Pattern recognition for water flooded layer based on ensemble classifier

ZhiQiang Geng, Xuan Hu, QunXiong Zhu, YongMing Han*, Yuan Xu, Yanlin He
College of Information Science & Technology, Beijing University of Chemical Technology, Beijing 100029, China;

Engineering Research Center of Intelligent PSE, Ministry of Education in China, Beijing 100029, China.

* (Corresponding author:E-mail: hanym@mail.buct.edu.cn)

***Abstract*—In order to establish an effective water flooded layer recognition model to deal with complex chromatogram data and correctly identify the water flooded layer in the oil and gas reservoirs, this paper proposes a modeling approach based on ensemble classifier. First, the proposed approach utilizes the function fitting method to obtain the effective chromatogram characteristic information (CCIs). Moreover, in order to transform the sparse classification problem into a general classification problem, the synthetic minority over-sampling technique (SMOTE) algorithm is used to process the unbalanced training sample as a general training sample. Compared with the traditional classification approach, the robustness and effectiveness of the ensemble classifier model composed of the model-free classification (MFBC) algorithm, the k-nearest neighbor (KNN) algorithm and the support vector machine (SVM) algorithm were validated through the standard data source from the UCI (University of California at Irvine) repository. Finally, the proposed model is validated through an application in a complex oil and gas recognition system of China petroleum industry. The CCIs and the prediction results are obtained to provide more reliable water flooded layer information, guide the process of reservoir exploration and development and improve the oil development efficiency.**

Keywords—ensemble classifier; model-free classification algorithm; k-nearest neighbor; support vector machine; synthetic minority over-sampling technique; water flooded layer identification

I. INTRODUCTION

At present, most of the oilfields in China are in the middle and late stage of development, the water content in the production fluid of oil wells is getting higher and higher [1]. In the combination of all kinds of fuzzy conditions, oil water-flooded behaves as none water flooded, weak water flooded, middle water flooded and strong water flooded [2]. According to the evaluation of water-flooded degree of water flooding reservoir, people make a mining plan to find the remaining oil. Therefore, the accurate identification and evaluation of water flooded layer is of great practical significance for the exploration and development of oil and gas fields [3, 4]. By the pyrolysis gas chromatography technology [5], we can get quantitative analysis data of reservoirs.In order to enhance the model generalization ability and reduce the over-fitting, we extract the process characteristic information through the function fitting method [6]. Since the training data set is an imbalanced data set, we use the synthetic minority over-sampling technique (SMOTE)[7, 8] to increase the sparse class.The traditional classification method is to find a classifier closest to the actual classification function in a space composed of various possible functions, such as model-free classification (MFBC) [9, 10], k-nearest neighbor (KNN) [11, 12] and support vector machine (SVM) [13, 14]. The classification accuracy can be improved by combining the prediction results of multiple classifiers, which is called the ensemble classifier [15]. Compared with various single classification models, the ensemble classifier model has better classification accuracy and more stable performance [16, 17]. By combining the above algorithm, this paper firstly extracted and preprocessed the raw chromatogram data as a training sample. Then we build an ensemble classifier consisting of MFBC, KNN, and SVM and verified its effectiveness through the standard data set from the UCI (University of California at Irvine) repository. Finally, we established a water flooded layer recognition model based on the ensemble classifier.

This research was partly funded by National Natural Science Foundation of China (61374166, 61603025), the Natural Science Foundation of Beijing, China (4162045) and National Key Technology Support Program (2015BAK36B04)

II. DATA PREPROCESSING

A. Data source and preprocessing

In this paper, a total of 413 chromatograms of the water flooded layer have been obtained from a oil field in China. Each chromatogram has 40000 lines of data. These chromatogram data have characteristics of high-dimension, complexity and noise [18]. For the thickened oil and heavy oil chromatograms, we get the area parameter by using the Fourier function to fit the interval minimum, the result is shown in Fig.1.

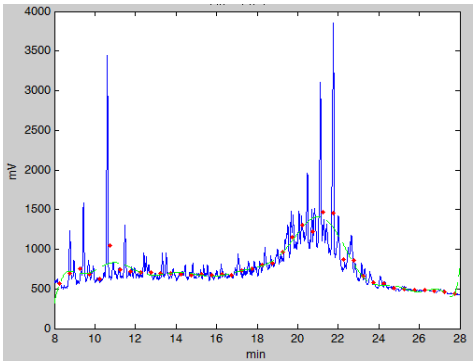

Fig. 1. The low yield reservoir of thickened oil

For the light oil chromatograms [19], we get the shape parameters by using the gamma distribution to fit the interval maximum, the result is shown in Fig.2.

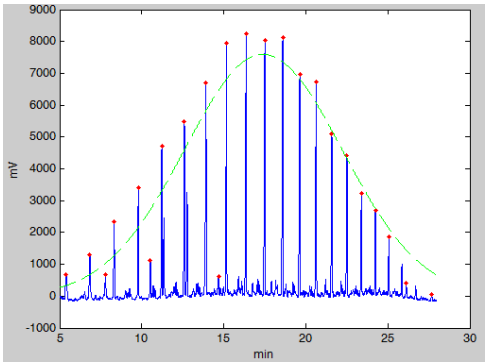

Fig. 2. The poor reservoir of light oil

In addition, we extract other parameters, including maximum, minimum, mean, standard deviation, variance, skewness and kurtosis. A part of preprocessed data are shown in Table 1:

TABLE I. CHROMATOGRAM PRE-PROCESSED DATA

parameters	***1***	***2***	***3***	4	5
category	oil layer	oil layer	oil layer	oil layer	oil layer
peak number	12	15	12	13	12
maximum	25800	1477	3905	24459	2614
minimum	1709	344	493	2058	472
mean	6447.07	475.27	1196.37	5854.59	863.26
median	25789	1477	3905	24225	2614
mode	2757	351	647	3122	477
standard deviation	3983.46	132.81	586.64	3340.25	346.82
variance	15867970	17640	344150	11157286	120281
skewness	1.42	2.61	1.32	2.11	1.34
kurtosis	5.15	10.04	4.62	6.96	4.28
Tmax	25.49	21.76	21.82	25.78	22.18
fit area	32273.38	2375.82	5981.05	29267.38	4316.12

B. Synthetic minority over-sampling technique (SMOTE)

Suppose there are sparse samples, each sample is x, search for its k nearest neighbor samples, If the rate of the up-sampling is N, N samples were randomly selected in their k nearest neighbor samples: $y_1, y_2, \ldots, y_n$, randomly linearly interpolated between sparse samples x and $y_j (j = 1,2,\ldots,n)$ to construct new sparse samples:

$$p_j = x + rand(0,1) \times (y_j - x) \quad , \quad j = 1,2,\ldots,n \ ,$$

where $rand(0,1)$ is a random number within the $(0,1)$. We add p_j to the old data set and get a new training set.

TABLE II. CLASS AND NUMBER OF SAMPLES USED IN THE EXPERIMENTS

ID	***Water flooded type***	***Number of samples***	***Number of samples after SMOTE***
1	Oil layer	12	48
2	Weak water flooded	57	57
3	Middle water flooded	49	49
4	Strong water flooded	52	52

III. THE ENSEMBLE CLASSIFIER METHOD

A. Model-free classification algorithm (MFBC)

Suppose M samples $X = [X_1 \ldots X_M]^T$, the corresponding category is $Y = [Y_1 \ldots Y_M]^T \ y_t \in D = \{d_1, d_2 \ldots d_t\}$, we need to predict the category of each sample X_t, and we can get the category of the sample by (1) :

$$d_t = \arg\max\left(p(y_t|X_t)\right) y_t \in \left\{d_1, d_1, \ldots, d_q\right\} \quad (1)$$

then we can get $p\left(y_t|X_t\right)$ by the Bayes theorem [20], as follows:

$$p\left(y_t|X_t\right) = \frac{p\left(X_t|y_t\right) * p\left(y_t\right)}{p\left(X_t\right)} \propto p\left(X_t|y_t\right) * p\left(y_t\right) \quad (2)$$

We use (3) :

$$p(X_t) = \frac{2^{-\left(r*\log_2(\tau_t+1)+\log_2\left(\frac{T*(M-1)}{r}\right)+\frac{r}{\ln(2)}\right)}}{Z} \quad (3)$$

For the sample $Y \sim Multinomial(\phi_1, \phi_2, \ldots, \phi_q)$, $p(Y = d_j) = \phi_j, \phi_j \in [0,1]$, we can (4), where $1\{*\}$ is indicator function.

$$p(y_t) = \sum_{j=1}^{q} \phi_j^{1\{y_t = d_j\}} \quad (4)$$

Then we get the logarithmic likelihood function, as follows:

$$\ell(\phi_1, \phi_2, \ldots, \phi_q) = \ln\left(\prod_{i=1}^{M} p(y_i|X_i)\right) = \sum_{i=1}^{M} \ln\left(p(y_i|X_i)\right) \quad (5)$$

We use (6) to maximize the logarithmic likelihood function:

$$\max_{\phi_1, \phi_2, \ldots, \phi_q} \ell(\phi_1, \phi_2, \ldots, \phi_q) = \max_{\phi_1, \phi_2, \ldots, \phi_q} \sum_{i=1}^{M} \ln\left(\sum_{j=1}^{q} \phi_j^{1\{y_i = d_j\}}\right) \quad (6)$$

Because $\sum_{j=1}^{q} \phi_j = 1$, and we need to get q-1 parameters, we can respectively take the partial derivatives for q-1 parameters, as shown in (7), we can get the parameters, as shown in (8).:

$$\nabla_\phi k = \begin{bmatrix} \frac{\partial k}{\partial \phi_1} \\ \vdots \\ \frac{\partial k}{\partial \phi_{\text{q-1}}} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \vec{0} \quad (7)$$

$$\ell(\phi_1, \phi_2, \ldots, \phi_q) = \sum_{i=1}^{M} \ln\left(\sum_{j=1}^{q-1} \phi_j^{1\{y_i = d_j\}}\right) + \sum_{i=1}^{M} \ln\left(\left(1 - \sum_{j=1}^{q-1} \phi_j\right)^{1\{y_i = d_j\}}\right) \quad (8)$$

So we can get

$$\begin{bmatrix} \frac{\partial\left(\sum_{i=1}^{M} 1\{y_i = d_1\} * \ln \phi_1 + \sum_{i=1}^{M} 1\{y_i = d_q\} * \ln\left(1 - \sum_{j=1}^{q-1} \phi_j\right)\right)}{\partial \phi_1} \\ \vdots \\ \frac{\partial\left(\sum_{i=1}^{M} 1\{y_i = d_{q-1}\} * \ln \phi_{q-1} + \sum_{i=1}^{M} 1\{y_i = d_q\} * \ln\left(1 - \sum_{j=1}^{q-1} \phi_j\right)\right)}{\partial \phi_{\text{q-1}}} \end{bmatrix} = \vec{0}$$

(9)

$$\phi_j = \frac{\sum_{i=1}^{M} 1\{y_i = d_j\}}{M}, (j = 1,2,\ldots,q) \quad (10)$$

Finally, $p(y_t)$ is obtained according to (4).

B. K-nearest neighbor (KNN)

In 1986, Vapnik proposed a dependency estimate based on empirical data. The KNN is a machine learning method based on the vector space model, which is used to classify unmarked test sets. The algorithm is as follows:

$$y(X, c_j) = \sum_{d_i \in KNN} f(X, d_i, c_j, b_j) \quad (11)$$

$$f(X, d_i, c_j, b_j) = \begin{cases} 1, sim(X, d_i)\delta(d_i, c_j) - b_j > 0 \\ 0, sim(X, d_i)\delta(d_i, c_j) - b_j \le 0 \end{cases} \quad (12)$$

Suppose sample X can be classified, d_i is the i^{th} of k nearest representative points, c_j is the category $\delta(d_i, c_j) \in \{0,1\}$, if d_i belongs to c_j , $c_j = 1$, otherwise $c_j = 0$. The b_j is the threshold of c_j . $sim(X, d_i)$ is the similarity between the sample X and the representative point d_i , which is measured by the cosine of the vector space corresponding to the two sample points. When $y(X, c_j) = \max_{j=1,2,\ldots,k}(y(X, c_j))$, the sample X belongs to c_j .

C. Support vector machine (SVM)

Support vector machine (SVM) is a statistical learning theory proposed by Vapnik, which has many unique advantages in solving small sample, nonlinear and high dimensional pattern recognition problems. The core idea of the SVM is to find an optimal classification hyperplane that satisfies the classification requirements. The hyperplane should maximize the classification interval while ensuring the classification accuracy. The algorithm is as follows:

Suppose there are training samples (x_i, y_i), $i = 1,2,\ldots,m$, $x \in R^n$, $y \in \{\pm 1\}$, the hyperplane is $(w \cdot x + k = 0)$. The classification plane needs to satisfy $y_i[(w \cdot x) + k] \ge 1], i = 1,2,\ldots,m$. Thus the classification interval is $2/\|\omega\|$, so the problem of constructing the optimal hyperplane is transformed into the following minimum problem with constraints :

$$\min \Phi(\omega) = \frac{1}{2}\|\omega\|^2 = \frac{1}{2}(\omega' \cdot \omega) \tag{13}$$

Lagrange function:

$$L = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{m} g_i y_i (\omega \cdot x_i + k) + \sum_{i=1}^{m} g_i \tag{14}$$

Where $a_i \geq 0$ is the coefficient of Lagrange, the constrained optimization problem is determined by the saddle point of the Lagrange function, and the solution of the optimization problem satisfies that the partial derivative of ω and k is 0 at the saddle point. The QP problem is transformed into the following dual problem $(a = (a_1, a_2, \ldots, a_l))$:

$$\max Q(g) = \sum_{j=1}^{m} g_j - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} g_i g_j y_i y_j (x_i x_j) \tag{15}$$

$$s.t. = \sum_{j=1}^{m} g_j y_j = 0, g_j \geq 0, j = 1,2,\ldots,m. \tag{16}$$

After calculation, the optimal weight vector ω^* and the optimal offset k^* are respectively:

$$\omega^* = \sum_{j=1}^{m} g_j^* y_j x_j \tag{17}$$

$$k^* = y_i - \sum_{j=1}^{m} y_j g_j^* (x_j \cdot x_i) \tag{18}$$

Where $j \in \{j | g_j^* > 0)$. Therefore, the optimal classification plane is $(w^* \cdot x + k^* = 0)$, the optimal classification function is $f(x) = \mathrm{sgn}\{(\omega^* \cdot x) + k^*\} = \mathrm{sgn}\left\{\left(g_j^* y_j (x_j \cdot x_i)\right) + k^*\right\}$, where $x \in R^n$.

D. The framework of the ensemble classifier

The ensemble classifier integrates several individual classifiers, and determines the final results by by combining the classification results of multiple classifiers to achieve better performance than a single classifier. Suppose there are K training samples with high dimensional input, which the set is represented as :

$S' = \{(x_{ij}, y_i) | i = 1,2,\ldots,K; j = 1,2,\ldots,J\}$, the input of each training sample contains J attributes, y_i is the label of the sample. For the training samples S' , the modeling process based on ensemble classifier is as follows:

Step 1: Construct the balanced data set. Using the SMOTE method in section 2.2 to to increase the sparse class. After sparse class samples increased, a new set of samples $S = \{(x_{ij}, y_i) | i = 1,2,\ldots,M; j = 1,2,\ldots,J\}$ are obtained, where M samples are training samples of the ensemble classifier.

Step 2: The training set S as the input of the ensemble classifier. The contribution of each feature is different due to the different description and the non-uniform dimension. In order to reduce the difficulty of training, we normalize the data set before training by using the following conversion formula:

$$x_{ij} = \frac{x_j^{\max} - x_{ij}}{x_j^{\max} - x_j^{\min}} \tag{19}$$

if $x_j^{\max} = x_j^{\min}$, $x_{ij} = -1$, where in (19) $x_j^{\max} = \max\{x_{1j}, x_{2j}, \ldots, x_{Mj}\}$, $x_j^{\min} = \min\{x_{1j}, x_{2j}, \ldots, x_{Mj}\}$, $i = 1,2,\ldots,M$, $j = 1,2,\ldots,J$.

Step 3: Training process. We establish the ensemble classifier model by using (1) - (10) to train the MFBC model, using (11) - (12) to train the KNN model, and using (13) - (18) to train the SVM model, respectively.

Step 4: Generalization process. Using a set of data sample $\hat{S} = \left\{(\hat{x}_{ij}, \hat{y}_i) | i = 1,2,\ldots,T; j = 1,2,\ldots,J\right\}$ which is different from the training samples as a generalization sample of the ensemble classifier model. $y1_i$ is the MFBC model classification result, $y2_i$ is the KNN model classification result, $y3_i$ is the SVM model classification result, we vote on the above three results as the result of the ensemble classifier model. We compare the generalization output y_i'' and the expected output $\hat{y}_i$, then calculate the generalization relative error.

The structure of water flooded layer recognition model is shown in Fig.3.

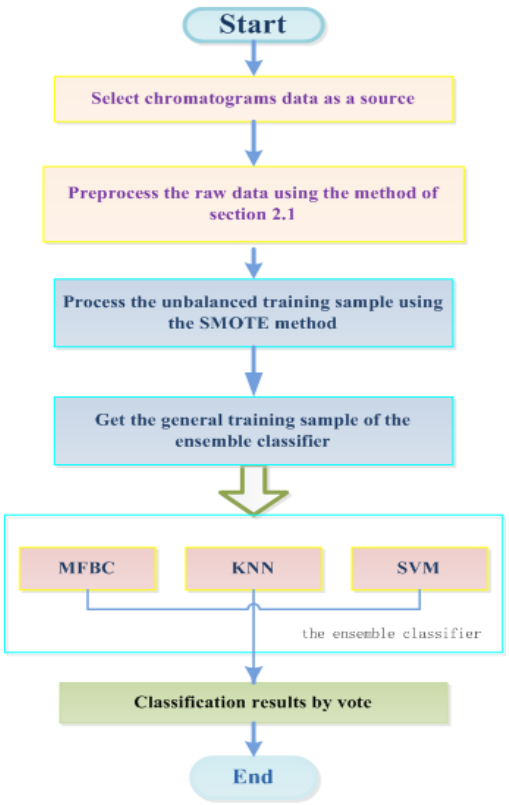

Fig. 3. The structure of the water flooded layer recognition model.

IV. ALGORITHM COMPARISON BASED ON STANDARD DATA SETS

In order to validate the effectiveness and robustness of the ensemble classifier model, this paper uses the breast cancer data set from UCI repository. This data set contains 9 attributes. We select 118 sets of data as training data, the other 20 sets of data as test data. We build four models: the MFBC, the KNN, the SVM and the ensemble classifier. The prediction accuracy of each model are shown in Table 3.

TABLE III. COMPARISON OF EACH MODE BY UCI DATA

Item	*SVM*	*KNN*	*MFBC*	*ensemble classifier*
Classification accuracy	90%	90%	90%	95%

It can be seen from Table 3 that the ensemble classifier model is better than the SVM, KNN, MFBC model for Breast Cancer data set. The prediction accuracy has been a certain increase by integrating three classifiers, its ability to promote the advantages of the more obvious. However, the computing time of the ensemble classifier model is longer than three other single classifiers. According to the Table 3, the effectiveness of the ensemble classifier model is validated.

We select 170 chromatograms to establish model. Firstly, we preprocess these raw chromatograms, and select 12 indicators as the input items of classifiers as shown in Table 1. Then, we extend the oil layer data samples by using the SMOTE algorithm as shown in Table 2. After processing by the SMOTE algorithm, the number of samples reached 206. We use the 10-fold cross validation method to test the accuracy of the classification model. The data set is divided into ten parts, in turn, nine of them as training data and the remaining one as test data. The correct rate is obtained for each test. The average of the correct rate of the 10 results is used as an estimate of the accuracy of the classification model. The comparison of different classification results and real results are shown in Table 4, and the prediction accuracy of each model by the 10-fold cross validation method are shown in Table 5.

TABLE IV. RESULTS OF THE DIFFERENT CLASSIFICATION MODEL

ID	*SVM*	*KNN*	*MFBC*	*ensemble classifier*	*Real result*
1	oil	oil	oil	oil	oil
2	oil	oil	oil	oil	oil
3	oil	oil	oil	oil	oil
4	oil	oil	oil	oil	oil
5	strong	oil	strong	strong	strong
6	strong	oil	strong	strong	strong
7	middle	strong	strong	strong	strong
8	weak	strong	weak	weak	weak
9	weak	weak	weak	weak	weak
10	weak	weak	middle	weak	weak
11	oil	oil	oil	oil	weak
12	weak	weak	weak	weak	weak

TABLE V. COMPARISON OF DIFFERENT MODEL BY CHROMATOGRAMS DATA

Item	*SVM*	*KNN*	*MFBC*	*ensemble classifier*
Classification accuracy	70.59 %	70.59 %	70.59 %	82.35 %

As we can see, the classification results of the ensemble classifier are more close to the actual value than SVM, KNN and MFBC. The classification accuracy of the ensemble classifier is 82.35%, however, the classification accuracy of SVM, KNN and MFBC are 70.59% respectively.

We also select the chromatogram data of the other five wells to make water flooded layer recognition. After the method in section 2 preprocessed, then we use the model in

section 3, 4 to classify and calculate the classification accuracy. The comparison of classification accuracy between the ensemble classifier model and three other classifiers model are shown in Fig.4.

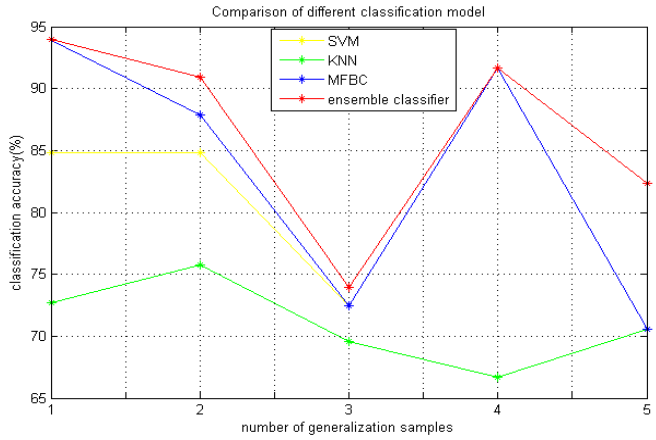

Fig. 4. Comparison of different classifiers model

It can be seen from Fig.4, the robustness of the ensemble classifier is better than that of a single classifier. Therefore, when we use the ensemble classifier to identify categories of water flooded layer, we can predict more accurate, more stable and with less error. With the recognition model, we can provide the reasonable exploration data so as to evaluate the characteristics of water flooded reservoirs accurately and it is of importance to the well site accurate evaluation.

CONCLUSIONS

This paper proposes an ensemble classifier approach that could help to build a recognition model of water flooded chromatogram data. The ensemble classifier approach solves the lag problem of the traditional quantitative data analysis. Moreover, in terms of accuracy, the robustness and effectiveness of the ensemble classifier model is better than the single classifier model through the standard data set from the UCI repository. Finally, the ensemble classifier approach is applied to classify the water flooded chromatogram data of well logging. The method successfully provides information about water flooded layers, and the well logging departments could utilize these results to effectively improve the reservoir development efficiency.

REFERENCES

[1] Jia, Chengzao, Y. Zhang, and X. Zhao. Prospects of and challenges to natural gas industry development in China. Natural Gas Industry B 1.1(2014):1-13.

[2] Zhang, F., She, Y., Chai, L., Banat, I. M., Zhang, X., Shu, F., & Hou, D. (2012). Microbial diversity in long-term water-flooded oil reservoirs with different in situ temperatures in China. Scientific Reports, 760-760.

[3] Zhang, M. L., Sun, B. G., Xie, L., & Hou, S. S. (2011). Water flooded layer identification of thick oil reservoir of fuyang oil layer with calcareous intercalation. Journal of Daqing Petroleum Institute.

[4] Hernandeztorres, J., Villamizar, G. A., Salgarchaparro, S. J., Silvaplata, B. A., Daza, O. D., Martinezperez, F., ... & Fuentes, J. L. (2015). Prokaryotic Community Characterization in a Mesothermic and Water- Flooded Oil Reservoir in Colombia. Geomicrobiology Journal, 33(2), 110-117.

[5] Aleardi, M. (2015). Seismic velocity estimation from well log data with genetic algorithms in comparison to neural networks and multilinear approaches. Journal of Applied Geophysics, 117, 13-22.

[6] Jana, S., Ojha, M., & Sain, K. (2015). Gas hydrate saturation from heterogeneous model constructed from well log in krishna−godavari basin, eastern indian offshore. Geophysical Journal International, 203(1), 246-256.

[7] Blagus, R., & Lusa, L. (2013). Smote for high-dimensional class-imbalanced data. BMC Bioinformatics, 14(1), 106.

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16(1), 321-357.

[9] Yuan, G., Lu, W., & Hao, H. Z. (2014). A model-free machine learning method for risk classification and survival probability prediction. Stat, 3(1), 337-350.

[10] Wang, Y., & Pedram, M. (2016). Model-free reinforcement learning and bayesian classification in system-level power management. IEEE Transactions on Computers, 65(12), 3713-3726.

[11] Hassanat, A. B., Abbadi, M. A., Altarawneh, G. A., & Alhasanat, A. A. (2014). Solving the problem of the k parameter in the knn classifier using an ensemble learning approach. Computer Science, 12(8), 33-39.

[12] Buana, P. W., Jannet, D. R., & Putra, I. K. (2012). Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News. International Journal of Computer Applications, 50(11), 37-42.

[13] Xu, P., Davoine, F., Zha, H., & Denoeux, T. (2015). Evidential calibration of binary svm classifiers. International Journal of Approximate Reasoning, 72(C), 55-70.

[14] Paul, S., Magdon-Ismail, M., & Drineas, P. (2016). Feature selection for linear svm with provable guarantees. Pattern Recognition, 60, 205-214.

[15] Cárdenas-Gallo, I., Sarmiento, C. A., Morales, G. A., Bolivar, M. A., & Akhavan-Tabatabaei, R. (2017). An ensemble classifier to predict track geometry degradation. Reliability Engineering & System Safety, 161, 53-60.

[16] Rahman, A., & Tasnim, S. (2014). Ensemble classifiers and their applications: a review. International Journal of Computer Trends & Technology, 10(1).

[17] Rokach, L. (2010). Ensemble-based classifiers. Artificial Intelligence Review, 33(1), 1-39.

[18] Mason, J. P., Kirk, I., Windsor, C. G., Tipler, A., Spragg, R. A., & Rendle, M. (2015). A novel algorithm for chromatogram matching in qualitative analysis. Journal of Separation Science, 15(8), 539-547.

[19] Hu, J., & Sheng, J. J. (2016). Numerical modeling on air injection in a light oil reservoir: recovery mechanism and scheme optimization. Fuel, 172, 70-80.

[20] Victor, J. D. (2002). Binless strategies for estimation of information from neural data.. Physical Review E, 66(5).