计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(9):138−144 [doi: 10.15888/j.cnki.csa.008060]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 AERF 模型的油井结蜡预测①

常益浩, 李庆云, 李克文

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 李克文, E-mail: likw@upc.edu.cn

摘　要: 油井结蜡是一种在开发以及开采油田时对油井正常产出造成了负面影响的现象, 该现象会引起油流通道堵塞, 导致油井开采过程中出油量降低. 对油井结蜡状况做出智能预警, 完成油井设备提前修复, 对油田提高产能效率、降低维护成本及智能化管理有非常关键的价值. 为了解决油井正常数据和结蜡数据严重不平衡问题, 本文引入了自适应合成抽样法 (ADASYN) 和最近邻规则欠抽样法 (ENN) 两种非均衡样本处理方法, 分别对类别为结蜡的样本和非结蜡的样本进行处理, 最终使用随机森林算法对新构成的数据集训练, 构造出 AERF 智能模型来预测油井结蜡. 实验结果表明, 提出的 AERF 模型在油井的结蜡数据集中预测效果更佳, 明显地提高了预测精度.

关键词: 结蜡预测; 不平衡数据; 样本均衡; 抽样

引用格式: 常益浩,李庆云,李克文.基于 AERF 模型的油井结蜡预测.计算机系统应用,2021,30(9):138–144. http://www.c-s-a.org.cn/1003-3254/8060.html

Prediction of Oil Well Wax Deposition Based on AERF Model

CHANG Yi-Hao, LI Qing-Yun, LI Ke-Wen

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Wax deposition in oil wells seriously affects the normal production of oil wells during the development and production of oilfields. This phenomenon will block oil flow channels and reduce oil production during the production of oil wells. Wax deposition prediction in oil wells and advance maintenance of oil well equipment are pivotal to higher production capacity, lower maintenance cost and more intelligent management. To solve the problem of serious imbalance between the normal data and wax deposit data of oil wells, this study introduces two processing methods of non-equilibrium samples, ADASYN and ENN, which deal with the non-paraffin and paraffin samples separately. Finally, the random forest algorithm is used to integrate the training data set, and the intelligent AERF model is constructed to predict the wax deposition in oil wells. The experimental results show that the AERF model proposed in this study has a better prediction effect in the wax deposition data set of oil wells, greatly improving the prediction accuracy.

Key words: wax deposition prediction; unbalanced data; sample equilibrium; sampling

油井在开发和采油过程中会出现某些对油井的常规生产造成干预的现象, 这种现象被称作油井结蜡, 严重时其会造成油流道堵塞, 导致油井开采过程中出油量降低. 更为严重时, 会造成井筒路径被卡死等生产性问题, 甚至会造成油井的停产. 随着油田信息化的提高, 数据采集传感器被广泛应用于油井中, 因此对于每个油井都会产出和记录海量的实时相关数据, 而目前这些数据没有真正利用起来. 影响井筒结蜡的因素很多, 例如原油含蜡系数、载荷、位移、油温、压力等数据, 目前的方法主要依靠示功图等方法对结蜡进行监测分

① 基金项目: 国家自然科学基金重大项目 (51991361); 国家自然科学基金 (61673396)
Foundation item: Key Program of National Natural Science Foundation of China (51991361); National Natural Science Foundation of China (61673396)
收稿时间: 2020-11-27; 修改时间: 2021-01-04; 采用时间: 2021-01-13; csa 在线出版时间: 2021-09-02

© 中国科学院软件研究所 http://www.c-s-a.org.cn

析, 而生成示功图的数据会有大量间接误差导致图形畸变、不正确, 同时人工对示功图的判断和解读会不够全面和精准, 造成对结蜡情况产生较大误判. 所以建立一个以石油专业知识为背景的科学与智能化的油井结蜡预测模型, 提前准确地识别结蜡现象, 降低油田的风险和损失拥有显而易见的价值.

1 相关工作

目前大量学者通过现场的操作以及应用, 在对于井筒中结蜡情况的预测模型的构建方面取得了不错的成绩, 并将其顺利应用于油田的生产中. 王利中[1] 将数学与石油知识结合, 实现了对于油井结蜡快慢和结蜡形成周期长短的计算; 另外对于实际工作的油井还需要考虑其他现实因素如管内侧本身就沉淀过的结蜡和抽油杆上经过长期工作挺溜的蜡. 孙百超等[2] 在石蜡沉积机理的基础上, 结合实际生产中石油油温高、粘度小、热流强度大等特点, 将热、动力学结合, 模拟出了管线长度与结蜡厚度的分布关系模型. 但此模型采用的许多常数仅对个别油管有效, 而对与不同的油流和油井, 参数需要重新进行计算, 不具有普适性. Gawas 等提出了单相湍流条件下的沉积预测模型, 对动态循环沉积数据进行了分析之后提出了新的剪切效应关系[3]. 最近几年, 人工智能技术如火箭般突破, 使其迅速成各行业各领域应用的焦点, 在油田领域, 段友祥等[4] 利用人工智能的分类算法, 建立了异常工况诊断模型, 对油井工况中的结蜡行为进行判别和诊断; Manshad 等[5] 利用人工智能预测算法, 建立了储层流体结蜡量预测模型. 然而, 石油行业的数据集一般是不平衡数据. 针对油井结蜡的问题, 在这种不平衡数据集中, 各类基础以及传统的机器学习算法大多仍局限于均衡的样本训练, 会导致算法将更多的精力用于多数类分类, 从而使得少数类分类错误率较高, 此时即使模型整体的分类结果较好, 但是实际上的结蜡分析效果不理想.

针对类似上述问题, 在机器学习相关的领域出现了许多针对不平衡数据处理的研究方法. 解决不平衡数据的分类问题, 大抵可以归类于两种办法, 一种方法是基于数据集的数据本身, 另一种方法是脱离数据集, 尝试对算法进行创新. 对于数据本身的方法, 大多是采取对不平衡数据中的少数类或多数类样本分别实现过采样或欠采样, 去提升数据集的均衡水平, 使分类器可以在相对平衡的数据集上进行学习. Chawla 等[6] 突破性得提出了一种 SMOTE 的方法, 计算所有的少数类样本情况, 据此再构造一定数量的相似的少数类样本, 完成过采样扩充数据集. 而 SMOTE 算法所暗藏的问题是没有对少数类样本之间邻近样本的不同进行思索, 对它们以同样的权重进行合成, 可能会造成较大重复. Kermanidis 等[7] 对于数据不平衡问题采取了一种单边抽样的方法, 并利用抽样技术使得样本分类变得更加准确. Gong 等[8] 使用一种新的循环神经网络来对少数类样本进行过采样, 使得最后的分类成果十分出色. Giraldo-Forero 等提出了基于距离度量的 SMOTE 类算法[9]. de Souto 等[10] 提出了一种新的多数类样本欠采样方法, 通过对两种算法 Tomek links 和 CNN 进行结合, 最终对分类结果产生了不错的提升. Laurikkala[11] 提出了范围清算 NCL (Neighborhood CLeaning rule) 的欠采样算法. 算法层面主要通过将自己所提出的创新或结合融会于一些基础和传统的机器学习的方法及分类模型, 从而对不均衡数据的分类从另一层面产生提高和进步. Thanathamathee 等[12] 将自己提出的处理边界数据的方法与传统的 AdaBoost 算法融会, 此措施同样属于样本的过采样方法的一种. Liang 等[13] 通过在不平衡数据中使用 bagging 算法, 对多个底层的分类器进行屡次采样, 提高了二分类模型的预测效果, 也相当有效地提高了模型的分类效果. 徐丽丽等提出了一种基于集成学习的不平衡数据处理方法, 通过将各类别以不同比例进行加权并且将模糊聚类和加权支持向量机模型 WSVM 结合. 但此方法的缺点是降低不平衡数据集误分类的损失相对较大[14].

基于上述问题, 本文引入 ADASYN 来代替 SMOTE 算法的过采样作用, 并改善了 SMOTE 算法生成新样本中的“傻瓜”操作; 引入 ENN 进行欠采样来删除大部分邻居中的样本都和自己本身不属于一类的多数类样本, 删除了少部分非常相似的多数类样本, 并将新的采样算法与随机森林算法相结合, 提出 ADASYN-ENN-RF (AERF) 算法模型来预测油井的结蜡情况. 多组实验结果表明, 本文采样后的样本更能代表数据集, 本文提出的采样算法与所选分类算法的结合对现有的算法进行了提升, 具有更好的分类效果, 证明了此算法的可行和提升.

2 ADASYN-ENN-RF 算法

2.1 ADASYN 算法

ADASYN 算法又叫自适应合成采样算法[15]. ADASYN

© 中国科学院软件研究所 http://www.c-s-a.org.cn

专门针对了 SMOTE 的缺点并加以改进, 它对少数类样本不再同权重对待, 并利用少数类样本的密度分布来计算少数类样本生成的数量, 使学习困难的少数类样本生成更多的合成样本; 它能根据样本的分布来进行采样, 从而提高了少数类样本在边缘区域的比例, 可以缓解边缘区域类分布不平衡的问题, 来增强分类模型的学习能力. 既能有效地克服 SMOTE 在生成少数类样本中的盲目性, 又能较好地改善 SMOTE 在处理边缘区域对象上的局限性, 从而合理的使样本比例达到相对均衡的效果, 缓解数据不平衡的问题. 因而在提升不平衡样本学习能力上具有非常显著的优势. ADASYN 的重点是取得一个概率分布 r_i, 然后根据 r_i 计算需要构造的样本个数.

对于训练集 T 包含 p 个样本 $\{x_i, y_i\}$, i=1, 2, 3,⋯, p, 其中 x_i 是 n 维特征空间 X 的一个样本, $y_i \in Y$={0, 1}代表不同类别, 其中 y=1 代表多数类样本, y=0 代表少数类样本. 它们的数量分别用 p_l 和 p_s 表示. 所以有 $p_s \le p_l$ 且 p_l+p_s=p.

算法流程如算法 1.

算法 1. ADASYN 算法

1) 计算不平衡度 $d = p_s / p_l$, $d \in (0, 1]$;
2) 计算应该构造的样本个数 $A=(p_l-p_s)\times\beta$, $\beta \in [0, 1]$, 当 β= 1 时, 即 A 等于两大类样本个数的差值, 经过新的构造, 它们的样本个数正好相等;
3) 使用欧式距离计算所有少数类样本的邻近样本数量, 设邻近样本为 m 个, $\triangle i$ 为 m 个邻近样本中多数类样本的个数, 设比值 r_i 为 $r_i = \triangle i/m$, $r_i \in [0, 1]$;
4) 在 3) 中获得所有少数类样本的概率分布 r_i, 用 $\widehat{r_i}=r_i/\sum_{i=1}^{p_s} r_i$ 运算获得所有少数类样本的邻近样本的构成情况;
5) 通过公式获得所有少数类样本需要构造的样本个数: $a_i=\widehat{r_i}\times A$;
6) 在任何一个需要构造的少数类样本周围都有 m 个邻居, 选取其中一个, 新样本构造计算如下: $s_i=x_i+(x_{zi}-x_i)\times\lambda$;
7) 持续步骤 6) 构造样本, 当达到步骤 5) 所要求构造的样本数量即可停止.

2.2 ENN 算法

ENN 算法即最近邻规则欠采样算法[16], 该算法根据多数类样本的邻近样本是否大部分和其本身一致来对其判断是否进行欠采样, 当它周围的样本中和它类别不同的样本占据主导地位时, 就可以判定此样本点是存在问题的样本, 处理方法为删除. 以邻近样本数量 K=5 为例子, 具体的 ENN 步骤如下:

设少数类为 S, 多数类为 L, 多数类样本点记为 p.

算法 2. ENN 算法

1) 从 p 的邻近样本中, 计算找出 5 个最近的样本;
2) 对于每个点 p, 如果 5 个最近样本中有 3 个或 3 个以上的样本点不是多数类样本 L, 就删除该样本点, 反之即无操作;
3) 遍历计算, 直到没有此种多数类样本即可停止.

2.3 ADASYN-ENN-RF 算法

ADASYN 是基于 SMOTE 算法的一种改进的算法, 改进了 SMOTE 在生成少数类样本中的盲目性, 又能较好的改善 SMOTE 在处理边缘区域对象上的局限性. 但只基于 ADASYN 的 RF 算法, 利用的是已经存在的少数类样本信息来增加样本数量, 没有产生任何新的不同的知识, 在训练样本严重不平衡时, 可能会因为少数类样本欠缺空间代表性导致分类器学习的决策域变小, 从而出现过学习. ENN 算法改善了随机欠采样不考虑每个多数类样本不同且独立的近邻分布, 造成可能会误删某些重要的多数类样本信息的问题. 但只基于 ENN 的 RF 算法因为多数类样本本身远多于其他, 它们之间往往也互相紧邻, 所以仅能删除非常有限的多数类样本, 并且可能会忽略掉许多多数类样本中的有用信息, 对不平衡数据集提升有限. 所以, 本文将上述两种采样方法相结合实现数据均衡, 并提出一种基于 ADASYN-ENN 和 Random Forest 相结合的算法 (AERF).

算法的主要思想是:

1) 将不平衡数据集通过 ADASYN 算法增加少数类样本, 调整 ADASYN 算法中的 β 值来确定合成后的新数据集的样本平衡度, β=1 时, 代表多数类样本和少数类样本数量一样, 本文中取 β=0.5.

2) 将 1) 步中生成的新的数据集, 通过 ENN 算法减少多数类样本, 最终生成一个多数类样本和少数类样本数量一样的新数据集.

3) 将处理完的数据集利用随机森林算法来进行分类, 调整参数使其分类性能达到最佳.

基于 AERF 的算法如图 1 所示.

算法描述:

对于训练集 T 包含 p 个样本 $\{x_i, y_i\}$, i=1, 2, 3,⋯, p, 其中 x_i 是 n 维特征空间 X 的一个样本, $y_i \in Y$={0, 1}代表不同类别, 其中 y=0 为少数类样本即结蜡数据, y=1 为多数类样本即非结蜡数据. 它们的数量分别用 p_s 和 p_l 表示. 所以有 $p_s \le p_l$ 且 p_l+p_s=p.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

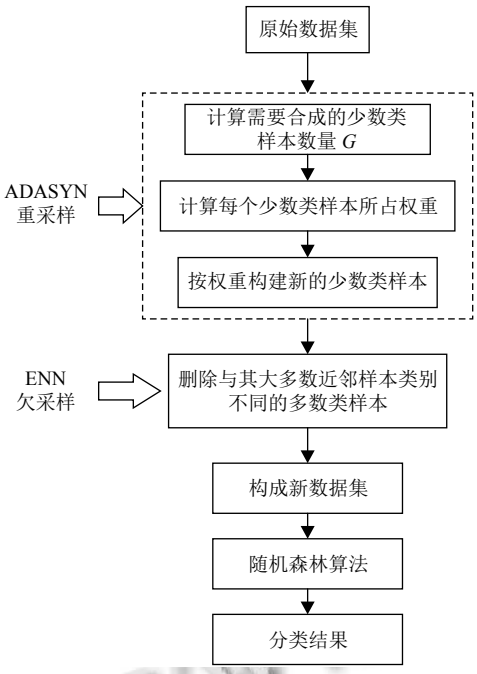

图 1 基于 AERF 的算法流程

算法步骤如算法 3.

算法 3. AERF 算法

1) 计算不平衡度 $d=p_s/p_l$, $d\in(0,1]$;
2) 计算应该构造的结蜡样本个数 $A=(p_l-p_s)\times\beta$, $\beta\in[0,1]$, 当 $\beta=1$ 时, 即 A 等于非结蜡样本减去结蜡样本的样本数量, 经过新的构造, 结蜡样本个数等于非结蜡样本个数;
3) 使用欧式距离计算所有结蜡样本的邻近样本数量, 设邻近样本为 m 个, $\triangle i$ 为 m 个邻近样本中属于非结蜡样本的个数, 记比例为 r_i 为 $r_i=\triangle i/m$, $r_i\in[0,1]$;
4) 对 r_i 进行标准化: $\widehat{r_i}=r_i/\sum_{i=1}^{ps} r_i$, $\widehat{r_i}$ 应满足式: $\sum_i \widehat{r_i}=1$;
5) 计算出所有结蜡样本需要合成的数据样本: $a_i=\widehat{r_i}\times A$;
6) 对结蜡样本生成的数据样本 $s_i=x_i+(x_{zi}-x_i)\times\lambda$ 其中 x_{zi} 是 x_i 的一个近邻样本, $x_{zi}-x_i$ 为全部属性差值, $\lambda\in[0,1]$, 直到满足数量 G 为止;
7) 使用 ENN 算法对处理过的样本集进行处理. 找出新样本集中的每个非结蜡样本的距它最近的 5 个样本点. 当在 5 个最近样本中有 3 个及其以上为结蜡样本, 则可判断此样本与周围大部分近邻样本不同, 视该非结蜡样本为噪声样本, 从数据集中删除;
8) 调用 RF 算法对平衡数据集进行分类.

3 实验结果及分析

3.1 数据来源

本文使用的数据集来源于胜利油田某采油厂的一百多万条抽油井生产数据. 原始的数据包括单井基础信息、示功图采集数据、示功图分析数据、油井实时数据、结蜡数据信息、油水分析数据, 数据时间范围是 2018 年至 2019 年, 其中单井基础数据主要存储井名、所属区域等信息, 示功图采集数据、示功图分析数据、油井实时数据主要记录各井每个时间点的示功图、油井实时参数, 结蜡数据信息主要对油井出现结蜡异常的情况进行了标记.

尽管以上数据时间跨度大、涵盖油井数量多、数据资源丰富, 但现实情况下提取和记录的数据依然存在以下缺陷: (1) 数据存在空缺值; (2) 数据存在无效值; (3) 数据收集的时间有中断, 存在没有采集到数据的时间; (4) 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度, 在建立油井结蜡预测模型之前需要进行数据预处理.

3.2 数据预处理

3.2.1 数据清洗

当示功图或油井数据中某一属性即某一列全部为空时, 对整列进行删除. 当示功图或油井数据中某时间一条数据全部为空时, 对该时间段记录进行删除. 对空缺值使用整个属性的平均值进行填补.

3.2.2 数据归一化

在模型训练与分类时, 其中的连续变量如果维度不同有可能影响到整个模型的精度. 所以在此之前, 需要对这些变量进行规范化的归一化处理, 将他们全部映射到 0~1 中. 对于分类变量, 同样需要对其做规范化的离散化处理, 将它们全部映射到 0,1 的向量空间中.

设 x_i 为来自于总体 X 的一个样本, 通过式 (1) 解决归一化, 生成新样本.

$$x_{new}=\frac{x_i-x_{min}}{x_{max}-x_{min}} \tag{1}$$

3.3 特征选择

对数据进行预处理过后, 数据集依然有 178 个特征. 许多冗余特征对模型的分类不能提供帮助甚至产生了干扰. 因此, 我们使用 mutual_info_classif 方法来对数据集进行特征选择, 选取 15 个对模型分类影响最大的特征. 最终我们将提供的结蜡数据信息中所标记为结蜡的时间, 在油井及示功图数据中将对应时间段的数据标记为 1, 其他数据标记为 0. 如表 1 所示, 繁杂的数据通过预处理过程, 构成了样本数据集, 我们从中挑选了 5 口井来进行实验.

3.4 实验设计

本文对样本数据集采取分区实验, 所用数据集如上表所示. 实验在 4 核 CPU、1.60 GHz 主频、16 GB

© 中国科学院软件研究所 http://www.c-s-a.org.cn

内存的 PC 机上进行, 使用 Python 完成了所有算法. 本文选择使用 RF, SMOTE-RF, ADASYN+RF, ADASYN-ENN-KNN, ADASYN-ENN-AdaBoost, ADASYN-ENN-RF 六种算法对相同的数据集进行测试, 采用十折交叉运算进行预测. SMOTE、ADASYN算法中领域值 K 值取 5, ADASYN 算法中 β 取 0.5. 在不平衡数据集中, 采用 Accuracy 作为评价指标不能客观的展示出模型的分类效果, 会倾向于多数类样本. 所以针对不平衡数据集, 我们采用评价指标 *F-value*、*G-mean* 和 *recall* 来对分类模型做出评价.

表 1 数据集基本信息

数据集	特征数	样本总数	少数类样本	多数类样本	少数类比例(%)
1号井	15	1719	99	1620	5.76
2号井	15	3567	45	3522	1.26
3号井	15	2273	74	2198	3.25
4号井	15	2139	11	2128	0.51
5号井	15	1560	87	1473	5.57

3.5 评价指标

在分类问题中, 使用 Accuracy 作为分类的评价指标的多是一般的平衡数据集, 但当碰到不平衡数据集[17] 时, 它的多数类少数类样本数量之差比较夸张, 所以使用 Accuracy 判断此类分类器的性能是不精准的. 因此, 本文采用不平衡分类中常用的评价指标 *F-value*、*G-mean* 和 *recall* 进行评估 (本文中定义少数类为正类, 多数类为负类). 而这 3 种指标均由混淆矩阵计算得出. 混淆矩阵如表 2 所示.

表 2 二分类的混淆矩阵

指标	预测正类	预测负类
实际正类	*TP*	*FN*
实际负类	*FP*	*TN*

表 2 中, *TP* 是指样本预测和实际特征均为正; *TN* 是指样本预测和实际特征均为负; *FN* 是指样本实际是正预测是负; *FP* 则与 *FN* 相反.

Recall、*F-value*、*G-mean* 计算公式如下:

(1) 召回率 (*recall*) 的计算公式为:

$$recall = \frac{TP}{TP+FN} \tag{2}$$

(2) 精确率 (*precision*) 的计算公式为:

$$precision = \frac{TP}{TP+FP} \tag{3}$$

(3) *F-value* 值为 *precision* 和*recall*的调和平均值, 计算公式为:

$$F\text{-}value = \frac{(1+\beta^2)\times recall \times precision}{\beta^2 + recall + precision} \tag{4}$$

参数 β 设置为 1, 如果 *precision* 和 *recall* 都高时, 很明显 *F-value* 值也会随之提升.

(4) *G-mean* 表示算法在正确正类和正确负类的平均性能:

$$G\text{-}mean = \sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}} \tag{5}$$

其值越大分类性能越好, 只有当正、负类样本的分类效果都比较好时, *G-mean* 值才会高.

3.6 实验结果

使用 RF, SMOTE-RF, ADASYN+RF, ADASYN-ENN-KNN, ADASYN-ENN-RF, ADASYN-ENN-AdaBoost 六种算法进行分类, 得到分类后的 *F*1、*recall* 和 *G-mean* 值, 从而对不平衡数据采样方法的处理效果以及分类模型的分类效果进行比较. 结果如表 3~表 5 所数据集进行处理的 RF 算法, ADASYN-ENN-RF 算法在 5 个数据集上的 3 个指标均拥有 5% 以上的提升. 对比仅使用了单一过采样算法处理数据集的 SMOTE-RF、ADASYN-RF 算法, ADASYN-ENN-RF 算法同样在 5 个数据集上的 3 个指标均产生至少 0.1% 以上的提升. 对比 ADASYN-ENN-KNN, ADASYN-ENN-AdaBoost, ADASYN-ENN-RF 三种算法, ADASYN-ENN-RF 算法在 5 个数据集上最终跑出的 *F*1 值, 均至少高出了 0.05%; 在 5 个数据集上最终跑出的 *recall* 值, 均至少高出了 0.01%; 在 5 个数据集上最终跑出的 *G-mean* 值, 均持平或高出. 结果如图 2 至图 4 所示.

表 3 各个模型在 5 个数据集上的 *F*1 值

数据集	RF	SMOTE-RF	ADASYN-RF	ADASYN-ENN-AdaBoost	ADASYN-ENN-KNN	ADASYN-ENN-RF
1	0.907	0.954	0.929	0.969	0.956	0.988
2	0.922	0.950	0.955	0.973	0.971	0.979
3	0.925	0.971	0.970	0.976	0.876	0.996
4	0.233	0.971	0.980	0.983	0.965	0.991
5	0.937	0.966	0.983	0.942	0.975	0.988

所以综合对比 *F*1、*recall*、*G-mean* 值这 3 个指标发现, AERF 算法能够有效地处理不平衡数据集并且在分类性能上有显著的提高.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

表 4 各个模型在 5 个数据集上的 *recall* 值

数据集	RF	SMOTE-RF	ADASYN-RF	ADASYN-ENN-AdaBoost	ADASYN-ENN-KNN	ADASYN-ENN-RF
1	0.853	0.961	0.911	0.945	0.953	0.976
2	0.919	0.981	0.983	0.975	0.983	0.991
3	0.933	0.955	0.960	0.968	0.880	0.976
4	0.3	0.956	0.957	0.967	0.977	0.978
5	0.943	0.966	0.984	0.983	0.958	0.991

表 5 各个模型在 5 个数据集上的 *G-mean* 值

数据集	RF	SMOTE-RF	ADASYN-RF	ADASYN-ENN-AdaBoost	ADASYN-ENN-KNN	ADASYN-ENN-RF
1	0.869	0.939	0.964	0.967	0.957	0.983
2	0.895	0.994	0.987	0.994	0.993	0.994
3	0.916	0.965	0.982	0.987	0.969	0.987
4	0.5	0.950	0.986	0.983	0.964	0.998
5	0.970	0.986	0.996	0.998	0.906	0.998

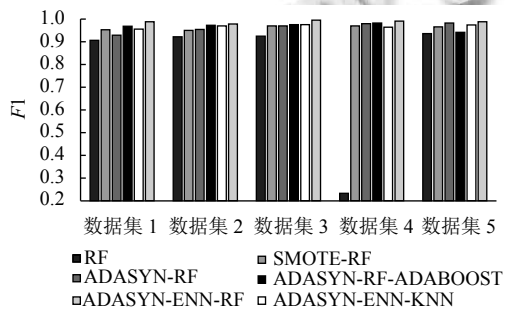

图 2 各个模型在 5 个数据集上的 *F*1 值对比

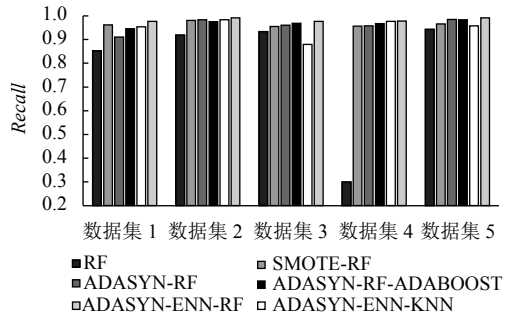

图 3 各个模型在 5 个数据集上的 *recall* 值对比

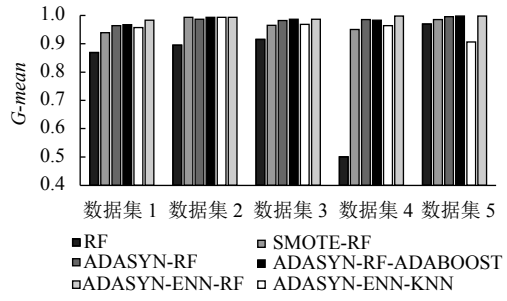

图 4 各个模型在 5 个数据集上的 *G-mean* 值对比

4 结束语

油井的运行过程中, 它的各项数据往往是极度不均衡的. 在绝大多数时间产生的都是运行正常时的样本, 而它出现问题造成结蜡的样本甚至几个月才有一次. 这很大程度增加了现在主流分类预测算法对其应用的难度. 主流算法以总体的准确性作为提升目标, 总会偏向于占比较大的类, 反而导致重要样本被忽视. 从而致使结蜡情况的出现难以被预测.

本文针对油井结蜡数据类别不平衡、判断结蜡情况不准确等问题, 提出一种 AERF 算法, 即使用 ADASYN 算法和 ENN 算法相结合的方法处理不平衡数据集, 再配合随机森林算法在平衡样本集中进行训练和学习, 从而得到效果优异的模型. 实验结果表明, 该算法在构造平衡数据集时更加合理, 特别是明显地提升了本来处于绝对弱势的少数类样本的影响力. 本文没有对 ADASYN 算法中 β 参数对于算法性能的影响并且没有对更加多元化的采样方式的组合来进行研究和实验, 这是下一步的研究方向.

参考文献

1 王利中. 油井结蜡速度及清蜡周期预测. 石油与矿业工程, 2003, 15(11): 54–55.

2 孙百超, 王岳, 尤国武. 含蜡原油热输管道管壁结蜡厚度的计算. 石油化工高等学校学报, 2003, 16(4): 48–51. [doi: 10.3969/j.issn.1006-396X.2003.04.013]

3 Gawas K, Karami H, Pereyra E, *et al*. Wave characteristics in gas–oil two phase flow and large pipe diameter. International Journal of Multiphase Flow, 2014, 63: 93–104. [doi: 10.1016/j.ijmultiphaseflow.2014.04.001]

4 段友祥, 李钰, 孙歧峰, 等. 改进的 Alexnet 模型及在油井示功图分类中的应用. 计算机应用与软件, 2018, 35(7): 226–230, 272. [doi: 10.3969/j.issn.1000-386x.2018.07.041]

5 Manshad AK, Ashoori S, Manshad MK, *et al*. The prediction of wax precipitation by neural network and genetic algorithm and comparison with a multisolid model in crude oil systems. Petroleum Science and Technology, 2012, 30(13): 1369–1378. [doi: 10.1080/10916466.2010.499403]

6 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic Minority Over-sampling TEchnique. Journal of Artificial Intelligence Research, 2002, 16: 321–357. [doi: 10.1613/jair.953]

7 Kermanidis K, Maragoudakis M, Fakotakis N, *et al*. Learning Greek verb complements: Addressing the class imbalance. Proceedings of the 20th International Conference

© 中国科学院软件研究所 http://www.c-s-a.org.cn

on Computational Linguistics. Geneva, Switzerland. 2004. 1065.

8 Gong ZC, Chen HH. Model-based oversampling for imbalanced sequence classification. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. Indianapolis, IN, USA. 2016. 1009–1018.

9 Giraldo-Forero AF, Jaramillo-Garzón JA, Ruiz-Muñoz JF, *et al*. Managing imbalanced data sets in multi-label problems: A case study with the SMOTE algorithm. Proceedings of the 18th Iberoamerican Congress on Pattern Recognition. Havana, Cuba. 2013. 334–342.

10 de Souto MCP, Bittencourt VG, Costa JAF. An empirical analysis of under-sampling techniques to balance a protein structural class dataset. Proceedings of the 13th International Conference on Neural Information Processing. Hong Kong, China. 2006. 21–29.

11 Laurikkala J. Improving identification of difficult small classes by balancing class distribution. Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe. Cascais, Portugal. 2001. 63–66.

12 Thanathamathee P, Lursinsap C. Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques. Pattern Recognition Letters, 2013, 34(12): 1339–1347. [doi: 10.1016/j.patrec.2013.04.019]

13 Liang G, Cohn AG. An effective approach for imbalanced classification: Unevenly balanced bagging. Proceedings of the 27th AAAI Conference on Artificial Intelligence. Bellevue, WA, USA. 2013. 1633–1634.

14 徐丽丽, 闫德勤. 不平衡数据加权集成学习算法. 微型机与应用, 2015, 34(23): 7–10. [doi: 10.3969/j.issn.1674-7720.2015.23.003]

15 He HB, Bai Y, Garcia EA, *et al*. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of 2008 IEEE International Joint Conference on Neural Networks. Hong Kong, China. 2008. 1322–1328.

16 Phua C, Alahakoon D, Lee V. Minority report in fraud detection: Classification of skewed data. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 50–59. [doi: 10.1145/1007730.1007738]

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. 计算机科学, 2015, 42(9): 249–252, 267. [doi: 10.11896/j.issn.1002-137X.2015.09.048]

© 中国科学院软件研究所 http://www.c-s-a.org.cn