 葡萄酒质量级别分布

Fig. 3 Wine quality grade distribution

表 2 K-Means 聚类结果

Table 2 K-Means clustering results

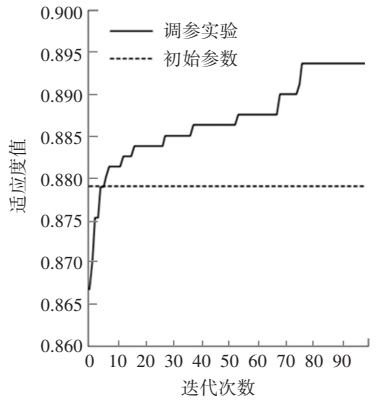
类别	3	2	1
聚类中心	3.84	5.48	7.08

3 实验结果分析

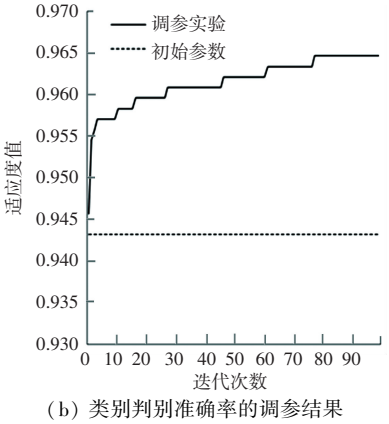
3.1 参数优化

本文利用遗传算法对 XGBoost 的 4 项主要参数进行优化,其中包括基树棵数(tree_num);学习率(learning_rate),即每次迭代更新权重时的步长;最大树深(max_depth);最小叶子权重(min_child_weight)。设置初始种群数量为 50,每个个体包含 4 个参数,参数在可选范围内随机生成,共迭代 100 次,分别以级别判别准确率和类别判别准确率作为适应度函数。每次迭代记录下,适应度值最高的最优个体,再进行选择、交叉、变异,繁殖后代获得更好的基因,且避免陷入局部最优解。两种适应度函数的 GA_XGBoost 调参实验过程如图 4 所示。

可以看出,随着迭代次数的增加,适应度函数不断增大,说明遗传算法对模型起到了优化作用。通过 GA_XGBoost 调参实验,在迭代 100 次后,判别准确率分别达到 89.36% 和 96.46%,相较初始参数的 87.90% 和 94.32%,GA_XGBoost 调参后的判别准确率均有明显提升。当前 XGBoost 模型的两组最优参数组合见表 3。



（a）级别判别准确率的调参结果



（b）类别判别准确率的调参结果

图 4　调参实验

Fig. 4　Tuning experiment

表 3　XGBoost 最优参数及初始值

Table 3　XGBoost optimum parameter and initial value

参数	初始值	最优值 1	最优值 2
tree_num	100	150	160
learning_rate	0.3	0.09	0.14
max_depth	6	8	10
min_child_weight	3	1	6

3.2　实验评估

本文选取级别判别准确率（*AUC*1）、类别判别准确率（*AUC*2）作为评价指标，对比不同模型的性能，比对结果见表 4。

表 4　不同模型 AUC1、AUC2 对比

Table 4　Comparison of AUC1 and AUC2 of different models %

	模型	*AUC*1	*AUC*2
组合模型	SMOTE_GA_XGBoost	89.36	96.46
	GA_XGBoost	71.25	88.75
	SMOTE_XGBoost	87.90	94.32
单一模型	XGBoost	67.50	86.25
	SVM	54.47	85.31
	RF	68.44	85.94

由上表预测结果准确率可知，采用改进的 SMOTE_GA_XGBoost 模型对葡萄酒质量级别进行预测，级别判别准确率为 89.36%，类别判别准确率为 96.46%，评价指标均高于其它对比模型。与其它模型相比，改进模型的拟合程度较高，降低了模型的预测误差，预测结果较好。因此，使用此模型来预测葡萄酒质量是可行且有效的。

4　结束语

在目前各类研究中，XGBoost 模型已被广泛应用于各类预测工作，然而单一模型的预测精度受到一定程度的限制。本文使用 UCI 葡萄酒数据集，针对样本不平衡、参数众多、调节繁琐等问题，利用 SMOTE 扩充样本和遗传算法全局优化，提出 SMOTE_GA_XGBoost 模型，用来进行葡萄酒质量预测。将 SMOTE_GA_XGBoost 模型与其他模型实验对比，其结果表明，SMOTE_GA_XGBoost 模型比单一模型预测精度有所提高，通过改进样本和模型的缺陷，进而缩小葡萄酒质量预测的误差。本文建立的 SMOTE_GA_XGBoost 模型为葡萄酒的质量评定提供了一种参考方式，有一定的借鉴意义。

参考文献

[1] 李伟康. 基于 GA-BP 神经网络对葡萄酒质量评估的研究[D]. 北京：北京工业大学，2018.

[2] 张志然，王恩辉，李兴元，等. 葡萄酒感官质量评价及其理化因子分析[J]. 现代食品，2022，28(7)：202-206.

[3] 程再冉，缪登科，王伊伟. 关于葡萄酒质量的分级评价模型[J]. 中国高新区，2018(3)：58.

[4] 夏铭泽，石春鹏，刘征宇. 基于支持向量机的葡萄酒质量预测[J]. 制造业自动化，2020，42(5)：57-60.

[5] 赵峰，姜胜兵. 基于优化的 GA-BP 及其在葡萄酒质量预测的应用[J]. 哈尔滨商业大学学报（自然科学版），2021，37(3)：307-313.

[6] 于尚琨. 小标注样本的葡萄酒质量评估模型[D]. 上海：上海师范大学，2020.

[7] CHEN Jiayao，HUANG Hongwei，COHN Anthony G，et al. Machine learning-based classification of rock discontinuity trace：SMOTE oversampling integrated with GBT ensemble learning[J]. International Journal of Mining Science and Technology，2022，32(2)：309-322.

[8] 毕艳亮，宁芊，雷印杰，等. 基于改进的遗传算法优化 BP 神经网络并用于红酒质量等级分类[J]. 计算机测量与控制，2016，24(1)：226-228.

[9] 曹睿，廖彬，李敏，等. 基于 XGBoost 的在线短租市场价格预测及特征分析模型[J]. 数据分析与知识发现，2021，5(6)：51-65.

[10] 张春富，王松，吴亚东，等. 基于 GA_Xgboost 模型的糖尿病风险预测[J]. 计算机工程，2020，46(3)：315-320.