TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 10/12 pp281–293
DOI: 10.26599/TST.2019.9010020
Volume 25, Number 2, April 2020

Valuable Data Extraction for Resistivity Imaging Logging Interpretation

Yili Ren*, Renbin Gong, Zhou Feng, and Meichao Li

Abstract: Imaging logging has become a popular means of well logging because it can visually represent the lithologic and structural characteristics of strata. The manual interpretation of imaging logging is affected by the limitations of the naked eye and experiential factors. As a result, manual interpretation accuracy is low. Therefore, it is highly useful to develop effective automatic imaging logging interpretation by machine learning. Resistivity imaging logging is the most widely used technology for imaging logging. In this paper, we propose an automatic extraction procedure for the geological features in resistivity imaging logging images. This procedure is based on machine learning and achieves good results in practical applications. Acknowledging that the existence of valueless data significantly affects the recognition effect, we propose three strategies for the identification of valueless data based on binary classification. We compare the effect of the three strategies both on an experimental dataset and in a production environment, and find that the merging method is the best performing of the three strategies. It effectively identifies the valueless data in the well logging images, thus significantly improving the automatic recognition effect of geological features in resistivity logging images.

Key words: machine learning; binary classification; multiclass classification; outlier detection; imaging logging

1 Introduction

Well logging is one of the most important means of identifying and evaluating oil and gas reservoirs in the process of oil exploration and development. It is also an important tool for helping to solve a number of geological problems. Mostly drawing on the principles of sound, electricity, nuclear, and magnetic fields, well logging also uses a large number of new materials, crafts, and technologies to measure dynamic oil and gas information. It is usually carried out in high temperature and high pressure wellbore environments. Logging technology is developing rapidly in response to the demand for oil exploration and development. Open hole logging technologies, such as microresistivity imaging logging, cross-dipole acoustic logging, nuclear magnetism logging, and logging while drilling, provide powerful measures for studying the anisotropy of strata. The use of imaging logging is particularly extensive in geology due to its high resolution, coverage, and visibility.

Imaging logging is a method for the physical parameter imaging of a borehole wall and surrounding objects based on the observation of the geophysical field in the borehole. Imaging logging includes borehole imaging, side-hole imaging, and cross-hole imaging technologies. Borehole imaging logging technology is the most mature, including borehole acoustic imaging and formation microresistivity scanning imaging. Side-hole imaging mainly refers to resistivity imaging[1], and uses the azimuth lateral logging and array induction logging methods. Cross-hole imaging, including

• Yili Ren, Renbin Gong, and Meichao Li are with the Institute of Computer Application Technology, PetroChina Research Institute of Petroleum Exploration and Development (RIPED), Beijing 100083, China. E-mail: renyili@petrochina.com.cn; gongrb@petrochina.com.cn; limeichao@petrochina.com.cn.
• Zhou Feng is with the Department of Well Logging & Remote Sensing Technology of RIPED, Beijing 100083, China. E-mail: fz@petrochina.com.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2018-12-22; revised: 2019-04-17; accepted: 2019-05-14

acoustic[2–4], electromagnetic, and resistivity imaging, has been widely used in engineering exploration, and has obtained some successful results in petroleum exploration.

Resistivity imaging logging is the most widely used of all the imaging logging methods. It can accurately describe the sedimentary characteristics of reservoirs, and can clearly reflect shale, clastic, calcite crystal, porosity, fracture, suture line, bedding, and biological disturbance. It is also of interest for lithology and sedimentary facies research[5–13]. The structural characteristics of fractures, dissolution cavities, faults，and folds can also be identified by resistivity imaging logging, and it is an effective tool for reservoir identification, well-side structure, and in-situ stress analysis[14, 15]. In addition, the response characteristics of resistivity imaging logging to dissolution voids and fractures are obvious. Under the constraints of core data and conventional curves, it can calculate porosity, permeability, and other parameters quantitatively[16–18].

Because of the huge amount of information contained in resistivity imaging logging images, the imaging interpretation must be carried out at the scale of 1:10 or 1:20, if it is to accurately identify the structural components and sedimentary structures of the strata. This makes manual interpretation highly time-consuming and inefficient. Furthermore, the accuracy of manual interpretation is limited by the restrictions of the naked eye and by empirical factors. Therefore, the automatic and intelligent processing of resistivity imaging logging images is of great benefit. Identifying the best methods to accurately classify, identify, and quantitatively extract all kinds of geological features from resistivity imaging logging images is the key to realizing this goal.

There are many challenges in the process of automatic recognition of geological features in resistivity imaging logging images. In particular, the proportion of geological features that are to be extracted is small in comparison with the amount of background and valueless data. The existence of this valueless data significantly affects the recognition effect. After extracting all the valuable data from resistivity logging images, it is relatively simple to classify these valuable data correctly. The extraction of valuable data is therefore a key part of the automatic recognition process of imaging logging interpretation.

In this paper, we propose a framework for the automatic extraction of geological features from resistivity imaging logging images. This framework includes image processing, image segmentation, feature extraction, and valuable data extraction and classification. Our experimental results show that this proposed framework is effective in extracting geological features from resistivity logging images.

Valuable data extraction is a crucial aspect of the framework. Valuable data extraction is similar to the multiclass classification problem, with one exception. In the multiclass classification problem, the training data set contains a number of categories, and each element of the test data belongs to one of these categories; in the valuable data extraction problem, some of the test data do not belong to any of the predefined categories of the training data, but they are not to be treated as outliers.

Many methods are available to solve the multiclass classification problem. Li et al.[19] presented a unifying framework for studying the solution of multiclass categorization problems by reducing them to multiple binary problems that are then solved using a margin-based binary learning algorithm. Goh et al.[20] used voting and combinations of approximate posterior probabilities to show the possibilities of simple generalizations of the binary classification. Hastie and Tibshirani[21] discussed a strategy for multiclass classification that involves estimating class probabilities for each pair of classes, and then coupling these estimates together. Dietterich and Bakiri[22] put forward the Error-Correcting Output Codes (ECOC) and studied code design techniques suitable for solving multiclass problems. Kong and Dietterich[23] carried out an investigation of the reasons why the ECOC method can improve the classification accuracy of multiclass classification.

Although there are many strategies and algorithms to solve multiclass classification problems, there are no effective methods for valuable data extraction. Clustering and outlier detection are the most common methods for solving this problem. Hodge and Austin[24] provided a broad overview of outlier detection, for which Local Outlier Factor (LOF)[25, 26] is a widely used algorithm. Li and Wong[27] and Hardin and Rocke[28] introduced a statistical method for the relationship between clustering and outlier detection, which has often been considered complimentary.

In this paper, we put forward three strategies, all based on binary classification, to solve the problem

of valuable data extraction: the intersection method, elimination method, and merging method. We then verify each of the three strategies with experimental data and in a production environment, with the results showing that the merging method is the best of the three.

In summary, the major contributions of this study are as follows.

• We propose a procedure for the automatic interpretation of resistivity imaging logging images. Since valueless data has a great influence on the effect of automatic interpretation, we include a valuable data extraction stage in the procedure.

• We put forward three strategies based on binary classification for valuable data extraction: the intersection method, elimination method, and merging method. These strategies can effectively solve the problem of valueless data in multiclass classification.

• We compare the effectiveness of these three strategies both using experimental data and in a production environment. The results show that the merging method performs best in effectively removing the valueless data from logging images.

The remainder of the paper is organized as follows. Section 2 gives the definition and examples of valuable data and valueless data, and defines the valuable data extraction problem. In Section 3, we propose three valuable data extraction strategies based on binary classification. Section 4 proposes a framework for the automatic extraction of geological features from resistivity imaging logging images. Section 5 presents our experimental setup and results.

2 Valuable Data Extraction Problem

2.1 Definition of valuable data and valueless data

Imaging logging interpretation aims to extract features with geological significance, such as dissolution pores and algal laminae. However, these significant features account for a very small proportion of the whole image, with most of the image contents made up of background and features without geological interpretation significance. After image segmentation, the background and most of the unexplained features can be removed, but there remain many redundant features. We call the data with geological significance valuable data and the remainder valueless data. The main difference between valuable data and valueless data is that valuable data belong to a category that is predefined in the training data while valueless data do not. In logging images, each kind of valuable data has its own rules, which we elaborate on in Section 2.2 below.

From the perspective of machine learning, data value can be defined as follows. In the classification task of machine learning, it is assumed that the distributions of training data and test data are consistent. Assuming that the training data set contains n categories, all data in the test data set can be classified into n categories through the classification model. However, in practical problems, test datasets often contain data that does not fit any of the n categories. We call data belonging to the n predefined categories of training data set valuable data, and data outside the n categories valueless data. The main reason for the existence of valueless data in the test data is inconsistency in the data distribution between training and test data.

Valuable data: Valuable data refers to the elements that belong to one of the categories predefined in the training data. Each kind of valuable data has its own rules and they can be assigned to their true class by a classifier.

Valueless data: Valueless data refers to the elements that do not belong to any of the categories predefined in the training data. They have no rules, and therefore a classifier is unable to assign them to a class.

2.2 Examples of valuable data and valueless data

Logging professionals are only interested in valuable data. There are five main types of valuable data in resistivity logging images: dissolution pores, algal laminae, thick mudstone, clay band, and induced fracture. Each of these types of data has its own graphic features that distinguish it from the others. Following is a description of the five types of data, along with sample images.

Dissolution pores have a different appearance from the other four types of data, as shown in Fig. 1.

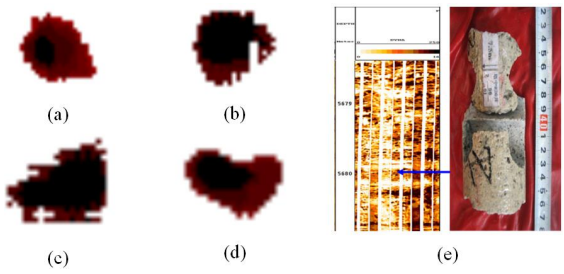


Fig. 1 Sample pictures of dissolution pores are shown from (a) to (d), (e) shows the coring result of dissolution pores.

Algal laminae, clay band, and thick mudstone share certain similarities. All are cross bar, although they have different thicknesses. Thick mudstone is the thickest and algal laminae is the least thick, as shown in Fig. 2.

Figure 3 shows the difference between algal laminae, clay band, and thick mudstone. Each image is shown in the same proportion, so it can be seen that these three types of geological features have different thicknesses.

Induced fracture is another common geological feature. Some sample images of induced fractures are shown in Fig. 4.

Valueless data are the geological features that are not of interest, and therefore are not to be extracted from resistivity logging images. They have no specific rules, and their shapes are irregular and random, but they do have a different appearance than valuable data. Figure 5 shows some sample images of valueless data, although it is far from exhaustive.

2.3 Problem definition

In the process of automatic extraction of geological features for resistivity imaging logging, the attribute values of geological features can be extracted after image segmentation. Therefore, we can identify every type of geological feature using a classifier. But the problem remains that there is still a large amount of valueless data after image segmentation.

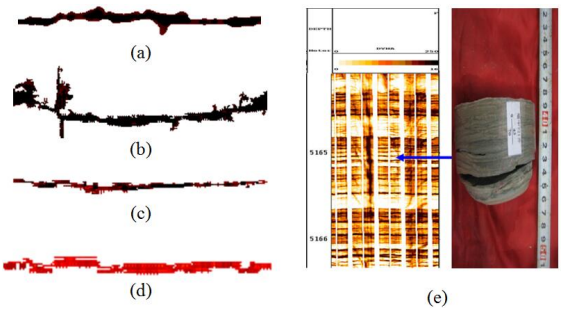


Fig. 2 Sample pictures of algal laminae are shown from (a) to (d), (e) shows the coring result of dissolution pores.

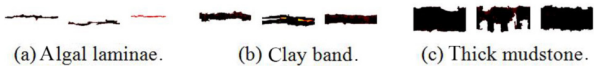


Fig. 3 Comparison pictures of algal laminae, clay band, and thick mudstone. These three are very similar. The difference between them is the thickness.

Fig. 4 Sample pictures of induced fracture.

Fig. 5 Sample pictures of valueless data.

Assuming that there are n categories in the training data and their class labels are $T = \{y_1, y_2, \ldots, y_n\}$, each element of the training dataset will have a class label y_i satisfying the following conditions $y_i \in T$. However, in the test dataset there is at least one element with a class label $y_i \notin T$; these elements make up the valueless data. A classifier will assign all of the test data a predicted class label y_t, which is predefined in the training data, such that $y_t \in T$. Valueless data will thus be misclassified by the classifier, so the elimination of valueless data before classification is an important and difficult problem.

3 Valuable Data Extraction Based on Binary Classification

Valuable data extraction is very similar to the multiclass classification and outlier detection problem, but not the same. The difference between valuable data extraction and multiclass classification is that valueless data do not belong to any predefined class in the training data, whereas for multiclass classification, all data in the test dataset belong to predefined categories. The difference between valuable data extraction and outlier detection is that outliers are small and scattered, and usually do not belong to a specific category, whereas valuable data usually belong to a specific category, but this category is not defined in the training data.

3.1 Multiclass problem and binary classification

When the input data is divided into two categories, the problem becomes a binary classification problem, which some classification techniques, such as Support Vector Machine (SVM) and AdaBoost, were originally designed for. When there are more than two categories,

it becomes a multiclass problem. There are several approaches for extending the binary classifier to handle multiclass problems. For the following, let $Y = \{y_1, y_2, \ldots, y_k\}$ be the set of classes in the input data.

• **One-against-rest (1−*r*) approach**

$1-r$ decomposes the multiclass problem into K binary problems[29]. For each class $y_i \in Y$, all of the instances that belong to y_i are considered positive samples and the remaining instances are considered negative samples. A binary classifier is then used to identify y_i and $\neg y_i$ (where $\neg y_i$ refers to these objects that do not belong to y_i).

• **One-against-one (1−1) approach**

$1-1$ constructs $C_k^2 = K(K-1)/2$ classifiers[29], each of which is used to distinguish between y_i and y_j. When constructing a classifier for y_i and y_j, all of the samples that do not belong to either y_i or y_j are ignored.

• **Error-Correcting Output Codes (ECOC)**

The $1-r$ and $1-1$ approaches are sensitive to the binary classification errors. ECOC provides a more robust method for dealing with multiclass problems[22,23]. For multiclass learning, each class y_i is represented by a unique bit string with a length of n, which becomes its code word. n binary classifiers are then trained to predict each binary of the code string.

We cannot simply apply multiclass strategies to our current problem, because the valueless data is not a class. Instead, we put forward three strategies for valuable data extraction that draw on the idea of the multiclass problem. These strategies are based on binary classification.

3.2 Strategies for valuable data extraction based on binary classification

Suppose that there are n categories of valuable data in the training dataset D. Let $T = \{y_1, y_2, \ldots, y_n\}$ be the set of class labels of valuable data. Our goal is to identify all the valueless data, to which we give a class label of y_0, so that we can eliminate them from the logging images. Following is a description of three strategies.

• **Strategy I: Intersection method**

The intersection method is very similar to the $1-r$ approach. We construct n binary classifiers, each of which divides the dataset D into two categories: y_i and $\neg y_i$ (where y_i refers to all the objects that belong to y_i, and $\neg y_i$ refers to all the objects that do not belong to y_i). We thereby identify $\neg y_1, \neg y_2, \ldots, \neg y_n$ once at a time. The set $I = \neg y_1 \cap \neg y_2 \cap \cdots \cap \neg y_n$ is the valueless data, and all the elements that belong to I are given a class label of y_0. The flowchart is shown in Fig. 6.

• **Strategy II: Elimination method**

We again construct n binary classifiers to identify the valuable data $y_1, y_2, \ldots, y_n$ in order, but the input dataset is constantly changing. For each classifier, $D_i = D_{i-1} \setminus D_{y_i-1}$, where D_{i-1} is the input data of the last classifier and D_{y_i-1} is the set of all valuable data y_{i-1} identified by the last classifier. Set $I = \neg y_n$ is the valueless data, and all the elements that belong to I are given a class label of y_0. The flowchart is shown in Fig. 7.

• **Strategy III: Merging method**

For the merging method, we regard all of the valuable data as a single class and use one binary classification to identify the valuable data y_e and the valueless data $\neg y_e$. Set $I = \neg y_e$ is valueless data, and all of the elements that belong to I are given a class label of y_0. The flowchart is as shown in Fig. 8.

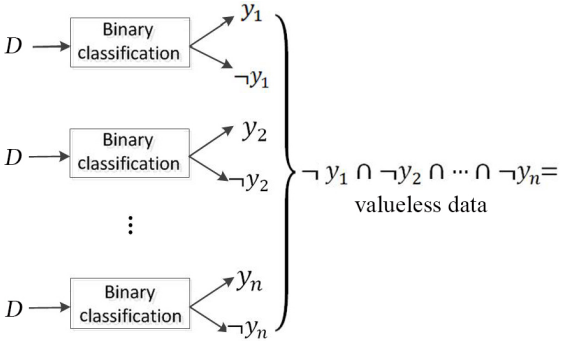


Fig. 6 Flowchart of intersection method.

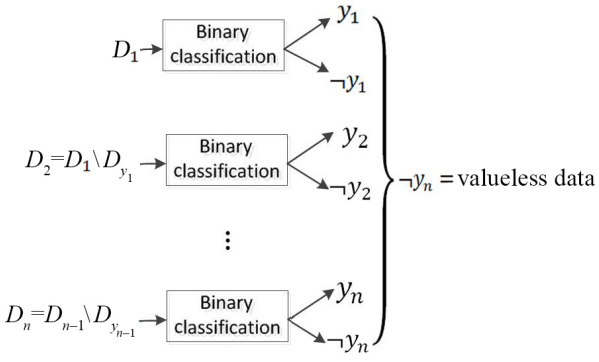


Fig. 7 Flowchart of elimination method, D_1 is equivalent to D in Figs. 6 and 8.

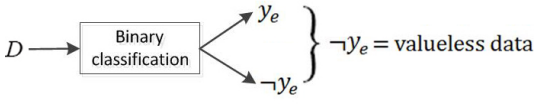


Fig. 8 Flowchart of merging method.

4 Automatic Extraction of Geological Features for Resistivity Imaging Logging

Traditional imaging logging interpretation depends on human expertise and experience, as people use the naked eye to identify geological features in imaging logging images. This manual method is inefficient and its accuracy is affected by human factors.

Machine learning technology has been applied to the intelligent interpretation of conventional logging curves, but as yet the results have not been ideal. A very important factor is that the existence of valueless data severely affects the classification results.

We establish a procedure for geological features extraction in resistivity imaging logging. This procedure is based on machine learning and is made up of visualization, image segmentation, feature extraction, and valuable data extraction and classification. The input data are resistivity imaging logging images and the output data are geological characteristics with interpretations. The procedure is shown in Fig. 9.

Compared to the traditional interpretation process, the automatic interpretation procedure proposed in this paper has the following advantages:

- It can reduce the workload of logging interpreters;
- It guarantees the stability of interpretation accuracy; and
- It improves the automation and intelligence level of imaging logging interpretation.

4.1 Image processing

The image processing stage is made up of data reconciliation, image representation, and image enhancement. Ideally, the drilling tool moves at a uniform speed; when the tool becomes slightly stuck in the borehole, the depth of logging records will deviate from the true measurement depth. Therefore, velocity correction must be carried out first. Image representation is the process of mapping the original data acquired by imaging logging into color or grayscale images. Image enhancement is mainly achieved through histogram equalization and other methods to achieve prominent features and eliminate noise.

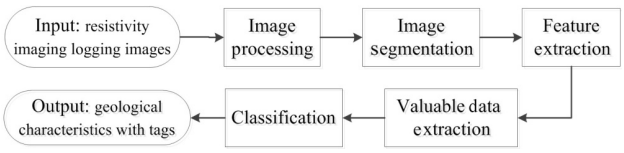


Fig. 9 Automatic extraction procedure of geological features for image well logging.

4.2 Image segmentation

Imaging data can reflect geological features precisely, and the purpose of image segmentation is to separate these features from the background. Feature segmentation is the basis for subsequent analysis. Firstly, the image is binarized to separate the features from the background, and then each region with a value of 1 is marked separately and recorded as a single feature. The simplest and most widely used binarization method is to set a threshold, which can be determined by a histogram of imaging data.

When the difference between the feature and background is obvious, image segmentation using the threshold method can achieve ideal results. However, if the feature and background are similar, the threshold method cannot distinguish them effectively. Therefore, we use K-means for image segmentation.

K-means is a typical clustering algorithm. First, it chooses K objects as initial centers. The distance from each object to the initial center is then calculated, and each object is assigned to the nearest cluster center. A cluster center and the objects assigned to it represents a cluster. The centroid of each cluster is updated according to the points assigned to the cluster. K-means will repeat the assignment and update steps until the cluster or the centroid does not change. Its goal is to minimize the Sum of Squared Errors (SSE), defined as

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \mathrm{dist}(c_i, x)^2,$$

where c_i represents a centroid of the cluster C_i, and K is the number of clusters. When K-means is used for image segmentation, each pixel is assigned to a different cluster. For each cluster, the value of the centroid then replaces the value of all the pixels in the cluster. Figure 10 shows the effect of several different segmentation methods, from which it can be seen that the clustering-based image segmentation algorithm is superior. It is not only able to separate features from background, but also to retain as many features as possible.

4.3 Feature extraction

After separating the features from the background, it is necessary to quantitatively analyze these features. This process is called feature extraction. The purpose of feature extraction is to obtain the attributes necessary for classification, which are of two types: shape and

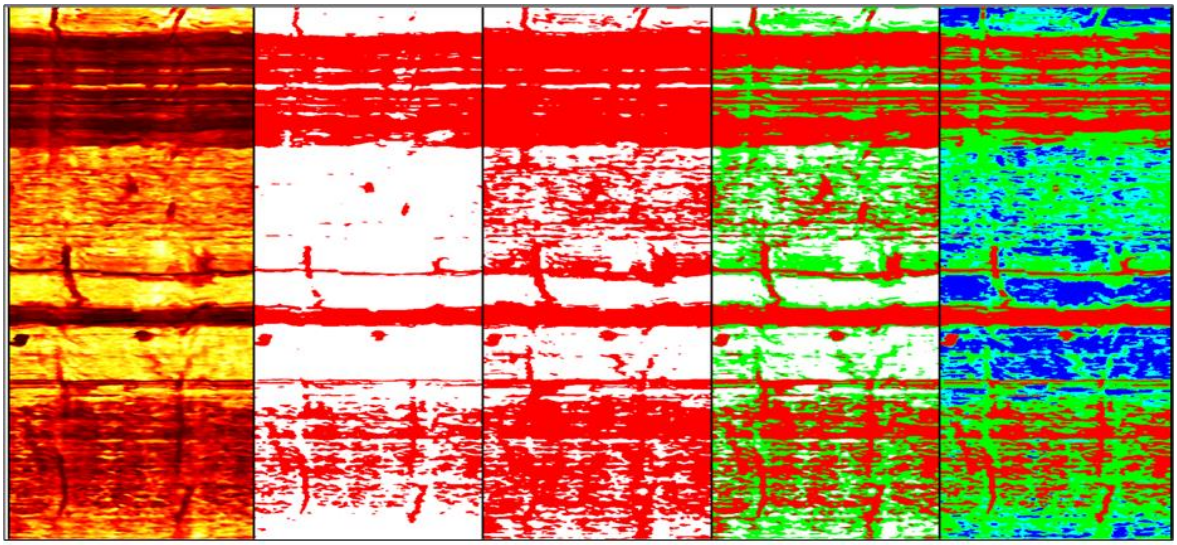


Fig. 10 Effect of several different segmentation methods, (a) represents original image, (b) – (e) are the segmentation effects of threshold, adaptive threshold, adaptive threshold (multi-objective), and clustering method, respectively.

texture. Each feature has various feature parameters that are important in distinguishing them. Figure 11 shows several common features in reef reservoir images. For different features, the importance of various feature parameters is also different. For example, we must consider the general distribution law to distinguish dissolution pores from muddy mass. But for algal lamina and clay band, thickness is the most important index.

It is difficult to judge the dissolution pores and muddy mass on the basis of individual characteristics, so the general distribution law must be taken into account. The morphology of algal lamina is similar to that of clay band, and thickness is an important parameter for distinguishing them. Sawtooth is a typical feature of suture and can be measured by curvature. For cracks, tendency and inclination are most important. Other descriptors of shape characteristics include appearance ratio, eccentricity, and sphericity.

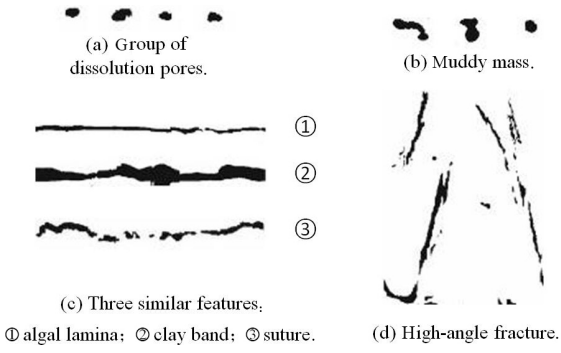


Fig. 11 Common features of reef beach reservoir images.

In addition to the above shape features, the texture features of images are also important. The gray level co-occurrence matrix is an effective method for texture feature extraction[30,31]. Based on the gray level co-occurrence matrix, texture feature parameters, such as second moment, entropy, and uniformity, can be calculated.

For an image S, if the function $f(x, y)$ defines some spatial relationship, then the elements in the gray level co-occurrence matrix P of S are defined as follows[30–32]:

$$P(g_1, g_2) = \#\{[(x_1, y_1), (x_2, y_2)] \in S | f(x_1, y_1) = g_1 \& f(x_2, y_2) = g_2\}/\#S,$$

where the molecule is the number of pairs of elements with spatial relations $f(x, y)$ and values g_1 and g_2, the denominator is the total number of pairs of elements in S (# represents the number).

4.4 Valuable data extraction

After image segmentation, all geological features are segmented from the background of the image. But there will remain some valueless data and outliers, and these will severely affect the classification effectiveness. There are two steps to extracting the valuable data: first, we identify all of the outliers and eliminate them; second, we extract the valuable data.

We use LOF to identify the outliers. Traditional distance-based noise processing algorithms are easily affected by density, and LOF presents a good solution to this problem. The basic principles of the LOF algorithm are as follows.

(1) The distance between two points p and o is $d(p,o)$.

(2) The definition of the k-distance $d_k(p)$ for point p is as follows. $d_k(p)$ equals $d(p,o)$ and meets the following conditions. First, there are at least k points except for p in the set $o' \in C\{x \neq p\}$ for which $d(p,o') \leqslant d(p,o)$. Second, there are up to $k-1$ points except p in the set $o' \in C\{x \neq p\}$ for which $d(p,o') < d(p,o)$. $d_k(p)$ is therefore the distance between p and k point ranked from near to far.

(3) The k-distance neighborhood of p, $N_k(p)$, is a set of points whose distance from p is less than or equal to $d_k(p)$. The number of points in $N_k(p)$ is greater than k: $|N_k(p)| \geqslant k$.

(4) The definition of the reach-distance between o and p is

$$\text{reach-distance}_k(p,o) = \max\{k - \text{distance}(o), d(p,o)\}.$$

(5) This means that reach-distance$_k(p,o)$ equals at least $d_k(o)$, otherwise it equals $d(p,o)$.

(6) The definition of local reachability density for point p is

$$\text{lrd}_k(p) = \frac{|N_k(p)|}{\sum\limits_{o \in N_k(p)} \text{reach-distance}_k(p,o)}.$$

(7) The local outlier factor of p is expressed as

$$\text{LOF}_k(p) = \frac{\sum\limits_{o \in N_k(p)} \frac{\text{lrd}_k(o)}{\text{lrd}_k(p)}}{|N_k(p)|} = \frac{\frac{\sum\limits_{o \in N_k(p)} \text{lrd}_k(o)}{|N_k(p)|}}{\text{lrd}_k(p)}.$$

If this ratio is close to 1, then the density of point p is approximately equal to the density of adjacent points, so p may belong to the same cluster as its adjacent points. If the ratio is less than 1, then the density of p is higher than the density of adjacent points, so p is a dense point. If the ratio is greater than 1, then the density of p is lower than the density of adjacent points, so p is more likely to be an outlier.

The main idea behind LOF is to compare the density of each point p and its neighborhood points to determine whether the point is an outlier. Point p is more likely to be identified as an outlier if its density is lower.

After eliminating all the outliers, it is still impossible to classify all the valuable data correctly because of the existence of valueless data. Valueless data elements are neither outliers nor do they belong to a specific category defined in the training data set. Until now, There has been no good method to eliminate such valueless data.

In this paper, we put forward three strategies for the identification of valueless data. These strategies are based on binary classification and we have introduced in detail in Section 3.2.

4.5 Classification

The last step is classification. After feature extraction, we obtain the attribute values of each geological feature, which can be used as a basis for the automatic classification of geological features. Since the geological data are nonlinear and small samples, we choose SVM as the recommended classifier.

Since the 1990s, and especially in the current century, the use of SVM has been gradually increasing in the natural and social sciences. Guangren Shi found that SVM performs better on geological data with strong nonlinearity and recommended using Multiple Regression Analysis (MRA) for dimensionality reduction and SVM to verify the effects[33].

The performance of SVM depends on model selection, which is mainly related to kernel functions and their parameters. Table 1 shows different kernel functions and their parameters. Different kernel functions and parameters have an important influence on SVM performance[34]. The technique for their selection breaks down into two main categories: data-independent and data-based[35]. The data-independent technique uses the a priori information of the problem rather than statistical information, so it is experimental and probabilistic, whereas the data-based technique draws on the training data. The latter technique has become the more popular of the two.

Because the performance of SVM is influenced by many factors, especially the penalty factor C, kernel functions and their parameters, we use k-cross validation[36,37] and grid search[38–40] to identify the optimal SVM parameters. We compare the

Table 1 Different kernel functions of SVM.

Kernel function	Formula	Parameter	Merit
Linear	$K(x,x_i) = x \cdot x_i$	-	It is only used when the sample is separable in low dimensional space.
Polynomial	$K(x,x_i) = [\gamma^*(x \cdot x_i) + \text{coe}f]^d$	γ, coe$f$, d	Global kernels
Radial Basis Function (RBF)	$K(x,x_i) = \exp(-\gamma^*\|\|x - x_i\|\|^2)$	γ	Good local performance
Sigmoid	$K(x,x_i) = \tanh(\gamma(x \cdot x_i) + \text{coe}f)$	γ, coe$f$	Needs to meet certain conditions

classification performance of different kernel functions in our experiments. The experiment results show that kernel functions are key factors in classification results, and the polynomial function offers the best performance.

• ***k*-cross validation:** The training dataset is split into k separate files of equal size, from which $k-1$ is selected as the training dataset and the remainder as a validation set. The modeling process is repeated k times and the average value of MSE after k iterations is obtained to estimate the expected generalization error.

• **Grid search:** A practical method of searching for data, grid search, is quite suitable for searching multidimensional data from different growing directions. To illustrate the principle, when choosing the RBF kernel function for SVM, two parameters need to be confirmed: the penalty parameter C and the kernel function parameter σ. Grid search chooses the change step C_s for $C \in [C_1, C_2]$ and σ_s for $\sigma \in [\sigma_1, \sigma_2]$. Each pair of parameters (C', σ') is then used for modeling and the best performing pair is selected as the final parameters.

We use the combination of k-cross validation and grid search to optimize parameters to avoid overfitting. First, we use grid search to set the range of parameters. For each pair of parameters, we use k-cross validation to calculate classification accuracy. Second, we further subdivide the grid according to the range of formal parameters to find the most accurate parameters. Finally, we adopt those parameters which have the highest classification accuracy.

5 Experiment and Analysis

5.1 Description of our data

After image segmentation, we arrived at structured data set with 19 attributes. The training data contains 600 instances and 6 categories. Their class labels are y_1, y_2, y_3, y_4, y_5, and y_6. We also collect some valueless data and add these to the training dataset with a class label of y_0. Note that although we add valueless data to the training data set, these are far from exhausting the characteristics of valueless data.

We use t-distributed Stochastic Neighbor Embedding (t-SNE)[41–43] to reduce dimensions and visualize data. t-SNE is a nonlinear dimension reduction algorithm that is very suitable for reducing high-dimensional data to two or three dimensions for visualization.

Figure 12 shows the data structure diagram after dimension reduction by t-SNE. Figures 12a and 12c are the structure diagrams after reduction to two dimensions, and Figs. 12b and 12d are the structure diagrams after the reduction to three dimensions. Figures 12c and 12d are the structure diagrams after transforming the problem into a two-class problem, with all valuable data as one category and valueless data as another.

We can see from the graphs in Fig. 12 that the class overlap is more serious when dealing with multiclass classification. When the problem is transformed into a binary classification problem, the difference between valuable data and valueless data is obvious, with just a few overlaps. When we consider all the valuable data as one category and the valueless data as another, the two are spatially separable.

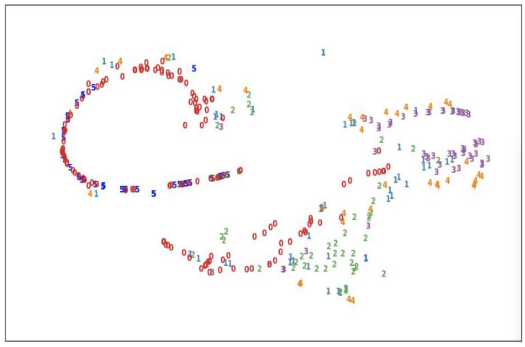
By analyzing the data structure, it is feasible to extract the valuable data using a binary classification method. We identify all the valueless data items and remove them from the resistivity imaging logging images, to ensure that all the data with geological interpretation significance can be identified by the classifier.

5.2 Experimental results

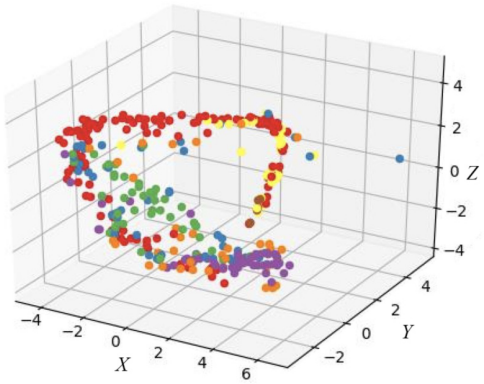
Our experiment verifies the effect of the three strategies using experimental data and in a production environment. We use F-measure, Receiver Operating Characteristic (ROC), and other indicators to evaluate the effect of the three strategies on the experimental data[44], whereas we observe the cleaning effect of the three strategies intuitively in the production environment.

We extracted some experimental data from the production environment. The data volume of the training dataset is 600 and that of test dataset is 400. These experimental data are structured, including 19 attributes: xpos, depth, ar, etc. There are six kinds of valuable data: dissolution pores, algal laminae, thick mudstone, clay band, induced fracture D, and induced fracture S. Among them, induced fracture D and induced fracture S are very similar. We also included some valueless data in the training dataset, but these do not cover all kinds of valueless data. The classifiers we chose were kNN, naivebayes, logistic regression, SVM, and decision tree. We use F-measure, precision, and recall to evaluate the classification performance. The experiment results are shown in Fig. 13.

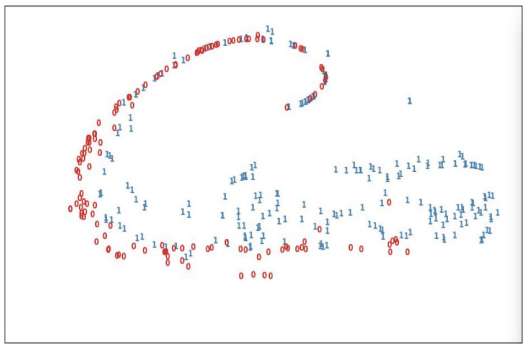
From the experimental results, we can see that



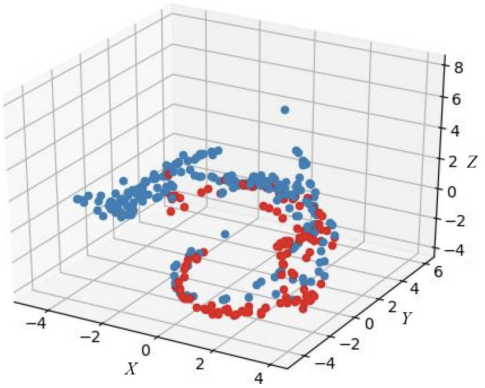
(a) Two-dimensional plan for multiclass.



(b) 3D stereogram for multiclass.

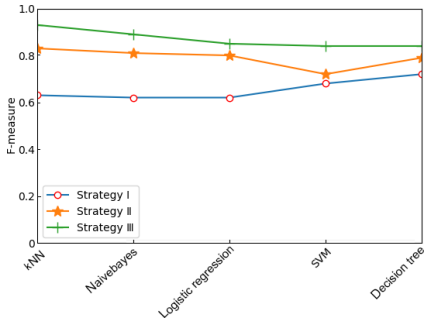


(c) Two-dimensional plan for two-class.

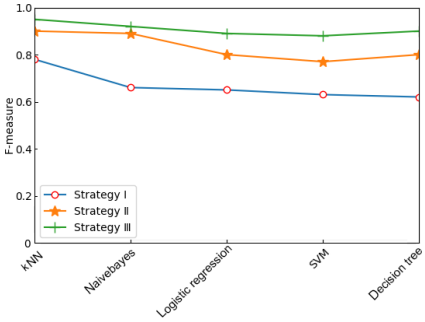


(d) 3D stereogram for two-class.

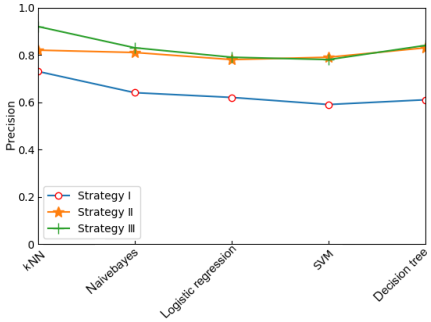
Fig. 12 Data structure diagram after dimension reduction of t-SNE. (a) and (b) are dimensionality reduction effects when they consider problems as multi-class problems. (c) and (d) are dimensionality reduction effects when converting problem into two-class problem.



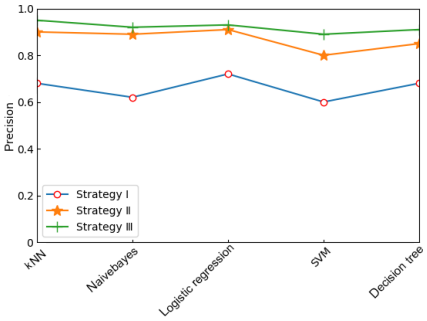
(a) F-measure for valuable data



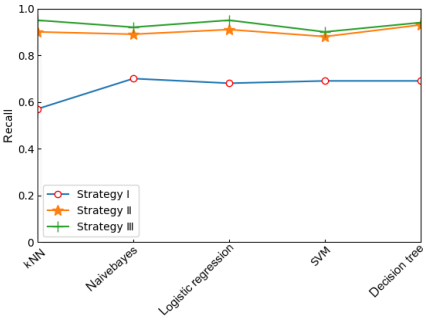
(b) F-measure for valueless data



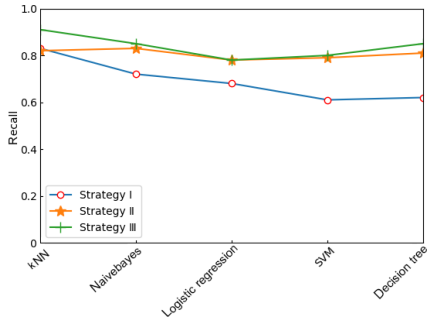
(c) Precision for valuable data



(d) Precision for valueless data



(e) Recall for valuable data



(f) Recall for valueless data

Fig. 13 Classification results on experimental data.

the merging method is the most effective of the three strategies; its F-measure, precision, and recall are higher than the other strategies. Among the five classifiers, kNN offers the best performance.

We encapsulate the three strategies into a program package and integrate them into the logging software CIFlog to verify the effectiveness of the three strategies in the production environment. CIFlog, a well logging interpretation software independently developed by the PetroChina Research Institute of Petroleum Exploration and Development (RIPED), has been widely used in the field of well logging interpretation in China and internationally. We use kNN as the classifier.

Figure 14 shows the valueless data elimination effect on a well in an oilfield. The first column shows the depth, the second shows the resistivity imaging logging image, and the third shows the result of image clustering segmentation.

We use the K-means method for image clustering segmentation. According to the gray value of an image, the K-means method is used to divide the image into five categories: brightest, brighter, intermediate, darker, and darkest. The intermediate, brighter, and darker parts generally represent the response of rock matrix. The brightest and darkest parts are the geological features that we are concerned about, such as gravel, fracture, muddy beds, muddy strips, and other areas of geological phenomena distribution. There are also many uninteresting image features that have no specific geological significance, influenced by the measurement and complexity of strata and rocks.

The fourth column shows the recognition result of image features. It can be seen that the algal laminae, clay band, thick mudstone, induced fracture, dissolution pores, and other features in the image are well recognized. However, there are still many uninteresting features, which have an impact on the subsequent analysis.

We apply the three valuable data extraction strategies mentioned earlier to this well, with the effects shown in the fifth to seventh columns. The fifth to seventh columns show the results of the merging method, intersection method, and elimination method, respectively. We can see that merging method can effectively remove valueless data from the image.

6 Conclusion

In the multiclass classification problem, the classification result is greatly affected by the presence of valueless data. To solve this problem, we propose three valuable data extraction strategies: the merging method, intersection method, and elimination method. These strategies attempt to identify interesting and valueless data using the binary classification method. The experimental results and effects when operating in the production environment show that the merging method can effectively extract the valuable data,

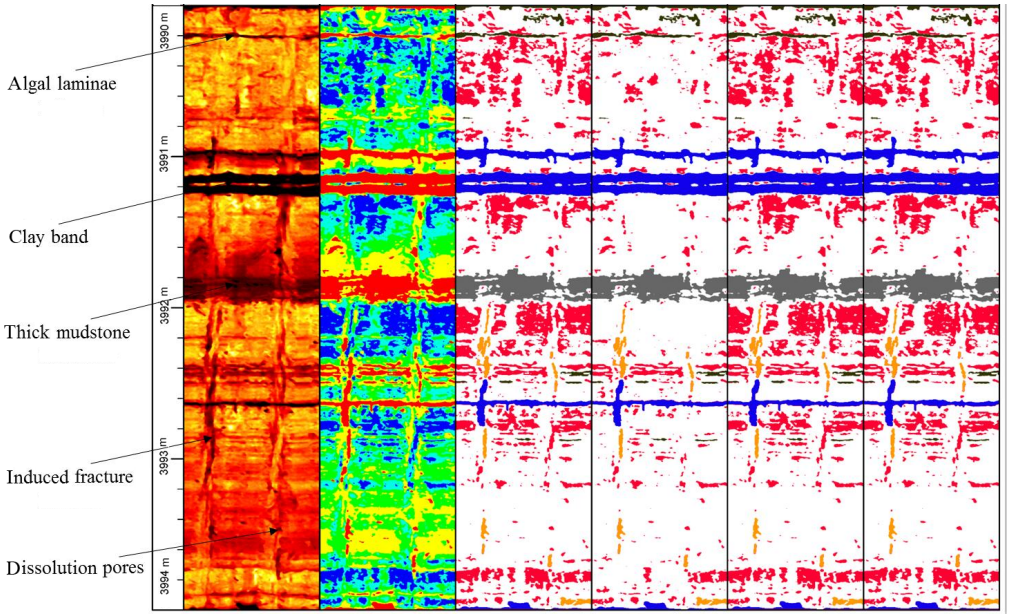


Fig. 14 Effect diagrams of three image cleaning strategies in practical production environment. Columns 1 to 3 refer to depth, the resistivity imaging logging image, and the result of image clustering segmentation. The fourth column is the recognition result of image features. Columns 5 to 7 refer merging method, intersection method, and elimination method, respectively.

to produce a clean image. Therefore, the merging method is well suited to improving the accuracy of the automatic recognition of geological characteristics in resistivity imaging logging images.

References

[1] H. Li and H. Wang, Investigation of eccentricity effects and depth of investigation of azimuthal resistivity LWD tools using 3D finite difference method, *Journal of Petroleum Science and Engineering*, vol. 143, no. 9, pp. 211–225, 2016.

[2] H. Wang and M. C. Fehler, The wavefield of acoustic logging in a cased hole with a single casing — part II: A dipole tool, *Geophysical Journal International*, vol. 212, no. 2, pp. 1412–1428, 2018.

[3] H. Wang and M. C. Fehler, The wavefield of acoustic logging in a cased-hole with a single casing — part I: A monopole tool, *Geophysical Journal International*, vol. 212, no. 1, pp. 612–626, 2018.

[4] H. Wang, T. Guo, and X. F. Shang, Understanding acoustic methods for cement bond logging, *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2407–2416, 2016.

[5] S. D. Russell, M. Akbar, B. Vissapragada, and G. M. Walkden, Rock types and permeability prediction from dipmeter and image logs; Shuaiba Reservoir (Aptian), Abu Dhabi, *AAPG Bulletin*, vol. 86, no. 10, pp. 1709–1732, 2002.

[6] D. V. Chitale, Borehole imaging in reservoir characterization: Implementation of a standard interpretation workflow for the clastic and carbonate reservoirs, in *Proc. 2005 IEEE Int. Conf. the 46th Annual Symposium of the SPWLA*, New Orleans, LA, USA, 2005, pp. 1–43.

[7] O. Faivre and G. Catala, Dip estimation from azimuthal Laterolog tools, in *Proc. 1995 IEEE Int. Conf. SPWLA 36th Annual Logging Symposium*, Paris, France, 1995, pp. 1–10.

[8] J. Prilliman, C. Bean, M. Hashem, T. Bratton, M. A. Fredette, and J. R. Lovell, A comparison of wireline and LWD resistivity images in the Gulf of Mexico, in *Proc. 1997 IEEE Int. Conf. 38th SPWLA Annual Logging Symposium*, Houston, TX, USA, 1997, pp. 1–11.

[9] R. Nurmi, M. Charara, M. Waterhouse, and R. Park, Heterogeneities in carbonate reservoirs: Detection and analysis using borehole electrical imagery (in geological applications of wireline logs), *Geological Society Special Publications*, vol. 48, no. 1, pp. 95–111, 1990.

[10] D. Moos, P. Peska, T. Finkbeiner, and M. Zoback, Comprehensive wellbore stability analysis utilizing quantitative risk assessment, *Journal of Petroleum Science and Engineering*, vol. 38, nos. 3&4, pp. 97–109, 2003.

[11] P. A. Pezard and S. M. Luthi, Borehole electrical images in the basement of the cajon pass scientific drillhole, California; fracture identification and tectonic implications, *Geophysical Research Letters*, vol. 15, no. 9, pp. 1017–1020, 1988.

[12] H. Q. Gao, H. F. Wang, Z. Feng, M. X. Fu, C. N. Ma, H. X. Pan, B. S. Xu, and N. Li, A novel texture extraction method for the sedimentary structures' classification of petroleum imaging logging, in *Proc. 2016 IEEE Int. Conf. on Pattern Recognition*, Chengdu, China, 2016, pp. 161–172.

[13] H. Chai, N. Li, X. L. Liu, D. L. Li, C. Z. Wang, D. C. Wu, and C. Xiao, Automatic discrimination of sedimentary facies and lithologies in reef-bank reservoirs using borehole image logs, *Applied Geophysics*, vol. 6, no. 1, pp. 17–29, 2009.

[14] H. F. Wang, Y. T. Wang, and H. Chai, State-of-the-art on texture-based well logging image classification, (in Chinese), *Journal of Computer Research and Development*, vol. 50, no. 6, pp. 1335–1348, 2013.

[15] Y. Q. Yang, W. P. Cui, and X. Zhang, A new method to detect formation occurrence from image recognition of electric imaging logging, (in Chinese), *Petroleum Geophysics*, vol. 56, no. 2, pp. 302–308, 2017.

[16] S. Z. Ke, Full 3-D numerical modeling of borehole electric image logging and the evaluation model of fracture, *Science in China Series D: Earth Sciences*, vol. 51, pp. 170–173, 2008.

[17] C. Barton, D. Moos, and K. Tezuka, Geomechanical wellbore imaging: Implications for reservoir fracture permeability, *AAPG Bulletin*, vol. 93, no. 11, pp. 1551–1569, 2009.

[18] B. M. Newberry, L. M. Grace, and D. O. Stief, Analysis of carbonate dual porosity systems from borehole electrical images, in *Proc. 1996 IEEE Int. Conf. on Permian Basin Oil and Gas Recovery Conference*, Midland, TX, USA, 1996, pp. 1–7.

[19] T. Li, S. H. Zhu, and M. Ogihara, Using discriminant analysis for multi-class classification: An experimental investigation, *Knowledge and Information Systems*, vol. 10, no. 4, pp. 453–472, 2006.

[20] K. S. Goh, E. Y. Chang, and B. Li, Using one-class and two-class SVMs for multiclass image annotation, *Journals & Magazines*, vol. 17, no. 10, pp. 1333–1346, 2005.

[21] T. Hastie and R. Tibshirani, Classification by pairwise coupling, *The Annals of Statistics*, vol. 26, no. 2, pp. 451–471, 1998.

[22] T. G. Dietterich and G. Bakiri, Solving multiclass learning problems via error-correcting output codes, *Journal of Artificial Intelligence Research*, vol. 2, no. 1, pp. 263–286, 1995.

[23] E. B. Kong and T. G. Dietterich, Error-correcting output coding corrects bias and variance, in *Proc. 1995 IEEE Int. Conf. on the 12th Machine Learning*, Tahoe City, CA, USA, 1995, pp. 313–321.

[24] V. J. Hodge and J. Austin, A survey of outlier detection methodologies, *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[25] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, OPTICS-OF: Identifying local outliers, in *Proc. 1999 IEEE Int. Conf. on the Third European Conference on Principles of Data Mining and Knowledge Discovery*, Prague, Czech Republic, 1999, pp. 262–270.

[26] M. M. Breunig, H. P. Kriegel, T. N. Raymond, and J. Sander, LOF: Identifying density-based local outliers, in *Proc. 2000 IEEE ACM SIGMOD International Conference on Management of Data*, Dallas, TX, USA, 2000, pp. 93–104.

[27] C. Li and W. H. Wong, Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *PNAS*, vol. 98, no. 1, pp. 31–36, 2001.

[28] J. Hardin and D. M. Rocke, Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator, *Computational Statistics and Data Analysis*, vol. 44, no. 4, pp. 625–638, 2004.

[29] Y. J. Zhang, *Image Analysis (Second Edition)*, (in Chinese). Beijing, China: Tsinghua University Press, 2005.

[30] G. Shi, *Data Mining and Knowledge Discovery for Geoscientists*, (in Chinese). Beijing, China: Petroleum Industry Press, 2013.

[31] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (Second Edition)*. USA: Prentice Hall, 2002.

[32] T. Blaschke, Object based image analysis for remote sensing, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 1, pp. 2–16, 2010.

[33] G. Shi, Superiorities of support vector machine in fracture prediction and gassiness evaluation, *Journal of Petroleum Exploration and Development*, vol. 35, no. 5, pp. 588–594, 2008.

[34] L. Bruzzone and D. F. Prieto, A technique for the selection of kernel-function parameters in RBF neural networks for classification of remote-sensing images, *Geoscience and Remote Sensing*, vol. 37, no. 2, pp. 1179–1184, 1999.

[35] V. Cherkassky and Y. Q. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks*, vol. 17, no. 1, pp. 113–126, 2004.

[36] J. E. Wang and J. Z. Qiao, Parameter selection of SVR based on improved k-fold cross validation, *Applied Mechanics and Materials*, vol. 462, no. 463, pp. 182–186, 2013.

[37] T. S. Joaquin, H. E. Carlos, and F. R. Mercedes, Improving adaptive boosting with k-cross-fold validation, in *Proc. 2006 IEEE Int. Conf. on Intelligent Computing*, Kunming, China, 2006, pp. 397–402.

[38] P. Lameski, E. Zdravevski, R. Mingov, and A. Kulakov, SVM parameter tuning with grid search and its impact on reduction of model over-fitting, in *Proc. 2015 IEEE Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, Tianjin, China, 2015, pp. 464–474.

[39] C. Liu, S. Q. Yin, M. Zhang, Y. Zeng, and J. Y. Liu, An improved grid search algorithm for parameters optimization on SVM, *Applied Mechanics and Materials*, vols. 644–650, pp. 2216–2219, 2014.

[40] R. Liu, E. Liu, J. Yang, M. Li, and F. Wang, Optimizing the hyper-parameters for SVM by combining evolution strategies with a grid search, in *Proc. 2006 IEEE Int. Conf. on Intelligent Computing*, Kunming, China, 20016, pp. 712–721.

[41] J. F. Kruiger, P. E. Rauber, R. M. Martins, A. Kerren, S. Kobourow, and A. C. Telea, Graph layout by t-SNE, *Computer Graphics Forum*, vol. 36, no. 3, pp. 283–294, 2017.

[42] A. Gisbrecht, A. Schulz, and B. Hammer, Parametric nonlinear dimensionality reduction using kernel t-SNE, *Neurocomputing*, vol. 147, no. 5, pp. 71–82, 2015.

[43] Z. Yang, C. W. Wang, and E. Oja, Multiplicative updates for t-SNE, in *Proc. 2010 IEEE Int. Conf. on Machine Learning for Signal Processing*, Kittila, Finland, 2010, pp. 19–23.

[44] M. A. Li, X. Y. Luo, and J. F. Yang, Extracting the nonlinear features of motor imagery EEG using parametric t-SNE, *Neurocomputing*, vol. 218, no. 19, pp. 371–381, 2016.

Yili Ren received the master degree from Beihang University in 2011. She is currently working in the PetroChina Research Institute of Petroleum Exploration and Development (RIPED). Her research interests generally focus on machine learning and its application in oil exploration and development.

Renbin Gong received the PhD degree from China University of Petroleum. He is a professor-level senior engineer of the Computer Applied Technology Research Institute of RIPED. His research interests include internet of things and the application of artificial intelligence in oil field.

Zhou Feng received the PhD degree from RIPED, in 2014. He is currently a senior engineer of RIPED. His research interests focus on well log processing and interpretation methods.

Meichao Li received the bachelor degree from Shenyang University of Technology in 2013. At present, he works in RIPED. His research interests mainly focus on machine learning and natural language processing.