Sedimentary environment prediction of grain-size data based on machine learning approach

Qiao Su[1], Yanhui Zhu[2], Fang Hu[3], and Xingyong Xu[1]

Abstract

Grain size is one of the most important records for sedimentary environment, and researchers have made remarkable progress in the interpretation of sedimentary environments by grain size analysis in the past few decades. However, these advances often depend on the personal experience of the scholars and combination with other methods used together. Here, we constructed a prediction model using the K-nearest neighbors algorithm, one of the machine learning methods, which can predict the sedimentary environments of one core through a known core. Compared to the results of other studies based on the comprehensive data set of grain size and four other indicators, this model achieved a high precision value only using the grain size data. We have also compared our prediction model with other mainstream machine learning algorithms, and the experimental results of six evaluation metrics shed light on that this prediction model can achieve the higher precision. The main errors of the model reflect the length of the conversation area of sedimentary environment, which is controlled by the sedimentary dynamics. This model can provide a quick comparison method of the cores in a similar environment; thus, it may point out the preliminary guidance for further study.

Introduction

The essence and formation of sediments are extremely relied on the distribution of grain size. At any location of sediment movement, the grain size varies greatly in space and time, thus controlling the physical transport and sedimentary processes (Switzer, 2013). Therefore, grain size analysis provides the significant clues to the sediment provenance, transport process, and sedimentary conditions (Blott and Pye, 2001; Zhang et al., 2018). Consequently, it is still a conventional procedure in sedimentology and engineering studies.

Although the analysis of grain size is usually used to identify sedimentary environments, landforms, and facies in various sedimentary settings, there are still noticeable differences in the interpretation techniques (Syvitski, 2007). Folk and Ward (1957) distinguish sedimentary environments using statistical parameters to characterize curves and bivariate plots of statistical measures: the mean, median, standard deviation, kurtosis, and skewness. Subsequently, two main ways of the parameter analysis of grain size were promoted: the graphic method and the moment method (Blott and Pye, 2001). Donato et al. (2009) consider that both ways have some disadvantages through comparison. The graphic method is quite insensitive to sediments with a large grain size range at the tail, which can be either merits or demerits depending on the specific problem under study. The moment method also overemphasized the importance of low frequencies with long tails, and under this circumstance, the Folk and Ward methods may more accurately describe the common characteristics of most samples.

Actually, most sediments, even when moderately well sorted, are polymodal, and the grain size generally follows a certain natural distribution. Some researchers argued that further information can be gained if the grain size and frequency scales are logarithmically transformed (Ashley, 1978; Gammon et al., 2017; Lu et al., 2018), yet Syvitski (2007) suggests that geologists should go beyond the sole use of the (log-) normal distribution as a model for grain size distribution. Specifically, the sediments of marine are composed of grain sizes with various sizes, which cannot be completely characterized using a single parameter or proxy (Bockelmann et al., 2018), whereas lacustrine sediments have a large uncertainty accounting for the origin components of grain size based on the polymodal distribu-

[1]First Institute of Oceanography, Ministry of Natural Resources, Key Laboratory of Marine Sedimentology and Environmental Geology, Qingdao 266061, China. E-mail: suqiao@fio.org.cn; xuxingyong@fio.org.cn.

[2]University of West Florida, Department of Mathematics and Statistics, Pensacola, Florida 32514, USA. E-mail: yzhu@students.uwf.edu.

[3]Hubei University of Chinese Medicine, College of Information Engineering, Wuhan 430065, China. E-mail: naomifang@hbtcm.edu.cn (corresponding author).

Manuscript received by the Editor 19 August 2019; revised manuscript received 16 January 2020; published ahead of production 16 March 2020; published online 14 May 2020. This paper appears in *Interpretation*, Vol. 8, No. 3 (August 2020); p. SL71–SL78, 4 FIGS., 6 TABLES.

http://dx.doi.org/10.1190/INT-2019-0153.1.

tions with the intrinsic complexity (Xiao et al., 2013). Most cores in the Bohai Sea area contain marine, lacustrine, and fluvial sediments; therefore, it is difficult to distinguish them, especially using only one indicator of grain size. According to experimental analysis of grain size, Sun et al. (2002) hold that the Weibull function is a suitable mathematical description in various sediments, whereas the normal function is also receivable in lacustrine and fluvial sediments. Despite that these methods exhibited well distribution of grain size in sediment, they usually require the personal experience of the researchers. Some other attempts have also been successful, but most are based on regional trends that could not be considered universal, and it is commonly used as supporting evidence for sedimentary environmental evolution studies (Gyllencreutz, 2005).

Recently, a great number of machine learning algorithms are applied in various research areas (Alaudah et al., 2019; Hu et al., 2018; Zhou et al., 2018). For example, these methods can reveal the potential relationship and knowledge hidden among marine data (Diesing et al., 2014; Cordier et al., 2017; Su et al., 2018). Furthermore, scientists are painfully aware of the challenges and complexities of stratigraphic analysis in larger, more complex, and heterogeneous data sets. The continuous progress of pattern recognition and machine learning provides a mean to handle these challenges (Jayaram et al., 2015). After analyzing the different characteristics of grain size in sediment samples and comparing many mainstream prediction algorithms, we constructed a prediction model using the K-nearest neighbors (KNN) algorithm (Bentley, 1975) based on the grain size data only, and then we used it to predict the sedimentary environments of one core through another known core.

Research area and materials

Geologic background

In this study, the grain size samples were obtained from the Bohai Sea area, which is a semienclosed inland sea in China. Through the narrow Bohai Strait, the Bohai Sea is connected to the Yellow Sea with an average water depth of 18 m (Figure 1). It is formed responding to the Cenozoic subsidence and is featured by tectonic quiescence and stable sedimentation rates from the Neogene to the present (Yi et al., 2015).

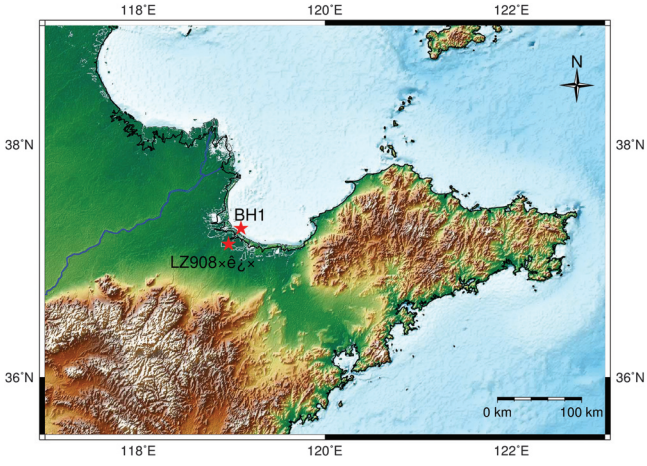

Figure 1. The locations of cores LZ908 and BH1.

The fluvial, lacustrine, and marine sediments that deposited in Bohai Sea basin range from 2000 to 3000 m. The Quaternary sediments can reach up to 400 m in thickness. In the Quaternary, the whole basin underwent four tectonic episodes. Before the late Quaternary, the sedimentary environment of Bohai Sea, a part of the Paleolake of North China, was dominated by the lacustrine and fluvial systems. Because the late Quaternary, sedimentary environments have been principally different among estuarine, delta, and tidal flat systems (Yao et al., 2014).

Cores

Cores LZ908 and BH1 were drilled in the summer of 2007 and the winter of 2008, respectively, by the First Institute of Oceanography, Ministry of Natural Resources, China. They are all located on shore near the southern Bohai Sea and are close to each other (Figure 1). The depth of LZ908 is 101.3 m with a 75% recovery rate; correspondingly, BH1 is 198.85 m and 85%. The interval of each grain size sample in LZ908 is 2 cm, yielding a total of 2141 samples for this study. The interval of BH1 is 10 cm, and only the upper 86.90 m (from 6.30 to 93.10 m) is used for this study; therefore, the total samples of core BH1 are 869.

The upper 54.3 m of LZ908 mainly involves coastal and marine sediments, whereas the lower 47.0 m is identified as fluvial and lacustrine sediments. The upper 47.0 m of BH1 contains coastal and fluvial sediments, and then lacustrine appears. The sedimentary environments of LZ908 and BH1 were well analyzed by (Yi et al., 2015, 2016) combining with various indices including grain size, tree pollen, magnetic susceptibility, radiocarbon dating, and optically stimulated luminescence dating. These papers hold that their deposition classification is correct and consider it the standard classification. The truth categories in this paper refer to their classification result.

Extraction of grain size samples

Before the analysis, all grain-size samples from two cores were removed the carbonate and organic matter by the chemical procedure. More specifically, approximately 2 g samples were preprocessed using 30% hydrogen peroxide (10–20 ml) to remove organic matter, washed with 10% hydrochloric acid to remove mollusk and carbonates fragments, rinsed using deionized water, and then placed in an ultrasonic vibrator for a few minutes to facilitate dispersion. One hundred grain size categories from 0.02 to 2000 were detected by a Malvern Mastersizer 2000 analyzer.

Data set of grain size

One hundred grain size categories were converted to ϕ values (Table 1) using the equation as follows:

$$\phi = \log_2 \frac{N}{1000}, \tag{1}$$

where N is the measurement values of grain size calculated from a 100 categories.

Then, the transformed ϕ values were divided into 51 properties from −0.50 to 11.75 according to the interval of 0.25. The 51 properties of each sediment sample represent the corresponding various magnitudes of grain sizes, of which the value ranges are $(0, 1)$. The digit describes the percentage of each magnitude occupying the total number of various grain sizes. Therefore, we built two data sets: a training set A with the 2141×51 matrix of LZ908 and a test set $\mathbf{X}$ with the 869×51 matrix of BH1, where $\mathbf{x}_{ij}$ represents the percentage of the jth magnitude in the ith sample. Descriptions of these two data sets are in Table 1.

Standard categories

For the Bohai Sea, researchers always classify the sedimentary samples based on grain size into three main categories: marine, fluvial, and lacustrine. Tables 2 and 3 show the ground-truth categories of the samples in LZ908 and BH1, respectively. To evaluate the precision and efficiency of our prediction model, we compared our predicted results of core BH1 with these standard categories.

Methods

Idea of sedimentary environment prediction

Based on the 2141 sediment samples with 51 properties (i.e., the various magnitudes of grain sizes) in LZ908, we constructed the sedimentary environment prediction model using the KNN algorithm, a nonparametric method, applied for classification and regression. In this model, we set the samples from LZ908 and BH1 as the training set and test set, respectively. First, we selected k neighbors and calculated the distances between the test set and the training set using the K-dimensional (KD) Tree searching algorithm (Zhou et al., 2008). We took the first k training samples with the closest distances to the test sample. Then, based on the categories to which the k nearest neighbors belong, we predicted the categories of the test samples.

Steps of sedimentary environment prediction

We have 2141 training records denoted as A, and we have 869 test records denoted as $\mathbf{X}$. These two data sets all have 51 properties: $a = \{a_1, a_2, \ldots, a_{51}\}$ and $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{51}\}$; here, a and $\mathbf{x}$ are one of the training and test samples, respectively. The number of sedimentary environment categories is three. The steps of sedimentary environment prediction are as follows:

Step 1: We initialize the sample $\mathbf{x}$ and randomly select $A[1] - A[k]$ as the initial neighbors of $\mathbf{x}$.

Table 2. Categories of training set (LZ908).

No.	Categories	Number of samples	Percentage (%)
1	Marine	338	15.787
2	Fluvial	781	36.478
3	Lacustrine	1022	47.735

Table 3. Categories of test set (BH1).

No.	Categories	Number of samples	Percentage (%)
1	Marine	138	15.880
2	Fluvial	372	42.808
3	Lacustrine	359	41.312

Table 1. The description of grain size samples in cores LZ908 and BH1. For the sample LZ0001, 0.17 represents the percentage of grain sizes with 11.75–11.50 magnitudes accounting for the total magnitudes.

No. of samples	Depth of top (m)	Depth of bottom (m)	>11.75	11.75–11.50	...	0.25–0.00	0.00 to −0.25	−0.25 to −0.50	< −0.50
LZ0001	2.45	2.47	0.00	0.17	...	0.00	0.00	0.00	0.00
...	...	...	...	...	...	...	...	...	...
LZ2141	101.32	101.34	0.00	0.16	...	0.00	0.00	0.00	0.00
BH001	6.30	6.40	0.00	0.09	...	0.00	0.00	0.00	0.00
...	...	...	...	...	...	...	...	...	...
BH869	93.10	93.20	0.13	0.43	...	0.00	0.00	0.00	0.00

Step 2: Using the KD tree searching algorithm, we calculate the distances, $d(\mathbf{x}, A[i])$, between $\mathbf{x}$ and the k neighbor samples $A[i]$; here, $i = 1, 2, \ldots, k$. We rank $d(\mathbf{x}, A[i])$ as an ascending sequence and mark the maximum distant $D = \max\{d(\mathbf{x}, A[i])\}$ between $\mathbf{x}$ and $A[i]$.

Step 3: We also calculate the distances $d(\mathbf{x}, A[j])$, between $\mathbf{x}$ and the other samples $A[j]$, where $j = k+1, \ldots, n$, and n is the number of training samples.

Step 4: If $d(\mathbf{x}, A[j]) < D$, set $D = A[j]$, and then go to step 6; otherwise, proceed to step 5.

Step 5: We also rank $d(\mathbf{x}, A[j])$ as an ascending order and mark the maximum distance $\max\{d(\mathbf{x}, A[j])\}$ as D.

Step 6: If traversing the test samples is unfinished, then, return to step 3; otherwise, proceed to step 7.

Step 7: By evaluating the probability of categories p, to which these k nearest neighbors belong, we can predict that $\mathbf{x}$ belongs to the category with the maximum probability.

The pseudocode of predicting sedimentary environment using KNN is shown in Algorithm 1.

Algorithm 1. Predicting sedimentary environment using KNN ($A, k, \mathbf{x}$)

Input:

A // Training set of grain size

k // Number of neighbors

$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{51}\}$ // One of the test set with 51 properties

Output:

C // Category $\mathbf{x}$ belonging to

Initialize:
$A[1] \sim A[k]$ // Neighbors of $\mathbf{x}$
$d(\mathbf{x}, A[i]), i = 1, 2, \ldots, k$ // Distances between $\mathbf{x}$ and $A[i]$
Rank $d(\mathbf{x}, A[i])$ as an ascending order
$D = \max\{d(\mathbf{x}, A[i])\}$ // Maximum distance between $\mathbf{x}$ and $A[i]$
for $j = k+1, \ldots, n$ **do**
  $d(\mathbf{x}, A[j])$ // Distances between $\mathbf{x}$ and $A[j]$
  if $d(\mathbf{x}, A[j]) < D$ **then**
    $D = A[j]$
  end if
  Rank $d(\mathbf{x}, A[j])$ as an ascending order
  $D = \max\{d(\mathbf{x}, A[j])\}$ // Maximum distance between $\mathbf{x}$ and $A[j]$
end for
Probability of categories p // k nearest neighbors belonging to
return $\mathbf{x} \rightarrow C$ // Maximum probability

Results and analysis

Predicted result evaluation

Grain-size analysis, one of the basic tools to classify different sedimentary environments, is rarely used for comparative analysis as a single indicator. Researchers usually combined it with other indicators for the sedimentary environment analysis. In the Bohai area, there exist three main categories of sedimentary environments: marine, fluvial, and lacustrine.

By using the prediction model trained by the grain size samples in LZ908, we predicted the sedimentary environments of BH1. Furthermore, we used six evaluation metrics (Hu et al., 2019): adjusted rand index (ARI) (Feizollah et al., 2014), adjusted mutual information (AMI) (Romano et al., 2016), normalized mutual information (NMI) (Ana and Jain, 2003; Danon et al., 2005), homogeneity (Rosenberg and Hirschberg, 2007), completeness (Rosenberg and Hirschberg, 2007), and precision (Biswas and Biswas, 2015), to verify the precision and efficiency of the predicted results of BH1. We showed the evaluation metric values of predicted results based on the KNN model with different neighbors in Table 4 and Figure 2.

In general, the predicted result is considered to be a better one, if the metric can get a higher value. Based on Table 4 and Figure 2, we showed that the value of precision, the most significant metric, can reach up to the maximum 0.921922 compared to the standard classification when neighbor = 1. At the same condition, these six indices, ARI, AMI, NMI, homogeneity, completeness, and precision, can reach up to almost the same values when neighbor = 1 and neighbor = 3, a fact that these values are higher than the metric values acquired by other numbers of neighbors.

Predicted result comparison

We also tested the data sets of LZ908 and BH1 using four other mainstream prediction algorithms: stochastic gradient descent (SGD) (Zhang, 2004; Tsuruoka et al., 2009), classification and regression trees (CART) (Breiman, 2001), AdaBoost (Rätsch et al., 2001), and gradient boosted regression trees (GBRT) (Zheng et al., 2008). The predicted results are shown in Table 5 and Figure 3.

We compared the values of ARI, AMI, NMI, homogeneity, completeness, and precision, acquired by five representative algorithms, to evaluate the predicted results of BH1. The specific evaluation results are shown in Table 5. From Figure 3, we showed that the KNN model with neighbor = 1 achieved the higher values of ARI, AMI, NMI, homogeneity, and precision than other four algorithms.

Predicted result analysis

The predicted results of BH1 are shown in Figure 4, and the red points in grain size ϕ of BH1 are the different points from the ground-truth classification.

According to our predicted results, 6.3–10.8 m and 21.3–51.5 m are fluvial, 12.4–18.6 m and 74.3–79.3 m

are marine, whereas 52.5–72.5 m and 80.1–93.2 m are lacustrine. They are the same as the results of standard classification. The main different samples appear at 10.9–12.3 m, 19.5–21.2 m, 51.6–52.4 m, 72.6–74.2 m, and 79.4–80.0 m, in which are all the transition areas of different sedimentary environments. The areas at 10.9–12.3 m, 19.5–21.2 m are the transitions of marine and fluvial; 51.6–52.4 m is the transition of fluvial and lacustrine; 72.6–74.2 m is the transition of lacustrine and marine; whereas 79.4–80.0 m is the transition of marine and lacustrine. The former of each transition area represents the upper sediment classification, and the latter denotes the lower classification, and each conversion area is where the prediction errors are concentrated, and the prediction errors all occur in the lower sedimentary environment.

As discussed by Syvitski (2007), the accurate prediction of sediment properties depends upon the empirical relationships between various physical properties and their expected variability. In most cases, the results acquired from different analytical methods of grain size are contradictory. Syvitski (2007) suggests that this fact is commonly correct and helpful as well because not all sediment masses were moved by only one transport agency. Many modern sediments represent mixed agencies or transition from one agency to another. The most important difficulty lies in the fact that certain sediments have been studied at a moment of transition; that is, when there is no single responsible agency or environment in fact. This difficulty may be the main problem for our wish to erect sharply defined class boundaries.

Sediment mass delivered by one agency to another may not lose the granulometric fingerprint of the former until some considerable time has passed; therefore, the analyst should not insist on selecting a single agency but instead be aware that there may be two or more. In this paper, most prediction errors occur in transition areas, a fact that shows our results conforming to this rule, essentially because these transition areas are originally the focus and difficulty of sedimentary environment research and usually need to be solved by encrypted sampling and dating. Furthermore, the transition zones usually have fluctuations in two types of depositions or more. Therefore, the areas where prediction errors appear represent the history of sedimentary conversion.

More specifically, below 80.1 m is the lacustrine environment, then the transgression occurs, but the lacustrine environment has not lost its granulometric fingerprint at all. Along with the transgression gradually strengthening, the sedimentary environment completely converts to marine at 79.4 m. The depth of 79–80.1 m

Table 4. Predicted result comparison with different neighbors in six evaluation metrics.

Evaluation indices	Number of neighbors					
	1	2	3	4	5	6
ARI	0.705855	0.679817	0.707183	0.658569	0.664181	0.651447
AMI	0.592659	0.573971	0.59665	0.557849	0.560503	0.550852
NMI	0.603029	0.587932	0.607756	0.572712	0.574407	0.565703
Homogeneity	0.612657	0.601266	0.618153	0.586969	0.58766	0.579935
Completeness	0.593553	0.574893	0.597533	0.558802	0.561453	0.55182
Precision	0.921922	0.907422	0.920713	0.894131	0.896548	0.889298

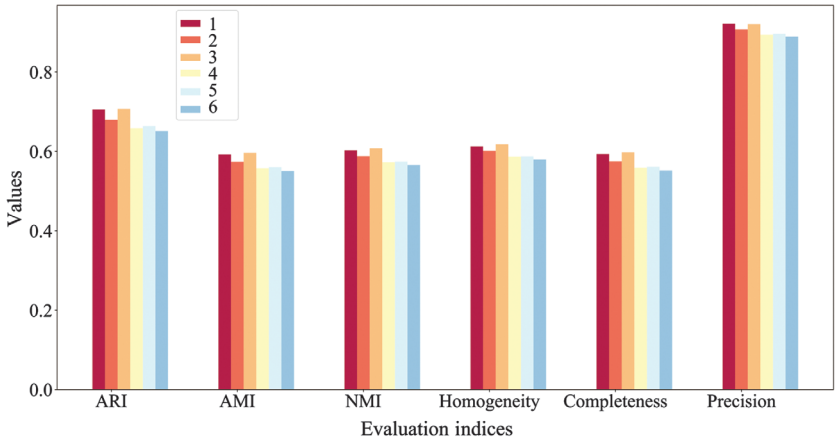

Figure 2. Predicted result comparison based on KNN model with different neighbors in six evaluation metrics.

Table 5. Predicted result comparison with different algorithms in six evaluation metrics.

Evaluation indices	Algorithms				
	SGD	CART	AdaBoost	GBRT	KNN
ARI	0.503781	0.672105	0.688467	0.550774	**0.705855**
AMI	0.494829	0.589014	0.542462	0.433969	**0.592659**
NMI	0.504612	0.598285	0.559071	0.447257	**0.603029**
Homogeneity	0.495974	0.589493	0.542979	0.434383	**0.612657**
Completeness	0.513401	0.607208	**0.799984**	0.639987	0.593553
Precision	0.714097	0.745685	0.835443	0.668354	**0.921922**

represents the period of the process of transformation from lacustrine to marine. Similarly, 72.6–74.2 m is the process from marine to lacustrine, 51.6–52.4 m is the process from lacustrine to fluvial, 19.5–21.2 m is the process from fluvial to marine, and 10.9–12.3 m is the process from marine to fluvial.

The change of grain size is often used to indicate sedimentary environment and transport dynamic conditions. The results of the KNN model give the main sedimentary classification and conversation area. The variation intervals between the sedimentary environments show the strength of hydrodynamic conditions. Usually, river and ocean have stronger hydrodynamic conditions than lake. Therefore, when the lacustrine is transformed into marine (79.4–80.1 m) or fluvial (51.6–52.4 m), the transformation speed is faster, and the corresponding conversation zone thickness is 0.7 and 0.8 m, respectively. On the contrary, it takes longer to transform from marine to lacustrine (72.6–74.2 m), and the thickness reaches up to 1.6 m. The transition between marine and fluvial, influenced by global sea level changes, also takes a relatively long time.

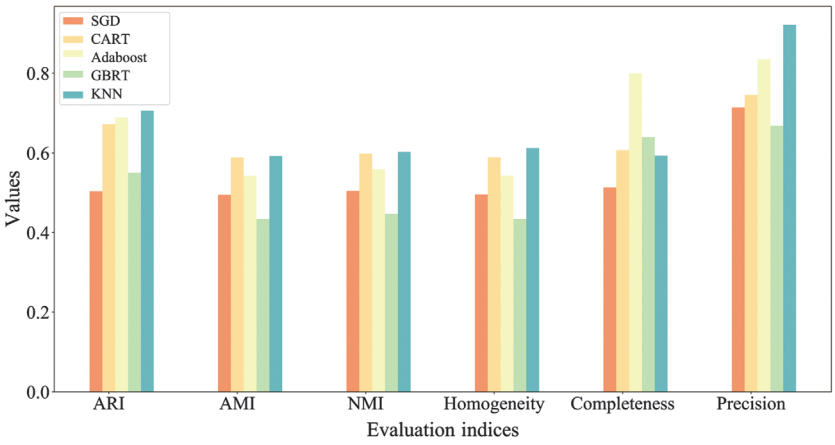

Figure 3. Predicted result comparison with different algorithms.

Conclusion

In previous studies, the sedimentary environments of cores LZ908 and BH1 were studied systematically. In this paper, we compared five mainstream machine learning prediction methods, and the results show that the KNN model is the best in predicting sedimentary environment using grain size only. The conclusions are summarized as follows:

1) When parameter neighbor = 1, the precision of the KNN model to predict the three main categories of core BH1 can reach up to 0.92192175. The KNN model can get the highest values of ARI, AMI,

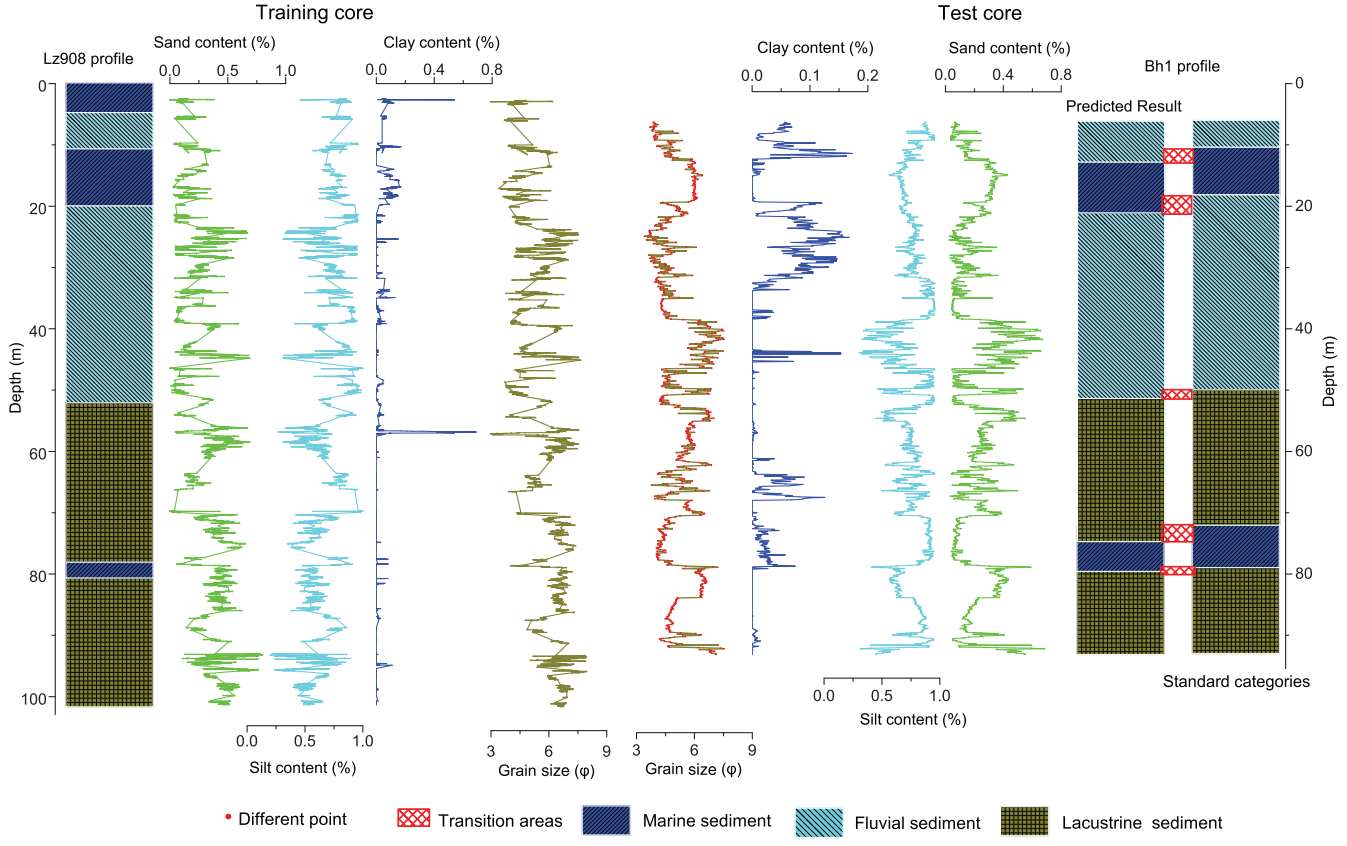

Figure 4. Predicted result analysis.

NMI, homogeneity, and precision than other four mainstream prediction algorithms.
2) The prediction errors mainly occur in the conversation areas of different sedimentary environments, which just need to be focused on and deeply researched. When the environment with a strong hydrodynamic force replaces the weak one, the conversation area is shorter; otherwise, it is longer.
3) Our analysis results can provide preliminary guidance for subsequent sample collection and testing. The length of the conversion area can be estimated according to the transformation of different sedimentary environments, then encrypt sampling and dating the conversion area.

Grain size is a very convenient and inexpensive index in sediment analysis; basically, it is measured for each core. When we get a new core, if there is a known core in a similar sedimentary environment, researchers can use the proposed prediction model to get the main sediment classifications and explore the conversation areas of sedimentary environments, so as to provide preliminary guidance for the next work. However, this model can only predict the main sedimentary classifications now, sedimentary microfacies needs more data and further study.

Acknowledgments

We acknowledge the funding support from the National Natural Science Foundation of China (U1806212 and 41406072) and the Natural Science Foundation of Hubei Province (2018CFB259).

Data and materials availability

Data associated with this research are available and can be obtained by contacting the corresponding author.

References

Alaudah, Y., P. Michalowicz, M. Alfarraj, and G. AlRegib, 2019, A machine learning benchmark for facies classification: Interpretation, **7**, SE175–SE187, doi: 10.1190/INT-2018-0249.1.

Ana, L., and A. K. Jain, 2003, Robust data clustering: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1–6.

Ashley, G. M., 1978, Interpretation of polymodal sediments: The Journal of Geology, **86**, 411–421, doi: 10.1086/649710.

Bentley, J. L., 1975, Multidimensional binary search trees used for associative searching: Communications of the ACM, **18**, 509–517, doi: 10.1145/361002.361007.

Biswas, A., and B. Biswas, 2015, Investigating community structure in perspective of ego network: Expert Systems with Applications, **42**, 6913–6934, doi: 10.1016/j.eswa.2015.05.009.

Blott, S. J., and K. Pye, 2001, GRADISTAT: A grain size distribution and statistics package for the analysis of unconsolidated sediments: Earth Surface Processes and Landforms, **26**, 1237–1248, doi: 10.1002/esp.v26:11.

Bockelmann, F.-D., W. Puls, U. Kleeberg, D. Müller, and K.-C. Emeis, 2018, Mapping mud content and median grain-size of North sea sediments — A geostatistical approach: Marine Geology, **397**, 60–71, doi: 10.1016/j.margeo.2017.11.003.

Breiman, L., 2001, Random forests: Machine Learning, **45**, 5–32, doi: 10.1023/A:1010933404324.

Cordier, T., P. Esling, F. Lejzerowicz, J. Visco, A. Ouadahi, C. Martins, T. Cedhagen, and J. Pawlowski, 2017, Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning: Environmental Science & Technology, **51**, 9118–9126, doi: 10.1021/acs.est.7b01518.

Danon, L., A. Diaz-Guilera, J. Duch, and A. Arenas, 2005, Comparing community structure identification: Journal of Statistical Mechanics: Theory and Experiment, **2005**, P09008, doi: 10.1088/1742-5468/2005/09/P09008.

Diesing, M., S. L. Green, D. Stephens, R. M. Lark, H. A. Stewart, and D. Dove, 2014, Mapping seabed sediments: Comparison of manual, geostatistical, object-based image analysis and machine learning approaches: Continental Shelf Research, **84**, 107–119, doi: 10.1016/j.csr.2014.05.004.

Donato, S., E. Reinhardt, J. Boyce, J. Pilarczyk, and B. Jupp, 2009, Particle-size distribution of inferred tsunami deposits in Sur Lagoon, Sultanate of Oman: Marine Geology, **257**, 54–64, doi: 10.1016/j.margeo.2008.10.012.

Feizollah, A., N. B. Anuar, R. Salleh, and F. Amalina, 2014, Comparative study of k-means and mini batch k-means clustering algorithms in Android malware detection using network traffic analysis: International Symposium on Biometrics and Security Technologies, IEEE, 193–197.

Folk, R. L., and W. C. Ward, 1957, A study in the significance of grain size parameters: Journal of Sedimentary Research, **27**, 3–26, doi: 10.1306/74D70646-2B21-11D7-8648000102C1865D.

Gammon, P. R., L. A. Neville, R. T. Patterson, M. M. Savard, and G. T. Swindles, 2017, A log-normal spectral analysis of inorganic grain-size distributions from a Canadian boreal lake core: Towards refining depositional process proxy data from high latitude lakes: Sedimentology, **64**, 609–630, doi: 10.1111/sed.12281.

Gyllencreutz, R., 2005, Late glacial and holocene paleoceanography in the Skagerrak from high-resolution grain size records: Palaeogeography, Palaeoclimatology, Palaeoecology, **222**, 344–369, doi: 10.1016/j.palaeo.2005.03.025.

Hu, F., M. Wang, Y. Zhu, J. Liu, and Y. Jia, 2018, A time simulated annealing-back propagation algorithm and its application in disease prediction: Modern Physics Letters B, **32**, 1850303, doi: 10.1142/S0217984918503037.

Hu, F., Y. Zhu, J. Liu, and Y. Jia, 2019, Computing communities in complex networks using the Dirichlet processing Gaussian mixture model with spectral clustering: Physics Letters A, **383**, 813–824, doi: 10.1016/j.physleta.2018.12.005.

Jayaram, V., P. A. Avseth, K. Azbel, T. Coléou, D. Devegowda, P. de Groot, D. Gao, K. Marfurt, M. Matos, T. Mukerji, and M. Poupon, 2015, Introduction to special section: Pattern recognition and machine learning: Interpretation, **3**, SAEi–SAEii, doi: 10.1190/INT2015-0918-SPSEINTRO.1.

Lu, Y., X. Fang, O. Friedrich, and C. Song, 2018, Characteristic grain-size component-a useful process-related parameter for grain-size analysis of lacustrine clastics: Quaternary International, **479**, 90–99, doi: 10.1016/j.quaint.2017.07.027.

Rätsch, G., T. Onoda, and K.-R. Müller, 2001, Soft margins for AdaBoost: Machine Learning, **42**, 287–320, doi: 10.1023/A:1007618119488.

Romano, S., N. X. Vinh, J. Bailey, and K. Verspoor, 2016, Adjusting for chance clustering comparison measures: The Journal of Machine Learning Research, **17**, 4635–4666.

Rosenberg, A., and J. Hirschberg, 2007, V-measure: A conditional entropy-based external cluster evaluation measure: The Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 410–420.

Su, Q., Y. Zhu, Y. Jia, P. Li, F. Hu, and X. Xu, 2018, Sedimentary environment analysis by grain-size data based on mini batch k-means algorithm: Geofluids, **2018**, 1–11, doi: 10.1155/2018/8519695.

Sun, D., J. Bloemendal, D. Rea, J. Vandenberghe, F. Jiang, Z. An, and R. Su, 2002, Grain-size distribution function of polymodal sediments in hydraulic and aeolian environments, and numerical partitioning of the sedimentary components: Sedimentary Geology, **152**, 263–277, doi: 10.1016/S0037-0738(02)00082-9.

Switzer, A. D., 2013, Measuring and analyzing particle size in a geomorphic context, in J. Shroder, A. D. Switzer, and D. M. Kennedy, eds., Treatise on geomorphology: Academic Press, **14**, 224–242.

Syvitski, J. P., 2007, Principles, methods and application of particle size analysis: Cambridge University Press.

Tsuruoka, Y., J. Tsujii, and S. Ananiadou, 2009, Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty: The Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 477–485.

Xiao, J., J. Fan, L. Zhou, D. Zhai, R. Wen, and X. Qin, 2013, A model for linking grain-size component to lake level status of a modern clastic lake: Journal of Asian Earth Sciences, **69**, 149–158, doi: 10.1016/j.jseaes.2012.07.003.

Yao, Z., X. Shi, Q. Liu, Y. Liu, J. C. Larrasoana, J. Liu, S. Ge, K. Wang, S. Qiao, X. Li, and F. Shi, 2014, Paleomagnetic and astronomical dating of sediment core BH08 from the Bohai sea, China: Implications for glacial-interglacial sedimentation: Palaeogeography, Palaeoclimatology, Palaeoecology, **393**, 90–101, doi: 10.1016/j.palaeo.2013.11.012.

Yi, L., C. Deng, L. Tian, X. Xu, X. Jiang, X. Qiang, H. Qin, J. Ge, G. Chen, Q. Su, and Y. Chen, 2016, Plio-Pleistocene evolution of Bohai Basin (East Asia): Demise of Bohai Paleolake and transition to marine environment: Scientific Reports, **6**, 29403, doi: 10.1038/srep29403.

Yi, L., C. Deng, X. Xu, H. Yu, X. Qiang, X. Jiang, Y. Chen, Q. Su, G. Chen, P. Li, and J. Ge, 2015, Paleo-megalake termination in the quaternary: Paleomagnetic and water-level evidence from South Bohai Sea, China: Sedimentary Geology, **319**, 1–12, doi: 10.1016/j.sedgeo.2015.01.005.

Zhang, T., 2004, Solving large scale linear prediction problems using stochastic gradient descent algorithms: 21st International Conference on Machine Learning, ACM, 116.

Zhang, X., A. Zhou, X. Wang, M. Song, Y. Zhao, H. Xie, J. M. Russell, and F. Chen, 2018, Unmixing grain-size distributions in lake sediments: A new method of endmember modeling using hierarchical clustering: Quaternary Research, **89**, 365–373, doi: 10.1017/qua.2017.78.

Zheng, Z., H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun, 2008, A general boosting method and its application to learning ranking functions for web search: Advances in Neural Information Processing Systems, 1697–1704.

Zhou, C., K. Yin, Y. Cao, B. Ahmed, Y. Li, F. Catani, and H. R. Pourghasemi, 2018, Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the three gorges reservoir area, China: Computers & Geosciences, **112**, 23–37, doi: 10.1016/j.cageo.2017.11.019.

Zhou, K., Q. Hou, R. Wang, and B. Guo, 2008, Real-time KD-tree construction on graphics hardware: ACM Transactions on Graphics (TOG), **27**, 126.

Biographies and photographs of the authors are not available.