基于 SCRF 的抽油井结蜡预测方法优化研究

王利君 支志英 贾 鹿 李 伟
(中国石油新疆油田分公司数据公司 新疆 克拉玛依 834000)

摘 要 在油田生产过程中,油井受各种因素的影响容易发生结蜡。油井结蜡通常会降低油井产生,造成油井阻塞,甚至会造成停井及烧电机等现象,大大增加采油成本。对抽油井结蜡状态进行提前预测,实现抽油井设备预见性维护对油田降本增效及智能化管理具有重要意义。针对基于不平衡数据集构建结蜡预测模型预测效果不理想的问题,文中提出了一种面向非平衡数据的集成学习方法 SCRF(SMOTE CLUSTER RANDOM FOREST)。该方法首先使用 SMOTE 方法对原数据集中的少数类进行过采样以增加少数类的数量,缩小不平衡比例;然后对新的数据集采用 CLUSTER 聚类方法分层欠采样,生成训练数据集;最后采用基于 bagging 技术的随机森林算法对训练数据集进行集成学习,从而生成预测模型。实验结果表明,样本均衡后模型预测效果更佳,预测精度和效率都有一定程度的提高。

关键词 不平衡数据分类,样本均衡,集成学习,结蜡预测模型
中图法分类号 TP3 **文献标识码** A

Study on Optimized Method for Predicting Paraffin Deposition of Pumping Wells Based on SCRF

WANG Li-jun ZHI Zhi-ying JIA Lu LI Wei
(Data Company of Petrochina Xinjiang Oilfield Company, Karamay, Xinjiang 834000, China)

Abstract In the production process of oil field, paraffin deposition is easy to occur for oil wells affected by various factors. Paraffin deposition usually causes blockage of oil wells, and even causes well stuck or overload burning of electric motors, which will greatly reduce oil well production and increase the cost of oil production. So predicting the paraffin deposition state of pumping wells in advance and realizing predictive maintenance for pumping wells equipment, can reduce the cost and increase efficiency for oil fields, which have great significance on intelligent management. In order to improve the accuracy of paraffin deposition prediction based on unbalanced data set for pumping wells, this paper proposed an integrated learning method named SCRF for unbalanced data. Firstly, SMOTE method is used to oversample a few classes in the original data set to increase the number of minority classes and reduce the unbalanced proportion. Then CLUSTER clustering method is used to stratify and undersample the new data set to generate the training data set. Finally, random forest algorithm based on bagging technology is used to integrate the training data set, so as to generate the prediction model. The experimental results show that the prediction effect of the model is better after sample equalization, whilethe prediction efficiency and accuracy are improved to a certain extent.

Keywords Unbalance dataset classification, Sample balance processing, Integration algorithm, Paraffindeposition prediction model

1 引言

随着智能油田建设的不断推进,油井的智能化管理需求日益突出,尤其是抽油井的结蜡问题经常造成油井产量降低甚至停产,而目前一般采用预防性固定周期作业,但洗井周期的确定大多基于人工经验,通常存在一定的随机性和盲目性,另外,还可能存在过度洗井现象,造成资源浪费,影响生产效益。所以,预测油井结蜡状态,确定合理优化的油井清洗周期,对于提高油井开采效率,降低生产成本,实现智能化精细化管理具有重要意义。

目前,油井洗井周期的确定方法大都基于油井结蜡速度的理论模型分析,以此预测油井结蜡周期。文献[1]基于地质参数、油藏性质定量分析了影响结蜡速度的部分因素,推导出建立在层流模型上的油井结蜡故障周期的定量计算方法。文献[2]假定油井的结蜡速度与原油的含蜡量、井温分布、油井产量、油管粗糙度等因素有关,定量化计算油井结蜡速度,从而预测油井的结蜡故障周期,但这些理论计算模型大多基于静态参数,而实际油井数据与理论值存在较大偏差,无法根据生产进行动态调整,导致油井结蜡故障周期预测的准确性受到很大的限制。近年来,大数据技术的快速发展,使其迅速成各行业各领域应用的焦点,在油田领域,文献[3]提出用大数据分析方法研究抽油井结蜡规律,建立多参数动态预测模型,提前预测周油井结蜡问题,实现了抽油井结蜡周期的动态优化,并取得较好的应用效果。然而,针对抽油井的结蜡问题,

本文受新疆油田公司 2018 年信息科研项目(1.2)资助。
王利君(1987—),女,硕士,工程师,主要研究方向为油田信息化规划研究、大数据分析、系统规划与设计,E-mail:wanglj_xj@petrochina.com.cn(通信作者);**支志英**(1966—),女,高级工程师,主要研究方向为油田信息化规划研究、系统规划与设计,E-mail:zzy@petrochina.com.cn。

通常采取计划式维护的方式,但存在抽油井只是有轻微的结蜡现象就采取洗井措施的现象,因此就造成在实际生产数据中抽油井结蜡记录相比于正常生产记录较少,即抽油井结蜡数据集往往存在类别不平衡的问题。在这种不平衡数据集中,传统的机器学习算法更容易把多数类分类正确,少数类分类出错率较高,这时虽然整体的预测符合率较高,但是实际上的结蜡分析效果不理想。

针对上述现象,机器学习、人工智能领域的研究者越来越关注对不平衡数据集的分类和预测研究[4]。目前,针对非均衡数据的分类方法主要分为两类:优化算法和平衡数据[5-6]。优化算法主要是对目前的经典分类算法进行改进,通过增加这类算法在处理非均衡数据时的敏感度来提高分类效果。王伟等[7]提出了针对不平衡数据的 C4.5 决策树改进方法。Wang 等[8]提出一种新的基于权重的改进 KNN 算法用于对不平衡数据集进行分类。于化龙等[9]通过引入模糊集的思想,对传统的加权支持向量机进行改进来提高类不平衡数据的分类性能。平衡数据主要是通过改变数据的分布情况使不同类别的数据达到平衡,然后再利用传统分类算法进行进一步分类研究。采样及代价敏感学习是处理不平衡数据分类的常用方法,即通过重构样本空间来实现数据平衡或通过提高少数类样本的权重来增加少数类被错分的代价。采样方法包括欠采样和过采样两种基本的实现方法,欠采样是通过减少多数类样本来提高少数类的分类性能。过采样是通过增加少数类样本来消除或减小数据的不平衡。Ren 等[10]提出重采样技术处理非平衡数据集。Chawla[11]提出了基于 SMOTE 的方法缩小少数类和多数类的样本差距,即通过对少数类线性插值增加少数类的数量。Liu 等[12]提出了基于多数类聚类的欠采样方法生成平衡的数据集。基于欠采样和过采样,有不少研究者[13]还提出了混合采样和集成学习的方法,混合采样通过应用欠采样和过采样两种方法,构建平衡的样本数据集。集成学习通过将不平衡数据集中的多数类划分为多个子集与少数类合并,得到多个平衡的数据集,从而训练多个基分类器进行集成。优化算法和平衡数据这两种方法各有优势,因此也有不少研究学者提出将这两种方法组合在一起以进一步提高非平衡样本中小类别样本分类的准确性[14-17]。以上针对非平衡数据的分类研究大多集中在 Boosting 算法方面,Boosting 算法属于串行算法[18],分类的时间开销较大。

鉴于上述问题,本文从样本均衡和优化集成学习算法着手,针对随机采样容易产生过拟合的问题,提出基于 KNN 思想的 SMOTE 过采样方法;针对 SMOTE 及 Boosting 算法导致模型训练时间长的缺陷,提出基于 K-means 的分层欠采样方法及基于 Bagging 技术的随机森林集成学习算法以降低时间成本。多组实验结果表明,本文所提方法能取得更高的准确率和精度。

2 基于 SCRF 的非均衡数据分类方法

结合样本均衡和优化集成学习算法,本文提出基于 SCRF 的非均衡数据预测方法(SMOTE+CLUSTER+RANDOM FOREST),首先使用 SMOTE 方法对原数据集中少数类进行过采样以增加少数类的数量,缩小类别差距;然后对新的数据集采用聚类方法分层欠采样,生成训练数据集;最后采用随机森林算法(RANDOM FOREST)对训练数据集进行集成学习,从而生成准确率高的预测模型。

2.1 样本均衡算法

SMOTE(Synthetic Minority Oversampling Technique)是一种合成少数类过采样算法[10],它也是基于随机过采样算法的一种改进方案。由于随机过采样采取简单复制样本的策略来增加少数类样本,这样容易产生模型过拟合的问题,使得模型学习到的信息过于特别而不够泛化。SMOTE 算法的基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本并将其添加到数据集中,即通过增加少数类样本的方法来达到数据平衡,但可能会因为增加数据量导致训练时间过长,从而降低模型的训练效率。因此本文在基于 SMOTE 过采样的基础上提出采用聚类方法对多数类分层欠采样,减少训练数据量,在保证总体分类精度的情况下提高训练效率。即通过对 SMOTE 过采样之后的数据集中的多数类进行 K 均值聚类,得到与少数类数量相同的 K 个聚类质心,然后与少数类合并,生成最终的平衡数据集。

2.2 基于随机森林的集成学习

随机森林是由一系列分类决策树组成的集成分类器,每一棵树的构建都是基于 bagging 思想,即随机选择样本数据(有放回)来构建基决策树,最终由所有基决策树投票决定集成分类器的分类结果。bagging 技术能够自动利用 CPU 的多线程进行并行计算,所以相比于 boosting 集成算法更加高效。因此,基于平衡数据集,我们利用 C&RT 算法来训练基分类器,通过基分类器权重投票得出最终输出结果。

基于 SCRF 的非均衡数据分类算法流程如图 1 所示。

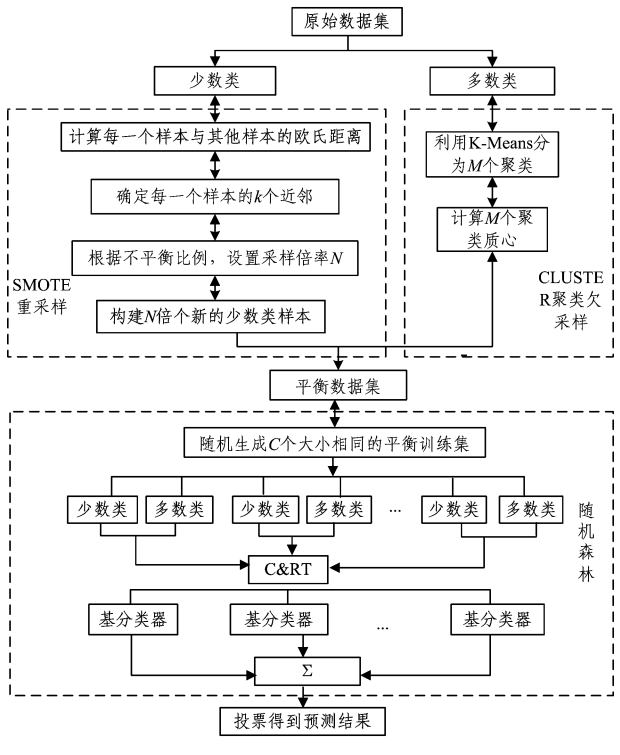

图 1 基于 SCRF 的算法流程

(1)给定一个样本,样本数为 M,其中少数类样本数为 A,多数类样本数为 B,$B \gg A$。

(2)对于少数类中每一个样本 X_i($i \leqslant A$),以欧氏距离为标准计算它到少数类样本集中所有样本的距离,得到

其 k 近邻，记为 $X_{in}(n\leqslant k)$。

(3)根据样本不平衡比例设置一个采样比例以确定采样倍率 N，对于每一个少数类样本 X_i，从其 k 近邻中随机选择 N 个近邻，其中 $N\leqslant k$。

(4)对于每一个随机选出的近邻，分别与原样本按照如下的公式构建 N 个新的样本。

$$X_{i\,\mathrm{new}}=X_i+rand(0,1)*(X_{\mathrm{in}}-X_i) \tag{1}$$

(5)重复步骤(2)—(4)，将新的样本与原始样本组成新的样本集，总样本数变为 $M+N*A$，其中少数类样本数为 $A+N*A$。

(6)对于多数类样本，采用 K-means 算法将其分为 M 个聚类，得到与少数类数量相同的 M 个聚类质心，然后与少数类合并，生成最终的平衡数据集。

(7)在平衡数据集中随机生成 C 个大小相同的训练集，基于 C 个训练集采用随机树集中构建 C 个分类树，依据构建好的多个分类器来预测新的未知样本，分类结果根据每个分类器投票结果的多数投票法来决定。

3　实验结果与分析

3.1　数据集

本文使用的数据集来源于新疆油田某采油厂上百万条抽油井生产数据。原始的数据包括组织机构数据、井基础信息、动态示功图数据、生产动态数据、油井日志数据、完井信息、热洗信息、化学清蜡信息、微生物清蜡信息，数据时间范围是 1999 年至 2017 年，其中井基础数据和组织结构数据表主要存储井名、所属机构等信息，动态示功图数据、生产动态数据、油井日志数据表主要存储示功图、电流图及动态参数，完井信息、热洗信息、化学清蜡信息、微生物清蜡信息表主要存储清防蜡措施及措施施工时间及内容。

尽管以上数据时间跨度大、数据资源丰富，但实际采集和存储的数据往往存在如下问题：1)数据记录的完整性不够；2)数据存在无效值；3)数据时间段不连续；4)数据标准不统一，不同数据库字段命名不一致；5)数据表中存在属性冗余等现象。

因此，为提高抽油井结蜡预测的准确性和精度，在建立抽油井结蜡预测模型之前需要进行数据预处理得到质量较好的记录和字段。

3.2　数据预处理

数据预处理是将与业务场景有关的来自不同数据库的数据抽取出来，中间经过清洗、转换，最后集成到统一的目标表中。

(1)数据清洗：主要用于处理由于仪器仪表损坏导致的数据异常和空值数据。对于空值数据，如果缺失严重(缺失个数大于列数 20%的数据项)则删除，相反则采用 K 邻近方法补全缺失值。动态功图数据表中电流字段空值现象较严重，无效值的比例较高，因此我们认为建模时该字段可利用的信息非常少，即视为严重缺失项。

(2)数据转换：主要包括数据归一化处理和特征构造操作。数据经过清洗后，共保留 35 个特征字段，考虑到抽油井结蜡使杆柱摩擦力增加，会造成上行载荷增加、下行载荷减小，功图面积增大(见图 2)，上下载荷会超出最大理论载荷和最小理论载荷线，即示功图载荷和面积会有较大的波动，基于此，我们通过对示功图进行切分处理，额外求取了面积、上行载荷平均值、下行载荷平均值、上行载荷变化率和下行载荷变化率 5 个新的特征值，同时构造结蜡标识字段，如果抽油井结蜡则标记为 1。

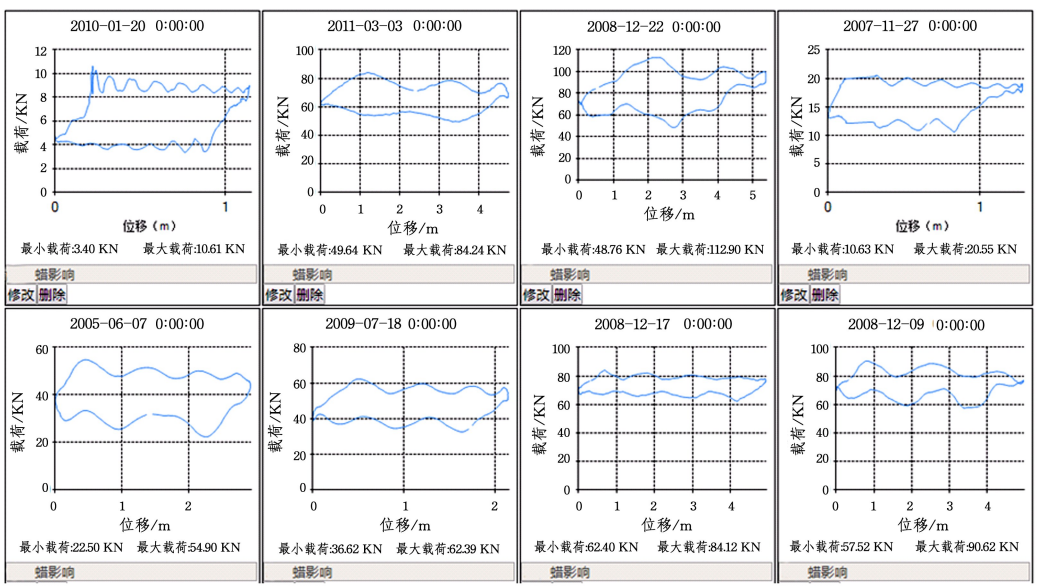

图 2　抽油井结蜡时的典型功图

原始数据经过预处理后，得到样本数据集，共包含 40 个特征参数和 1 个目标参数。为方便表示，本文将少数类样本称为正类样本，多数类称为负类样本。在所有样本中，正类样本为 3250 个，负类样本 282288 个，正类样本与负类样本比例为 1 ∶99，如表 1 所列，明显可以看出，样本存在严重的类别不平衡现象。

表 1　预处理后的样本情况

样本	数量	比例
全部	285538	1
负类	282288	0.99
正类	3250	0.01

3.3　实验设置

本文对样本数据集采取分区实验，选取其中 70%作为训

练集，其余 30%为测试样本集，实验在 4 核 CPU、3.30 GHz 主频、8GB 内存的 PC 机上进行，整个过程主要分为 3 个部分：第一部分是对样本数据集进行平衡处理解决类别不平衡的问题，第二部分是采用多种算法建立结蜡预测模型，第三部分是选取合适的评价标准综合评估模型性能。

首先，采取 4 种策略解决类别不平衡的问题。

(1)数据集 1，即负类样本与正类样本比例为 1∶87 的原始不平衡数据集，主要用于实验对比；

(2)数据集 2，聚类分层欠采样，即采用聚类算法对正类样本聚类，得到与负类样本数相同的聚类数，然后取每个聚类的中心点与负类合并为样本集；

(3)数据集 3，SMOTE 重采样，即采用 SMOTE 算法对负类样本进行分析并根据正负类样本比例人工合成新的负类样本，再将其添加到数据集中，从而避免随机采样的过拟合问题；

(4)数据集 4，SMOTE 重采样与聚类分层欠采样相结合，根据样本类别的不均衡比例，首先使用 SMOTE 方法以 3 的采样倍率对原数据集中少数类进行过采样，然后对新的数据集中的多数类采用聚类方法分为 13000 个聚类，最终生成平衡的训练数据集。

表 2 4 种策略下的样本数据集

数据集	正类/负类	属性个数	总样本个数	正类个数	负类个数
数据集 1	1∶87	41	285538	3250	282288
数据集 2	1∶1	41	6500	3250	3250
数据集 3	1∶1	41	549575	269750	279825
数据集 4	1∶1	41	26000	13000	13000

其次，我们采用 CHAID、Logistic、C5.0、神经网络和随机森林 5 种算法基于上述 4 种策略下的样本数据集构建抽油井结蜡预测模型，这样就得到 20 个模型。针对每一个模型的训练都采用同样的思路，即利用 T 时刻 40 项特征参数作为模型的输入，选取 t 作为滑动时间窗口(根据数据采集频率，这里 $t=5$)，以 $T+t$ 时刻的目标值作为模型的输出，对模型进行训练。

最后，在平衡数据的分类学习中，常采用分类的准确率，即分类正确的样本个数占总样本个数的百分比，作为评价标准；而对于非平衡数据集，多数类样本和少数类样本的分类准确率往往存在很大的差别，即模型在进行分类时会倾向于多数类样本，从而影响整体性能，但是在实际应用中，我们更关注少数类的分类准确性。针对不平衡数据集分类的评价标准主要有 F 值、G-mean、AUC[19-20]，为综合评估模型的性能，本文将采用 F 值、G-mean、AUC 来共同评价，而这 3 个指标通过二分类问题的混淆矩阵计算得出。混淆矩阵如表 3 所列，其中 TP 和 TN 分别是表示正确分类的正类和负类，FP 表示实际为负类但被预测为正类样本的个数，FN 表示实际为正类但是被分为负类样本的个数。

表 3 混淆矩阵

实际类别	预测结果	
	预测为正类	预测为负类
正类	TP	FN
负类	FP	TN

F 值、G-mean、AUC 的计算公式如下：

$$\text{F 值}=\frac{(1+\beta^2)\times \text{Recall}\times \text{Precision}}{\beta^2\times \text{Recall}+\text{Precision}} \tag{1}$$

其中，Recall 代表所有实际上为正类的样本中被正确分为正类所占的比例，Precision 反应所有被分为正类的样本中实际为正类样本所占的比例，两者都是针对正类的评价标准。“β”用来衡量 Recall 相对于 Precision 的重要度，若 $\beta>1$，则 Recall 相对重要，相反则 Precision 相对重要，为了尽可能降低 FN 的值，这里将 β 设置为 2。

$$Recall=TP/(TP+FN) \tag{2}$$

$$Precision=TP/(TP+FP) \tag{3}$$

$$G\text{-mean}=\sqrt{\text{Recall}*(\text{TN}/(\text{TN}+\text{FP}))} \tag{4}$$

AUC 是 ROC 曲线与坐标轴围成的面积，一般情况下，AUC 值越大，模型的预测性能越好。

3.4 实验结果

表 4－表 7，图 3－图 6 分别是 20 种模型在测试集上的 F2 值、G-mean 值、AUC 值、时间开销及其对比结果。由表 4－表 7 可知，CHAID、Logistic 算法对不平衡数据集的预测结果较差，对平衡的数据集预测性能有所提升，虽有改进，但对少数类和整体的预测效果仍较差。与基于多层感知器的神经网络对不平衡数据集的预测指标相比，CHAID 、Logistic 都有一定程度的提高，但模型的泛化能力较弱。C5.0 模型对平衡数据集有较好的预测能力，而对不平衡数据集的预测结果较差。随机森林模型在不同数据集下的 F2 值、G-mean 值、AUC 值都始终表现更佳，尤其是针对数据集 4 的表现最好，此外由于数据集 4 的数据量相对数据集 3 偏少，所以预测时间开销相对较小，因此在保证少数类预测准确性的同时，预测效率也有一定程度提升。

表 4 所有模型的 F2 值

数据集	CHAID	Logistic	C5.0	MLP	RANDOM FOREST
数据集 1	0.026	0.023	0.027	0.032	0.036
数据集 2	0.876	0.816	0.946	0.93	0.954
数据集 3	0.885	0.824	0.986	0.93	0.947
数据集 4	0.963	0.897	0.972	0.949	0.975

表 5 所有模型的 G-mean 值

数据集	CHAID	Logistic	C5.0	MLP	RANDOM FOREST
数据集 1	0.537	0.561	0.439	0.682	0.738
数据集 2	0.86	0.864	0.939	0.93	0.944
数据集 3	0.893	0.835	0.983	0.931	0.97
数据集 4	0.957	0.887	0.974	0.936	0.982

表 6 所有模型的 AUC 值

数据集	CHAID	Logistic	C5.0	MLP	RANDOM FOREST
数据集 1	0.678	0.699	0.5	0.752	0.786
数据集 2	0.943	0.866	0.958	0.977	0.988
数据集 3	0.952	0.909	0.99	0.978	0.987
数据集 4	0.984	0.946	0.978	0.978	0.988

表 7 所有模型的时间成本

数据集	CHAID	Logistic	C5.0	MLP	RANDOM FOREST
数据集 1	30.5	29.8	28.3	29.9	25.4
数据集 2	15.2	15	14.8	14.6	11.7
数据集 3	65.1	64.3	55.6	50.5	46.2
数据集 4	20.6	20.4	18.7	19.3	13.8

注：以分钟为单位

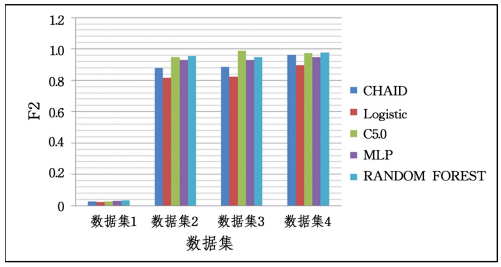

图 3　所有模型的 F2 值对比

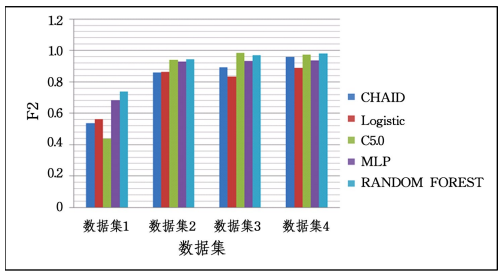

图 4　所有模型的 G-mean 值对比

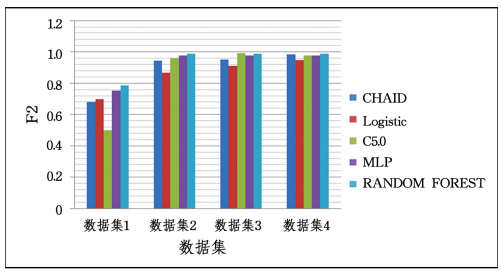

图 5　所有模型的 AUC 值对比

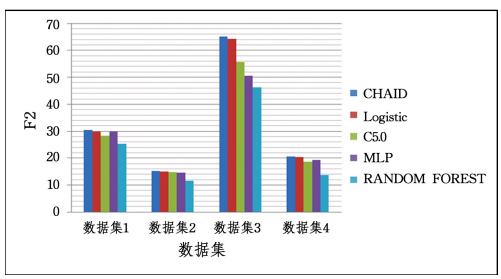

图 6　所有模型的时间成本对比

综上所述，本文提出的基于 SCRF 的非均衡数据预测方法能有效处理不平衡数据集，并且对于少数类和整体的预测都有较高的精度。

结束语　本文针对抽油井结蜡数据类别不平衡、结蜡判断不精准等问题，通过研究分析国内外对不平衡数据集的预测算法，结合实际情况，提出了一种基于 SCRF 的非均衡数据预测方法，即利用 SMOTE 方法及聚类欠采样相结合的方法处理不平衡数据集，利用随机森林算法（RANDOM FOREST）对平衡数据集进行集成学习，从而生成准确率高的预测模型。实验结果表明，本文方法在保证预测精度的同时，预测效率也有所提高。但目前在平衡数据的处理过程中采样频率及聚类数都是固定不变的，后续研究将考虑预测结果对采样频率和聚类个数的影响，从而让模型更加关注错分样本。

参 考 文 献

[1] 吴大康，吴学庆，李媛. 油井清蜡周期预测方法探讨[J]. 广东化工，2013，39(16)：53-55.

[2] 王利中. 油井结蜡速度及清蜡周期预测[J]. 西部探矿工程，2003，15(11)：54-55.

[3] 支志英，王利君，蔡志强. 基于大数据分析的抽油井结蜡预测方法研究[J]. 信息化建设，2016(2)：28-29.

[4] 向鸿鑫，杨云. 不平衡数据挖掘方法综述[J]. 计算机工程与应用，2019，55(4)：1-16.

[5] JIANG K，LU J，XIA K L. A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE[J]. Arabian Journal for Science & Engineering，2016，41(8)：3255-3266.

[6] 李艳霞，柴毅，胡友强，等. 不平衡数据分类方法综述[J]. 控制与决策，2019，34(4)：673-688.

[7] 王伟，谢耀滨，尹青. 针对不平衡数据的决策树改进方法[J]. 计算机应用，2019(3)：623-628.

[8] WANG C X，PAN Z M，MA C S，et al. Classification for Imbanlanceddataset of Impoved Weighted KNN Algorithm[J]. Computer Engineering，2012，38(20)：160-163.

[9] 于化龙，祁云嵩，杨习贝，等. 类不平衡模糊加权极限学习机算法研究[J]. 计算机科学与探索，2017，11(4)：619-632.

[10] REN S，LIAO B，ZHU W，et al. The Gradual Resampling Ensemble For Mining Inbalanced Data Steams With Concept Drift [J]. Neurocomputing，2018，286：150-166.

[11] CHAWLA N V，BOWYER K W，HALL L O，et al. Smote：Synthetic Minority Over-SamplingTechnique[J]. Journal of Artificial Intelligence Research，2002，16(1)：321-357.

[12] LIN W C，TSAI C F，HU Y H，et al. Clustering-Based Under-Sampling In Class-Inbalanced Data [J]. Information Sciences，2017，409/410：17-26.

[13] GEAPA B，RC P，MC M. A study of the behavior of several methods for balancing machine learning training data[J]. ACM Sigkdd Explorations Newsletter，2004，6(1)：20-29.

[14] IRTAZA A，ADNAN S M，AHMED K T，et al. An ensemble based evolutionary approach to the class imbalance problem with applications in CBIR[J]. Applied Sciences，2018，8(4)：495.

[15] GALAR M，FERNANDEZ A，BARRENECHEA E，et al. EUSBoost：enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling [J]. Pattern Recognition，2013，46(12)：3460-3471.

[16] 魏勋，蒋凡. 基于大规模不平衡数据集的糖尿病诊断研究[J]. 计算机系统应用，2018，27(1)：219-224.

[17] 李克文，杨磊，刘文英，等. 基于 RSBoost 算法的不平衡数据分类方法[J]. 计算机科学，2015，42(9)：249-252，267.

[18] 于玲，吴铁军. 集成学习：Boosting 算法综述[J]. 模式识别与人工智能，2004，17(1)：52-59.

[19] GAO S. An ensemble classifier learning approach to ROC optimizationPattern Recognition；Patttern Recognition [C] // 18th International Conference on ICPR. 2006：679-782.

[20] HAND D J，TILL R J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems [J]. Machine Learning，2001，45(2)：171-186.