doi：10.3969/j.issn.2095-0411.2024.03.005

Lost circulation prediction based on deep AUC maximization

LUO Junru[1], DING Yanrui[1], XU Minghua[1], HU Chao[1], LIU Bingguan[2], KONG Weijun[2], MA Qiangwei[3], SHI Lin[1]

(1. Aliyun School of Big Data, Changzhou University, Changzhou 213164, China; 2. Jiangsu Oil Field Branch, China Petroleum and Chemical Corporation, Yangzhou 225009, China; 3. Jiangsu Rutong Petro-Machinery Co., Ltd., Nantong 226400, China)

Abstract: Lost circulation is a significant challenge in oil and gas drilling, which can lead to various costly and time-consuming problems. It is of great significance to use artificial intelligence technology to accurately predict the risk of lost circulation. The lost circulation prediction problem was converted into an imbalanced classification problem, which pose challenges to traditional deep learning models due to the imbalance between categories and the lack of high correlation between drilling features. Accuracy is not an appropriate measurement for imbalanced classification algorithms. A deep AUC maximization (DAM) algorithm, which is called FAUC-S, is introduced in this paper. It trains a combination deep learning model by focusing on the AUC loss of hard samples (FAUC-S). Several traditional deep learning methods are also applied to classify lost circulation risk during oil exploration in the experiments. The result shows that the FAUC-S method achieved the highest accuracy, recall, and F_1 score among the other three models. This confirms that the FAUC-S model has superior classification performance. Therefore, the successful implementation of this deep model can help drilling teams effectively solve drilling problems.

Key words: lost circulation; imbalanced classification; deep learning; AUC maximization

基于深度 AUC 最大化算法的井漏风险预测

罗俊如[1]，丁言瑞[1]，徐明华[1]，胡　超[1]，刘炳官[2]，孔维军[2]，马强维[3]，石　林[1]

（1. 常州大学 阿里云大数据学院，江苏 常州 213164；2. 中国石油化工股份有限公司 江苏油田分公司，

收稿日期：2024-02-19。

基金项目：中国石油-常州大学创新联合体资助项目(2021DQ06)；江苏省双创博士研究资助项目(JSSCBS20210885)；常州大学阿里云大数据学院研究资助项目(ZMF21020012)。

作者简介：罗俊如(1987—)，男，江西吉安人，博士，讲师。通信联系人：石林(1979—)，E-mail：slcczu@cczu.edu.cn

引用本文：LUO J R, DING Y R, XU M H, et al. Lost circulation prediction based on deep AUC maximization[J]. 常州大学学报(自然科学版)，2024，36(3)：34-44.

江苏 扬州 225009；3. 江苏如通石油机械股份有限公司，江苏 南通 226400）

摘要：井漏是石油天然气钻井过程中经常面临的一项重要挑战，它的发生会极大的降低钻井效率，增加钻井成本。运用人工智能技术实现井漏风险的精准预测具有重要意义。文章将钻井液泄露分类问题转化为不平衡分类问题。类别间的非均衡性和钻探特征间缺乏高度相关性，对传统的深度学习模型提出了挑战。在这种情况下，传统的准确率度量很难正确评估模型性能。此外，研究引入了一种称为 FAUC-S 的深度 AUC 最大化（DAM）算法实现井漏风险预测，该算法是通过关注困难样本的 AUC 损失来训练组合深度学习模型。实验中还应用了一些经典的深度学习模型实现井漏风险的分类。实验结果表明，与其他 3 个模型相比，FAUC-S 获得了最高的精确度、召回率和 F_1 分数，实验验证了 FAUC-S 模型优越的分类性能。因此，该深度模型的成功应用可以有效地帮助钻井队解决钻井问题。

关键词：井漏；非均衡分类；深度学习；AUC 最大化

中图分类号：TP 391.4　　**文献标志码：**A　　**文章编号：**2095-0411(2024)03-0034-11

Lost circulation is a common complex downhole problem encountered in oil and gas drilling operations, where the drilling fluid (also known as mud) enters the formation instead of returning to the surface through the annulus[1]. The incidence of lost circulation in the world accounts for about 20% to 25% of the total number of drilling operations, with annual costs for plugging leaks reaching as high as 4 billion dollars. China's oil drilling engineering also faces the challenge of lost circulation[2]. According to the statistics of China Petroleum Group Oilfield Technology Service Co., Ltd., during 2017—2018, the total loss time of complex accidents in domestic and overseas blocks of China Petroleum Natural Gas Corporation Limited (abbreviated as "China Petroleum") was caused by well leaks, accounting for more than 70% of the total loss time, with an average annual economic loss exceeding 4 billion yuan; in the deep layers of the Tarim Kuqa Mountain front, a total of 76 wells were completed from 2017 to 2019, with 354 well leak incidents, of which 65% were severe well leaks.

This results in a decrease in drilling efficiency, increased costs, and potential damage to the wellbore and formation, and even blowouts[3]. To address the issue of lost circulation, various techniques have been developed, such as using lost circulation pills, water-based muds, and synthetic-based muds[4-5]. These techniques aim to reduce the fluid loss and improve the drilling efficiency. However, these methods can be time-consuming and expensive, and may not always be effective in all conditions.

Lost circulation can be divided into permeability leakage, fractures Sexual leakage and cave leakage[3]. There are several factors that can cause lost circulation, including formation pressure, formation porosity, and the rheology of the drilling fluid[6]. The early warning of lost circulation is of great significance. The occurrence of lost circulation incidents can be avoided by adjusting the wellbore trajectory or drilling parameters, thus saving on drilling time and costs[7].

To detect drilling fluid leaks accurately, it is crucial to design a classification algorithm that can classify different types of drilling fluid leaks appropriately[8]. However, for large oil fields, implementing algorithms that can efficiently process large datasets consisting of tens of thousands of data records is the most fundamental challenge[9]. Moreover, the classification problem of drilling fluid leakage is typically an imbalanced data classification problem, where the number of data records in the no/low loss category is significantly higher than in the severe/complete loss category[10]. This poses a significant challenge to the accuracy of traditional models. As a result, developing an accurate classification algorithm that can detect lost circulation is crucial.

1 Background

In recent years, the application of artificial intelligence (AI) in the oil and gas industry has gained significant attention[8, 11]. Data-driven methods have been used to predict and prevent lost circulation in drilling operations. Operators can make more informed decisions regarding drilling fluids and techniques, and optimize their operations to reduce the risk of lost circulation. For example, AI can be used to select the most suitable drilling fluid properties, such as viscosity and filtration control, based on the formation characteristics and expected pressure conditions. Additionally, AI can also help operators identify the most effective lost circulation control techniques, such as using lost circulation pills or changing the fluid loss properties of the drilling fluid[11-12].

Machine learning algorithms analyze large amounts of data collected from drilling operations, including wellbore pressure, fluid properties, and formation characteristics, to identify patterns and predict the likelihood of lost circulation occurrences[6, 12]. The tree models and network models are proposed to fit the relation between the features and the lost circulation.

Lost circulation is classified based on the loss rate. The classification includes no loss, seepage loss, partial loss, severe loss, and complete loss of drilling fluid[8]. The severity of the leakage depends on the loss rate, and each category has its unique characteristics. The drilling team needs to address these leakage issues immediately to avoid more severe problems[9-10].

While in the lost circulation prediction problems, the distribution of data is imbalanced, where there are few samples labeled with lost circulation accident. The traditional learning algorithms have poor performance. Deep AUC maximization (DAM) is a popular machine learning method for imbalanced datasets, suitable for large-scale data[13-15]. It learns an imbalanced classifier by maximizing the AUC metric which is an important assessment performance for imbalanced learning problems. It can improve classification of highly imbalanced data[16]. However, DAM has not been explored for drilling fluid leak classification in oil exploration. It is a crucial area that needs accurate classification algorithms. Applying DAM to the problem can improve safety and efficiency of oil exploration. The integration of AI in oil and gas drilling operations has the potential to improve drilling efficiency, reduce costs, and minimize the environmental impact of lost circulation.

2 Deep AUC maximization

2.1 AUC maximization problems

The area under the receiver operating characteristic curve (AUC), which is also called AUC score, is a basic performance metric for imbalanced classification algorithm. A learning classifier can be obtained by minimizing the surrogate loss of misclassification error[17-18]. The AUC metric is then used to evaluate the performance of the learned classifier. The difference between evaluation metrics and optimization metrics restricts the performance of learning algorithms. In this paper, direct method is took by maximization the AUC to fit the data. Directly maximizing the AUC score is a better choice than optimizing the accuracy metric.

The receiver operating characteristic curve (ROC) is an important graphical evaluation metric for classification problems[19-20]. In the categorical prediction phase, samples with predicted values greater than the threshold are classified as positive, otherwise they are classified as negative. As the threshold increases, the number of samples predicted to be positive decreases. Therefore, both true positive rate (TPR) and false positive rate (FPR) decrease as the threshold increases and increases as the threshold decreases. ROC reflects this co-directional variation of FPR and TPR (Fig.1).

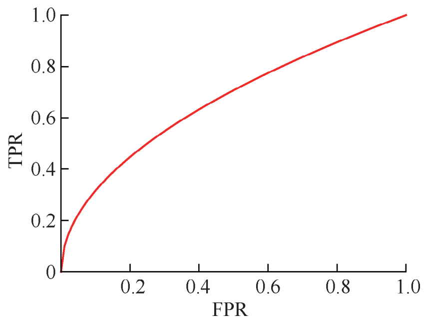

图 1 ROC 曲线

Fig.1 Curve of receiver operating characteristic

AUC is an ROC-based scalar that can be used to directly compare the predictive performance of classifiers[16]. It provides a single number that summarizes the model's performance across all possible classification thresholds. AUC is a way to measure how well a binary classification model works[21]. It looks at whether the model can identify positive cases (data that belong to the studied group) before negative cases (data that doesn't belong to the group). AUC scores range from 0 to 1, with 1 being the best score. The higher the value of AUC (λ_{AUC}), the better the classifier performance.

Given a training dataset $\{(\boldsymbol{x}_i, y_i)\}_i^n$, $\boldsymbol{x}_i \in \mathbb{R}^d$, where S_+, S_- represent the index of positive samples and negative samples, respectively, and $P = |S_+|$, $N = |S_-|$. AUC score of a classifier $f(\boldsymbol{x})$ can be calculated with the Wilcoxon-Mann-Whitney statistics[22]. That is

$$\lambda_{\mathrm{AUC}} = \frac{1}{N}\frac{1}{P}\sum_{i\in S_+}\sum_{j\in S_-} I[f(\boldsymbol{x}_i) > f(\boldsymbol{x}_j)] \tag{1}$$

where $I[\cdot]$ is the indicator function, returning 1 for the true event and 0 otherwise. Maximizing the AUC score is equivalent to minimizing the AUC risk, which is $1-\lambda_{\mathrm{AUC}}$. The corresponding optimization problem can be defined as follow

$$\min_f \frac{1}{N}\frac{1}{P}\sum_{i\in S_+}\sum_{j\in S_-} I[f(\boldsymbol{x}_i) \leqslant f(\boldsymbol{x}_j)] \tag{2}$$

The objective AUC loss is written as a sum of pairwise 0—1 loss, and this formulation makes it difficult to be optimized. The non-convex of the pairwise 0—1 loss makes the optimization of AUC risk be a difficulty problem, unfortunately.

2.2 FAUC-S method

Many methods attempted to replace the non-convex 0—1 loss function with a surrogate pairwise convex function, including pairwise hinge loss and pairwise square loss. Due to the need for pairwise comparisons between samples, the objective function has the computational complexity of $(N+P)^2$, which makes it a great difficulty in large scale imbalanced learning problems. YING et al.[23] transformed the AUC optimization problem into an equivalent min-max optimization with a square loss as the surrogate loss of AUC, and proposed a surrogate saddle loss with linear space and per-iteration time complexities of $O(d)$. The surrogate loss opens a door of deep AUC maximization method[16].

The use of squared losses produces certain negative effects. The easy samples, which are classified correctly, may suffer larger losses than those without correct classification, which is unreasonable in optimization problems. Adjusting the importance of the sample can help to construct a more reasonable way to calculate AUC loss. Recently, XU et al.[24] proposed a focus AUC loss based on samples (FAUC-S) by constructing a differentiable weight function that identifies hard and easy samples, which makes easy samples have small losses, and hard samples have large losses. It is more reasonable than the traditional AUC loss while having the same advantages on large-scale datasets. The weights of hard samples which are difficult to classifier correctly are increasing and the weights of easy samples which are classified correctly are decreasing. Let $z_i = (x_i, y_i)$, the optimization problem of the FAUC-S method can be represented as

$$\min_{w,a,b}\max_{\alpha}\sum_{i=1}^{n}\Lambda(w;z_i)F(w,a,b,\alpha;z_i) \tag{3}$$

where $F(w,a,b,\alpha;z_i)=(1-p)(f(w;x_i)-a)^2I(y_i=1)-p(1-p)\alpha^2+2\alpha(p(1-p)+pf(w;x_i)I(y_i=0)-(1-p)f(w;x_i)I(y_i=1))+p(f(w;x_i)-b)^2I(y_i=0)$ is the surrogate saddle loss of AUC metric, w, a, b are the primal parameters and α is the dual parameter. The weight function is given as

$$\Lambda(w;z_i)=\begin{cases}(1-C(f(w;x_i)-a))^{\gamma} & \text{if} \quad y=1\\(1-C(b-f(w;x_i)))^{\gamma} & \text{if} \quad y=0\end{cases} \tag{4}$$

which is used to focus on hard samples. C and γ are hyperparameters. The primal-dual stochastic compositional adaptive (PDSCA) method is employed to minimize the objective loss function[15,24]. The deep AUC maximization algorithms, including FAUC-S, can be applied in large-scale complex imbalanced learning problems, including image classification and so on. In this paper, The FAUC-S method is applied in lost circulation recognition problems due to its imbalanced distribution.

In cases with multiclassification problems, a method called "one vs all" is used. This method trains the model to recognize one group at a time, while grouping the other groups together as "not that group"[25]. This process is repeated for each group, and the scores are added to find the overall performance of the model.

3 Result

3.1 Dataset

The lost circulation dataset is a collection of a significant 65 376 data records gathered from 20 wells and is used to evaluate the effectiveness of a proposed DAM model. It is sourced from the drilling parameter dataset already published online in the southern portion of the Azadegan field in southwest Iran[8, 12]. It contains information on different categories of drilling fluid leakage losses, which are categorized as no loss, seepage loss, partial loss, severe loss, and complete loss. The number of data records belonging to each category is listed in Table 1.

表 1 井漏数据集中不同标签的样本数量

Table 1 Number of samples with different labels in the lost circulation dataset

Classification	Leakage	Fluid-loss rate	Number
0	No loss	0	49 545
1	Seepage loss	<10	12 880
2	Partial loss	10—100	2 647
3	Severe loss	>100	270
4	Complete loss	No return	34

It's worth noting that the data records belonging to class 4, which represent complete loss, account for only 0.05% of the total collected data, with just 34 complete loss records. In contrast, class

0 and class 1, which represent no loss and seepage loss, respectively, account for over 95% of all data. This distribution of data is typical in real drilling processes, and it poses significant challenges to traditional classification models due to the extreme imbalance in the number of data records in the dataset[17]. The misclassifications made to minority classes can lead to serious mining problems, making it crucial to develop models that can effectively handle imbalanced datasets.

Exploring for oil is a complex process in reality[11]. If a model can accurately identify the classification of current drilling fluid leaks, it can help drilling teams make informed decisions about appropriate plugging strategies for different leak classifications. This ensures the efficiency and safety of exploration. However, traditional models often struggle to identify minority classes, especially class 3 and class 4. Therefore, improving the precision and recall of these classes has become particularly important.

3.2 Data preparation

When collecting drilling parameter information using professional equipment in the real world, it is not uncommon to encounter outliers and missing data records for certain variables. If this raw data is used directly without being processed, it can negatively impact the model's performance and result in poor classification outcomes[18].

The original drilling dataset contains 22 variables, each of which provides information about the drilling process. However, some of these variables have missing values. It is necessary to remove incomplete data records and transform text descriptions of drilling fluid leaks into numerical types to facilitate model classification training. In addition, in order to focus on the issue of drilling fluid leakage, data records such as mechanical equipment failures were removed from the original dataset.

After filtering the data, Table 1 contains 65 376 data records, each of which has 17 features and a dependent variable for drilling fluid leakage classification, as shown in Table 2. These features include important information such as the drilling depth, the pump rate, and the mud weight.

表 2　井漏数据集中 17 个变量的取值范围

Table 2　Value range of 17 features variables in lost circulation dataset

Feature	Range
Hole section/m	0.11—0.66
Depth/m	14—4 285
Rate of penetration/($m \cdot h^{-1}$)	0—616
Weight on bit/kg	0—35 834
Rotation/($r \cdot min^{-1}$)	0—457
Torque/($N \cdot m$)	0—1 701
Standpipe pressure/MPa	0.062—26.99
Flow in/($m^3 \cdot h^{-1}$)	0—633.58
Flow out/%	0—100
Pump strokes/min^{-1}	6—242
Mud weight/($kg \cdot cm^{-3}$)	1 056—2 032
Funnel viscosity/($h \cdot m^{-3}$)	8.48—22.22
Plastic viscosity/($Pa \cdot s$)	0.003—0.048
Yield point/Pa	0.957 6—22.503 6
Gel strength(10 s)/Pa	0.478 8—12.448 8
Gel strength(10 min)/Pa	0.478 8—12.927 6
Solid/%	4—65

To avoid the bias of model numerical calculation weighting caused by excessively large values of a single feature variable, the minimum and maximum values of the 17 features data were standardized. This is done to eliminate the influence of dimensionality, so that the features can be compared in the same numerical interval. By eliminating the influence of feature magnitude caused by dimensionality on analysis results, the features of different dimensions become comparable, and the model can better capture the relationships between them[26].

The input features were normalized, and gradient descent optimization models were used. Normalizing input features can prevent some features from dominating others, while gradient descent optimization models can speed up the convergence process of the algorithm.

3.3 Classification model for lost circulation

Based on our analysis of the feature and class distribution of drilling datasets, the classification problem of drilling fluid leakage is essentially a problem of imbalanced classification. To address this issue, a deep neural network model was developed that used a compositional training approach. The backbone of the model is resnet-20, and cross-entropy loss is used as its internal function. Additionally, the method called FAUC-S is used[24], which maximizes the AUC score for hard samples, to help us learn a robust classifier that can handle minority classes better.

To solve the imbalanced lost circulation classification problem, three traditional network models are applied to test the dataset. The first model is a convolutional neural network (CNN) that creates a one-dimensional convolutional layer consisting of 64 filters with a kernel size of one[27]. The second model is a long short-term memory (LSTM) layer containing 32 neurons. The last model is a network similar to LSTM, which contains 32 gated recurrent units (GRU) nodes[28].

In the next experimental section, these four network models will be applied to the lost circulation classification problem to further verify the effectiveness of all methods. Our goal is to develop a comprehensive and robust classification system that can accurately identify drilling fluid leakage, even in cases where the classes are imbalanced.

In the domain of lost circulation prediction problem, a structured dataset was created and subjected to preprocessing. The outcome was standardized numerical data that was utilized for training a machine learning model. To retrieve input data, the Python programming language was employed in conjunction with the sklearn library. The dataset was then bifurcated into a training and testing set in a 7 ∶ 3 ratio, with 65 376 data records being utilized. It was ensured that the testing set did not participate in the training process[29]. Each model was trained with a total of 100 epochs and a learning rate of 0.01. All models achieved convergence results. The final performance of the five classification models was calculated using the one vs all training method[25].

3.4 Model comparison results

Binary classification problems, the confusion matrix plays an important role in assessing a learning classifier, which can be rewritten as a contingency table (Table 3).

表 3 混淆矩阵

Table 3 Confusion matrix

Confusion matrix		Predicted	
		Positive	Negative
Actual	Positive	T_P	F_N
	Negative	F_P	T_N

To make it easier to compare the performance of different classifiers, extract some statistics scalarsare needed from the confusion matrix. In the context of traditional classification models, accuracy (A) is commonly utilized to evaluate the performance of the classification process, which is defined as

$$A=\frac{T_P+T_N}{T_P+F_P+T_N+F_N} \tag{5}$$

Where true positive (T_P) and true negative (T_N) to represent correctly predicted positive and negative labels, respectively. And false positive (F_P) and false negative (F_N) represent incorrectly predicted positive and negative labels, respectively. However, in real-world scenarios, the problem of imbalanced classification may arise where the distribution of the classes is not uniform. In such cases, accuracy alone may not be sufficient to reflect the classification information correctly. Hence, precision (P_{re}), recall (R), and F_1 score are often considered as additional performance metrics to evaluate the classification process[30].

$$P_{re}=\frac{T_P}{T_P+F_P} \tag{6}$$

$$R=\frac{T_P}{T_P+F_N} \tag{7}$$

$$F_1=\frac{2\times(R\times P_{re})}{R+P_{re}} \tag{8}$$

Where P_{re} represents the proportion of samples predicted to be positive that are truly positive, R represents the proportion of positive samples that are accurately predicted, and F_1 score is the harmonic average of precision and recall.

The performance of four different models in classifying data is evaluated using precision, recall, and F_1 scores, which are calculated separately for each model on the test data. The scores of these models for different metrics on the test data are presented in Table 4. The one-vs-all method are used to train the multi-class lost circulation prediction problem, and the average performance metric are recorded in Table 5.

The experimental results reveal that FAUC-S outperforms the other models in this task. It shows exceptional performance in no loss, seepage loss, partial loss, and complete loss categories. In addition, among the majority classes of no loss and seepage loss, it still performs the best.

表 4　4 种不同模型得到的性能指标

Table 4　Performance measurements recorded under four different models

Methods	Metrics	Classification				
		0	1	2	3	4
LSTM	Precision	0.98	0.80	0.74	0.78	1.00
	Recall	0.94	0.88	0.89	0.86	0.80
	F_1	0.96	0.84	0.81	0.82	0.89
GRU	Precision	0.98	0.80	0.75	0.76	1.00
	Recall	0.94	0.87	0.88	0.85	0.40
	F_1	0.96	0.84	0.81	0.80	0.57
CNN	Precision	0.99	0.94	0.75	0.85	0.88
	Recall	0.98	0.90	0.95	0.94	0.70
	F_1	0.99	0.92	0.84	0.89	0.78
FAUC-S	Precision	0.99	0.96	0.85	0.77	1.00
	Recall	0.98	0.97	0.87	0.88	0.99
	F_1	0.98	0.97	0.86	0.82	0.99

Furthermore, on an extremely small number of complete losses, FAUC-S scores far exceed those of the other models. Thus, it can be concluded that FAUC-S is the best model for the classification task.

It has been observed that in cases of "severe loss", the performance of FAUC-S is not as good as that of CNN. This could be because the model is not able to fully distinguish between "partial loss" and "severe loss". Additionally, CNN performs better in the class 0—3 as compared to LSTM and GRU. However, LSTM outperforms CNN only in cases of "complete loss".

表 5 4 种不同模型的多分类井漏风险预测性能指标

Table 5 Multiclass performance measurements recorded under four different models

Methods	Multiclass metrics	Value
LSTM	Precision	0.934 0
	Recall	0.925 8
	F_1	0.929 7
GRU	Precision	0.934 3
	Recall	0.923 1
	F_1	0.929 4
CNN	Precision	0.969 8
	Recall	0.962 7
	F_1	0.969 6
FAUC-S	Precision	0.977 5
	Recall	0.973 2
	F_1	0.972 5

It is important to note that LSTM, GRU, and CNN are all affected by the imbalanced distribution of data. This results in a skewed classification performance towards the majority class and poor classification performance towards the minority class. Despite this, in cases where the other three models show low recall and F_1 scores, FAUC-S results in higher recall and F_1 scores. Overall, the performance of FAUC-S is better suited for the distribution of imbalanced data in cases of drilling fluid leaks.

To further evaluate the performance of the CNN and FAUC-S in imbalanced classification, we have opted to compare the test AUC score. The FAUC-S model achieved a higher AUC score than the CNN model, as demonstrated in Table 6. This result provides further evidence of the effectiveness of the FAUC-S method in addressing the challenge of drilling fluid leakage.

表 6 测试数据集的 AUC 得分统计

Table 6 Test AUC scores of CNN and FAUC-S models

Methods	Classification				
	0	1	2	3	4
CNN	0.98	0.97	0.97	0.99	0.97
FAUC-S	0.99	0.99	0.98	0.98	0.99

4 Discussion

The distribution of data pertaining to drilling fluid leakage in oil exploration is characterized by its complexity and imbalance, which poses significant challenges to traditional classification models. Accurately classifying this data is essential for drilling teams to mitigate any potential adverse effects during industrial operations. With this task in mind: ① To ensure successful training of subsequent models, it is imperative that filter, clean, and standardize the collected raw data. The quality of data has a significant impact on the outcome of the models. ② To achieve comprehensive classification, it is important to consider not only accuracy but also precision, recall, F_1, and other metrics. ③ While traditional classification models can accomplish classification, they are often influenced by the majority class and struggle to accurately identify minority classes. ④ Selecting a deep AUC maximization method that is specifically designed for imbalanced classification tasks can greatly enhance model performance and should be considered.

5 Conclusions

Artificial intelligence technology has a very important and wide application in oil drilling engineering, which can help drilling enterprises effectively reduce drilling costs and improve drilling efficiency. This paper discusses the application of the latest deep AUC maximization algorithm to solve the lost circulation risk prediction problem. The lost circulation dataset is imbalanced and the traditional methods can't effectively and accurately identify the risk of well leakage. Directly optimizing the imbalanced classification performance metric is helpful. FAUC-S proposed a differentiable weight function that identifies hard and easy samples, which makes hard samples with larger losses. Experimental results show that FAUC-S has better prediction performance than the traditional method in lost circulation risk prediction problem.

References:

[1] CHEN G H, HE Y L, QIU Z S, et al. Research and application of fine identification method of lost circulation characteristics in drilling[J]. Petroleum Drilling Techniques, 2024, 52(1): 1-6.

[2] SUN J S, BAI Y R, CHENG R C, et al. Research progress and prospect of plugging technologies for fractured formation with severe lost circulation[J]. Petroleum Exploration and Development, 2021, 48(3): 732-743.

[3] 李大奇，康毅力，刘修善，等. 裂缝性地层钻井液漏失动力学模型研究进展[J]. 石油钻探技术，2013，41(4)：42-47.

[4] APALEKE A S, AL-MAJED A, HOSSAIN M E. Drilling fluid: state of the art and future trend[C]// SPE North Africa Technical Conference and Exhibition. Cairo: SPE, 2012: SPE-149555-MS.

[5] 陈卓，耿立军，岳明，等. 渤中 13-2 区块复杂薄弱层高强度承压堵漏技术研究及应用[J]. 常州大学学报(自然科学版)，2023，35(4)：77-86.

[6] KALHOR MOHAMMADI M, RIAHI S, BOEK E S. An insight review on formation damage induced by drilling fluids[J]. Reviews in Chemical Engineering, 2023, 39(3): 387-415.

[7] KIRAN R, TEODORIU C, DADMOHAMMADI Y, et al. Identification and evaluation of well integrity and causes of failure of well integrity barriers (a review)[J]. Journal of Natural Gas Science and Engineering, 2017, 45: 511-526.

[8] MARDANIRAD S, WOOD D A, ZAKERI H. The application of deep learning algorithms to classify subsurface drilling lost circulation severity in large oil field datasets[J]. SN Applied Sciences, 2021, 3(9): 785.

[9] SULE I, IMTIAZ S, KHAN F, et al. Risk analysis of well blowout scenarios during managed pressure drilling operation[J]. Journal of Petroleum Science and Engineering, 2019, 182: 106296.

[10] WOOD D A, MARDANIRAD S, ZAKERI H. Effective prediction of lost circulation from multiple drilling variables: a class imbalance problem for machine and deep learning algorithms[J]. Journal of Petroleum Exploration and Production Technology, 2022, 12(1): 83-98.

[11] KUANG L C, LIU H, REN Y L, et al. Application and development trend of artificial intelligence in petroleum exploration and development[J]. Petroleum Exploration and Development, 2021, 48(1): 1-14.

[12] KAMGUE LENWOUE A R, LI Z H, TANG C F, et al. Recent advances and challenges of the application of artificial intelligence to predict wellbore instabilities during drilling operations[J]. SPE Drilling & Completion, 2023, 38(4): 645-662.

[13] ZHU X Y, HAO J, GUO Y H, et al. AUC maximization in imbalanced lifelong learning[C]//Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence. Pittsburgh: ACM, 2023: 2574-2585.

[14] YUAN Z N, YAN Y, SONKA M, et al. Large-scale robust deep AUC maximization: a new surrogate loss and empirical studies on medical image classification[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 3020-3029.

[15] YUAN Z N, GUO Z S, CHAWLA N, et al. Compositional training for end-to-end deep AUC maximization[C]//International Conference on Learning Representations. [S. l. : s. n.], 2022.

[16] YANG T B, YING Y M. AUC maximization in the era of big data and AI: a survey[J]. ACM Computing Surveys, 2022, 55(8): 172.

[17] WANG X Y, LYU Y L, JING L P. Deep generative model for robust imbalance classification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 14112-14121.

[18] JAN B, FARMAN H, KHAN M, et al. Deep learning in big data analytics: a comparative study[J]. Computers and Electrical Engineering, 2019, 75: 275-287.

[19] CARRINGTON A M, MANUEL D G, FIEGUTH P W, et al. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 329-341.

[20] JOLOUDARI J H, MAREFAT A, ALI NEMATOLLAHI M, et al. Effective class-imbalance learning based on SMOTE and convolutional neural networks[J]. Applied Sciences, 2023, 13(6): 4006.

[21] YANG Z Y, XU Q Q, BAO S L, et al. AUC-oriented domain adaptation: from theory to algorithm[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14161-14174.

[22] HANLEY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143(1): 29-36.

[23] YING Y, WEN L, LYU S. Stochastic online AUC maximization[C]//Neural Information Processing Systems. 2016, 29: 451-459.

[24] XU S K, DING Y R, WANG Y H, et al. FAUC-S: deep AUC maximization by focusing on hard samples[J]. Neurocomputing, 2024, 571: 127172.

[25] SAITO K, SAENKO K. OVANet: one-vs-all network for universal domain adaptation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 8980-8989.

[26] MWANGI B, TIAN T S, SOARES J C. A review of feature reduction techniques in neuroimaging[J]. Neuroinformatics, 2014, 12(2): 229-244.

[27] RAMSUNDAR B, ZADEH R B. TensorFlow for deep learning: from linear regression to reinforcement learning [M]. Sebastopol: O'Reilly Media, Inc. , 2018.

[28] GAO Y, GLOWACKA D. Deep gate recurrent neural network[EB/OL]. (2016-04-11)[2024-03-04]. http://arxiv.org/abs/1604.02910.pdf.

[29] YU K P, TAN L, MUMTAZ S, et al. Securing critical infrastructures: deep-learning-based threat detection in IIoT [J]. IEEE Communications Magazine, 2021, 59(10): 76-82.

[30] LUO J R, QIAO H, ZHANG B. A minimax probability machine for nondecomposable performance measures[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(5): 2353-2365.

（责任编辑：谭晓荷）