文章编号：2095-2163(2022)02-0079-05 中图分类号：TP183;TE624 文献标志码：A

基于随机森林的汽油精制过程中辛烷值损失模型

薛 洁

(北京信息科技大学 经济管理学院，北京 100192)

摘 要：目前，随着汽车尾气排放污染日趋严重，汽油质量标准日益严格，中国大力发展以催化裂化为核心的重油轻质化工艺技术，对汽油进行精制处理，实现汽油清洁化。在实现汽油清洁化的过程中，会不可避免地降低辛烷值(RON)，亦会同时出现较大损失值单位，无疑给企业增加了生产成本，减少了收益。为此，本文通过建立基于随机森林的汽油精制过程中 RON 损失预测模型，对 RON 及其指标进行预测。首先，命名建模变量并计算矩阵相关性，利用随机森林法对降低 RON 损失模型所涉及的 158 个变量进行二次降维，提取前 30 个主要变量；其次，基于随机森林法对样本数据进行划分，建立损失预测模型并对模型进行验证，得到预测值与真实值曲线对比图，保证所建模型合理化；最后，运用遗传算法对主要变量进行优化，力求将 RON 损失值降幅控制在 15%以上，以此确保损失预测模型真实有效。

关键词：RON 损失预测；随机森林法；遗传算法

Octane loss model in gasoline refining process based on random forest

XUE Jie

(Economics and Management College, Beijing Information Science & Technology University , Beijing 100192, China)

【Abstract】At present, with the increasingly serious automobile exhaust pollution and increasingly strict gasoline quality standards, heavy oil lightening process technology with catalytic cracking as the core, refining gasoline, and realizing gasoline cleaning are vigorously developed. In the process of realizing gasoline cleaning, the octane number (RON) will inevitably be reduced, and a large loss value unit will also appear at the same time, which will undoubtedly increase the production cost of the enterprise and reduce the income. To alleviate this problem, this paper predicts RON and its indicators by establishing a random forest-based prediction model for RON loss during gasoline refining. First, modeling variables are named, the matrix correlation is calculated, and the random forest method is used to perform secondary dimensionality reduction for the 158 variables involved in the RON loss reduction model to extract the first 30 main variables. Secondly, the sample data is divided based on the random forest method, the loss prediction model is established and verified, and the curve comparison between the predicted value and the actual value is obtained to ensure the rationalization of the model. Finally, the genetic algorithm is used to optimize the main variables and attempt to control the loss of RON to more than 15% to ensure that the loss prediction model is true and effective.

【Key words】RON loss prediction; random forest; genetic algorithm.

0 引 言

近年来，随着汽车尾气污染问题日趋严重，世界各国都制定了严格的汽油质量标准。为此，中国大力发展以催化裂化为核心的重油轻质化工艺技术，对汽油进行精制处理，以实现汽油清洁化。

经研究发现，辛烷值(RON)作为反映汽油燃烧性能最重要的指标，在实现汽油清洁化的过程中，却不可避免地出现较大的损失值单位。据统计，RON 每降低 1 个单位，相当于每吨损失约 150 元，这对于一个企业来说，无疑是增加了其生产成本，减少了收益。以一个 100 万吨/年的催化裂化汽油精制装置为例，若能降低 0.3 个单位的 RON 损失，其经济效益将达到 4 500 万元，因此，降低汽油 RON 损失具有重要的意义[1]。

本文以某石化企业为例，研究其 RON 损失值的诸多问题。经广泛收集各类相关数据，并进行相应处理，综合运用随机森林、遗传算法等统计知识建立并优化相关问题的损失预测模型，利用 SPSS(Statistical Product and Service Solutions)、Matlab(Matrix&laboratory)等软件对汽油精制过程中的 RON 损失进行可视化展示及分析，力求降低其损失值 15%以上，增加企业效益。

1 主要变量降维

1.1 建模变量命名

为了方便统计与计算，将所需的 354 个操作变量以“M+变量编号”命名，如 1 号位点氢油比命名为“M1”。同样，将 13 个材料性质以“A+变量编号”命名，如原料的 RON 命名为“A2”，依次据此方式对 366 个变量进行命名。

作者简介：薛 洁(1995-)，女，硕士研究生，主要研究方向：数字技术驱动。

收稿日期：2021-09-30

1.2 计算相关性矩阵

因样本中存在许多特征相同的变量,冗余程度较高,而相关性较强的变量较多会影响随机森林模型的准确性,使得随机森林的优势被削弱;同时,高相关度的属性会挤占其他属性被选择的机会,导致其他具有不同特征信息的属性无法得到评估,所以在使用随机森林降维之前,需对相关度较高的变量进行剔除,以此提高随机森林的泛化能力。

计算 366 个变量的相关性矩阵,按照相关度矩阵的值进行填色[2]。如图 1 所示,亮黄色和深蓝色表示变量间存在强相关性,本文定义为相关度大于 0.8,对于强相关的变量,保留其一即可,删除冗余变量后,剩余 158 个变量,再进行随机森林的构造,进行再一次降维。

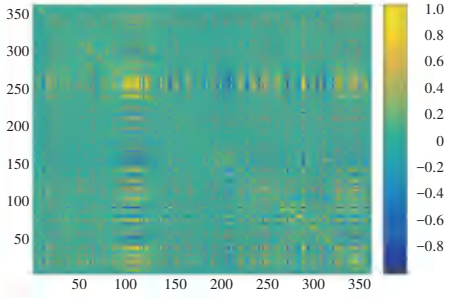

图 1 366 个变量的相关性矩阵

Fig. 1 Correlation matrix of 366 variables

1.3 随机森林降维

使用随机森林算法找出剩余 158 个变量的统计结果中信息量最大的特征子集,从而进行降维,重复 10 次实验,对 158 个变量的重要程度求平均值后进行排序,得出前 30 个主要变量,如图 2 所示。

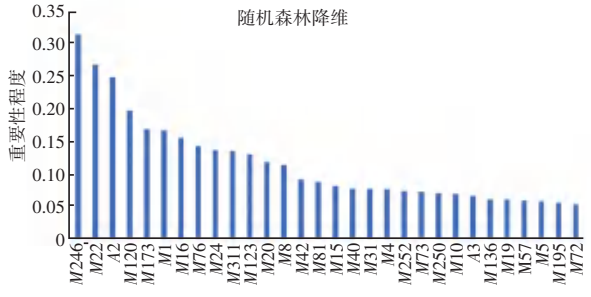

图 2 随机森林算法计算出前 30 个主要变量

Fig. 2 The first 30 main variables calculated by the random forest algorithm

对前 30 个主要变量再次进行筛选,本文保留重要性程度在 0.1 以上的主要变量,如图 2 中的 *M*246 ~ *M*8,共 13 个变量,而后使用 SPSS(Statistical Product and Service Solutions)软件对前 6 个变量进行相关性计算,得出表示相关关系强弱情况的皮尔逊相关性与显著性(双尾)计算结果,见表 1。

表 1 皮尔逊相关性与显著性(双尾)计算结果

Tab. 1 Pearson correlation and significance (two tailed) calculation results

		RON 损失	*M*246	*M*22	*A*2	*M*120	*M*173	*M*1
RON 损失	皮尔逊相关性	1	−.252**	−.185**	−0.037	0.091	.303**	.291**
	显著性(双尾)		0.000	0.001	0.501	0.100	0.000	0.000
	个案数	325	325	325	325	325	325	325
*M*246	皮尔逊相关性	−.252**	1	.127**	.161**	−0.054	−0.092	−0.094
	显著性(双尾)	0.000		0.022	0.004	0.328	0.096	0.090
	个案数	325	325	325	325	325	325	325
*M*22	皮尔逊相关性	−.185**	.127**	1	.152**	−0.089	−.398**	−.334**
	显著性(双尾)	0.001	0.022		0.006	0.109	0.000	0.000
	个案数	325	325	325	325	325	325	325
*A*2	皮尔逊相关性	−0.037	.161**	.152**	1	.347**	−.281**	−.170**
	显著性(双尾)	0.501	0.004	0.006		0.000	0.000	0.002
	个案数	325	325	325	325	325	325	325
*M*120	皮尔逊相关性	0.091	−0.054	−0.089	.347**	1	0.083	0.036
	显著性(双尾)	0.100	0.328	0.109	0.000		0.137	0.512
	个案数	325	325	325	325	325	325	325
*M*173	皮尔逊相关性	.303**	−0.092	−.398**	−281**	0.083	1	.677**
	显著性(双尾)	0.000	0.096	0.000	0.000	0.137		0.000
	个案数	325	325	325	325	325	325	325
*M*1	皮尔逊相关性	.291**	−0.094	−.334**	−170**	0.036	.677**	1
	显著性(双尾)	0.000	0.090	0.000	0.002	0.512	0.000	
	个案数	325	325	325	325	325	325	325

**.在 0.01 级别(双尾),相关性显著.

2 基于随机森林的损失预测模型

2.1 随机森林预测

随机森林是一种分类和预测集成的学习算法,其预测模型对部分变量坏值的容忍度较高,能够更好地利用不同变量与预测值之间的特征信息进行预测[3]。预测步骤如下:

(1)划分训练集与测试集:对原始样本进行划

分，选出训练集与测试集。

（2）训练预测模型：使用带有输出的训练集训练随机森林模型。

（3）对测试集进行测试：删除测试集中的输出结果，将测试集输入模型，得到测试集样本的预测值。

（4）模型评价：对模型预测的误差进行计算，得到更接进于真实值的最佳测量结果。

2.2　建立 RON 损失预测模型

首先对样本的 366 个变量进行处理，删除冗余变量，保留主要的 13 个变量；再将某石化企业的 325 个数据样本以 6∶4 的比例进行划分，随机选出训练集与测试集；构建随机森林模型，以训练集的 RON 损失值作为标签，以 13 个主要变量作为特征值输入训练模型；最后，将测试集中的 13 个变量输入到训练好的模型中，得到测试集样本的预测值，以测试集中预测值与真实值的均方对数误差作为评价指标，对模型预测的误差进行计算。随机森林模型预测值与真实值曲线对比，如图 3 所示。

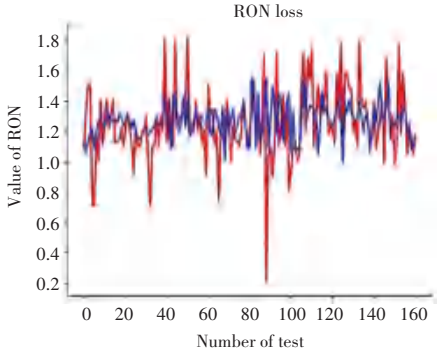

图 3　随机森林模型预测值与真实值曲线对比图

Fig. 3　Comparison of predicted value and true value curve of random forest model

3　基于遗传算法的优化预测模型

3.1　主要变量操作方案的优化

在 13 个主要变量中，除原料的 RON 是固定值以外，依次对其他 12 个操作变量进行编码，并在不同取值范围内进行限幅。将最大迭代次数设置为 100，将预测样本 RON 损失值的倒数作为个体的适应度函数，对 325 个数据样本逐一进行交叉、遗传、变异、选择等优化操作；而后运用随机森林预测模型进行封装，但个别样本的适应度在 100 次迭代内出现了明显提高，遗传算法 100 次迭代适应度变化曲线如图 4 所示。大部分数据无法在迭代内得到优化，效果并不理想，没有产生降幅大于 15%的样本。

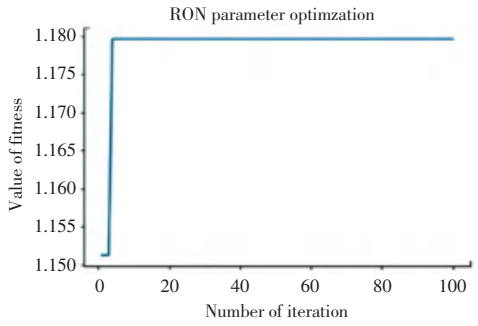

图 4　遗传算法 100 次迭代适应度变化曲线

Fig. 4　The fitness curve of 100 iterations of genetic algorithm

受计算速度和计算时间的限制，无法对全部数据增加优化的迭代次数，因此只能对小部分样本进行再一次优化[4]。如：对 129 号样本在 500 次迭代内先后进行 2 次优化，迭代适应度变化曲线如图 5 所示，其 RON 损失值由 0.9 降低至 0.78，降幅为 13.3%，依然没有产生降幅超过 15%的优化数据。

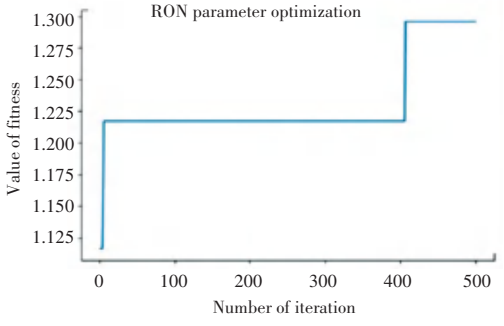

图 5　129 号样本 500 次迭代适应度变化曲线

Fig. 5　The fitness curve of sample No. 129 during 500 iterations

对 170 号样本在 1 000 次迭代内先后进行 3 次优化，迭代适应度变化曲线如图 6 所示，其 RON 损失由 0.98 降低至 0.81，降幅为 17.3%，实现了降幅超过 15%的优化目标。

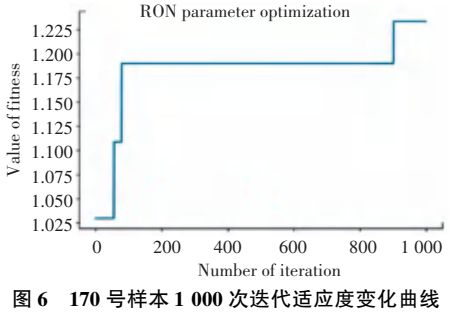

图 6　170 号样本 1 000 次迭代适应度变化曲线

Fig. 6　The fitness curve of sample No.170 during 1 000 iterations

3.2　优化预测模型的部分可视化展示

为了工业装置稳定高效运行，优化后的主要变量只能逐步调整到位。因此，若只改变一种变量，保持其他变量不变，便可得出该变量在优化调整过程中所对应的 RON 损失变化轨迹。以 133 号样本为例，其 RON 损失变化曲线如图 7 所示。

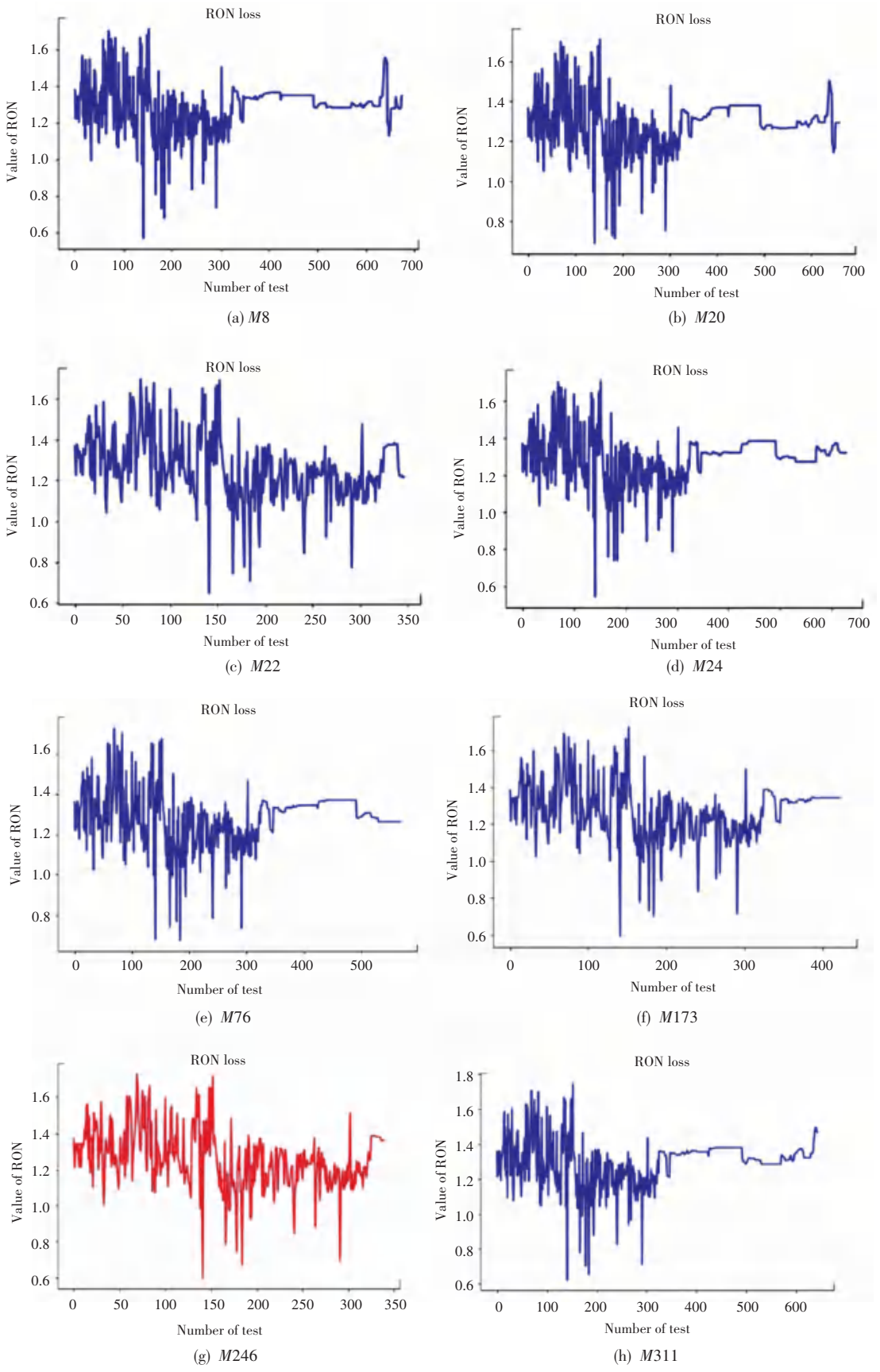

图 7 133 号样本的 RON 损失变化曲线

Fig. 7 RON loss curve of sample No.133

(下转第 90 页)