Research Article

Sedimentary Environment Analysis by Grain-Size Data Based on Mini Batch K-Means Algorithm

Qiao Su ¹, Yanhui Zhu,² Yalin Jia,² Ping Li,¹ Fang Hu ^{2,3} and Xingyong Xu¹

¹Key Laboratory of Marine Sedimentology and Environmental Geology, First Institute of Oceanography, State Oceanic Administration, Qingdao 266061, China

²College of Information Engineering, Hubei University of Chinese Medicine, Wuhan 430065, China

³Department of Mathematics and Statistics, University of West Florida, Pensacola 32514, USA

Correspondence should be addressed to Fang Hu; fhu1@uwf.edu

Received 26 April 2018; Accepted 23 September 2018; Published 2 December 2018

Academic Editor: Umberta Tinivella

Copyright © 2018 Qiao Su et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

During the last several decades, researchers have made significant advances in sedimentary environment interpretation of grain-size analysis, but these improvements have often depended on the subjective experience of the researcher and were usually combined with other methods. Currently, researchers have been using a larger number of data mining and knowledge discovering methods to explore the potential relationships in sediment grain-size analysis. In this paper, we will apply bipartite graph theory to construct a *Sample/Grain-Size* network model and then construct a *Sample* network model projected from this bipartite network. Furthermore, we will use the Mini Batch K-means algorithm with the most appropriate parameters (*reassignment ratio* $\epsilon = 0.025$ and *mini batch* = 25) to cluster the sediment samples. We will use four representative evaluation indices to verify the precision of the clustering result. Simulation results demonstrate that this algorithm can divide the *Sample* network into three sedimentary categorical clusters: *marine*, *fluvial*, and *lacustrine*. According to the results of previous studies obtained from a variety of indices, the precision of experimental results about sediment grain-size category is up to 0.92254367, a fact which shows that this method of analyzing sedimentary environment by grain size is extremely effective and accurate.

1. Introduction

Data mining, knowledge discovery, and machine learning algorithms have virtually permeated into research in various fields [1–4]. The complex network as a significant method of data mining gives top priority to discovering concealed information between things. Therefore, a great number of researchers from various research fields, including mathematics, physics, biology, chemistry, and oceanology, used the complex network to explore the potential relationships between data [5–9]. The complex network has some characteristics: self-similarity, self-organization, scale-free, small-world, community structure (cluster), and node centrality. The community structure is one of the most important traits because it can objectively reflect the potential relationships between nodes. A community is made of one group of nodes

within which the links between nodes are densely connected but between which they are sparsely connected with other clusters [10, 11].

The grain-size analysis is one of the basic tools for classifying sedimentary environments, an analysis which can provide important clues to the provenance, transport history, and depositional conditions [12]. In general, the representative statistical parameters of grain-size analysis involve median, mode, mean, separation parameter, skewness, and kurtosis [13]. During the last few decades, two computing methods of grain-size parameters were developed: the graphical method and the moment method [12]. Blott and Pye (2011) presented that these two analysis methods had some advantages and disadvantages in computing sediment grain-size samples with various parameters. As most sediments are polymodal, curve shape and statistical measures

usually simply reflect the relative magnitude and separation of populations. Polymodal grain-size spectrum can be considered as a result of the superposition of several unimodal components [14]. Many works have shown that different grain-size distribution is related to special transport and deposition process [15]. Three kinds of functions are commonly used to fit the grain-size distribution: Normal function, Lognormal function, and Weibull function [15]. Based on experimental results, Sun et al. [15] found that the Weibull function was appropriate for the mathematical description of the grain-size distribution of all kinds of sediments while the application of Normal function for fluvial and lacustrine sediments was also acceptable. Although these methods, especially Weibull function, performed well in sediment in fitting grain-size distribution, they often need subjective experience of the researchers, and the definite criteria for environmental determination have not been given. Based on the data of borehole Lz908, Yi et al. analyzed the evolution of the sedimentary environment. Besides grain-size data, they also used the data of magnetic susceptibility, tree pollen, radiocarbon dating, and optically stimulated luminescence (OSL) dating [16, 17]. Can the same conclusion be obtained by using only the grain-size data which are the relatively convenient and low-priced indices?

In this paper, we introduce complex network into the data modeling of sediment grain-size data. Based on the theory of bipartite graph [18], we construct the *Sample/Grain-Size* bipartite weighted network model which can objectively reflect the association relationships between sediment samples and grain sizes. By using projection, we will construct the *Sample* network model from the bipartite network. After repeatedly testing based on tens of representative clustering algorithms, we have selected the Mini Batch K-means algorithm [19], an optimization algorithm combined with the K-means algorithm [20], and the classical batch algorithm [21] to split the *Sample* nodes into their categories and find the relationships between the sedimentary environment and grain size. After 400 tests, we can find the most appropriate parameters in Mini Batch K-means algorithm. Finally, we will use four evaluation indices AMI, NMI, completeness, and precision to verify the accuracy and efficiency of clustering divisions.

2. Evaluation Functions

In the research field of complex networks, researchers always use several representative performance evaluation indices AMI, NMI, completeness, and precision to verify the accuracy and efficiency of clustering divisions. It is universally acknowledged that the higher the value of one index is, the better the result of clustering division is. Therefore, we will also use these four evaluation indices to verify the clustering result of sediment grain-size samples.

2.1. NMI and AMI. Normalized Mutual Information (NMI) [22, 23] is an approach to measure shared information between two data distribution by the information theory in which entropy is defined as the information included in a distribution [24].

$$MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i, j) \log \left(\frac{P(i, j)}{P(i)P'(j)} \right), \quad (1)$$

where $P(i, j) = |U_i \cap V_j|/N$ represents the probability that an object picks at random falls into both classes U_i and V_j . The two label assignments U and V have the corresponding entropy $H(U)$ and $H(V)$ defined as follows:

$$H(U) = - \sum_{i=1}^{|U|} P(i) \log (P(i)), \quad (2)$$

where $P(i) = |U_i|/N$ is the probability that an object picked at random from U falls into class U_i . The equation $H(V)$ has the similar definition $P'(j)$ with $P(i)$.

$$H(V) = - \sum_{j=1}^{|V|} P'(j) \log (P'(j)). \quad (3)$$

The NMI and the Adjusted Mutual Information (AMI) [25] are defined as

$$NMI(U, V) = \frac{MI(U, V)}{\sqrt{H(U)H(V)}}, \quad (4)$$

$$AMI(U, V) = \frac{MI - E[MI]}{\max(H(U), H(V)) - E[MI]},$$

where $E[MI]$ is the expected value for MI. The range of NMI and AMI is $[0, 1]$ and $[-1, 1]$, respectively.

2.2. Completeness. Based on the standard partitioning precious study results of known grain-size samples, conditional entropy analysis is used to define some intuitive measures. The completeness assigns all nodes in the given class to the same cluster [26, 27]. The completeness is formally given by

$$c = 1 - \frac{H(K|C)}{H(K)}, \quad (5)$$

where $H(K)$ is the entropy of the classes and $H(C|K)$ is the conditional entropy of the classes given the cluster assignments.

2.3. Precision. Precision [28] (P) is the number of true positives (T_p) over T_p plus the number of false positives (F_p). The precision is given by

$$P = \frac{T_p}{T_p + F_p}. \quad (6)$$

3. Dataset of Sediment/Grain-Size

The sediment samples of this study came from borehole Lz908 (37° 09'N, 118° 58'E), which is located in the southern Bohai Sea, China (Figure 1). The borehole was drilled to 101.3 m below the surface in 2007, and the recovery rate

FIGURE 1: The location of borehole Lz908.

TABLE 1: The dataset of “Sediment/Grain-Size” in the borehole Lz908.

No. of samples	No. of layers	Depth of top (m)	Depth of bottom (m)	$>11.75(\varphi)$	11.75~11.50	...	0.25~0.00	0.00~-0.25	-0.25~-0.50	$<-0.5 (\varphi)$
Lz01-01	0001	2.45	2.47	0.00	17%		0.00	0.00	0.00	0.00
...	...	...	...	...	...	...	...	...	...	...
Lz59-77	2141	101.32	101.34	0.00	16%		0.00	0.00	0.00	0.00

(The size of sediment samples is 2141, and the number of various magnitudes of grain size is 51; 0.17 denotes the percentage of grain sizes with the magnitude of 11.75~11.50 accounting for the total number of 51 magnitudes of grain size in sample Lz01-01).

reached 75%. The existing research results show that this region developed into three transgressive layers from late Pleistocene, and the thickness of *fluvial*, *lacustrine*, and *marine* sediments reached 2000–3000 m in this basin [16]. We extracted 2141 sediment samples of grain size from the borehole based on a 2 cm interval. We tested the grain size by using a thorough pretreatment method at the First Institute of Oceanography, State Oceanic Administration, China. The measuring instrument for grain size was a Mastersizer 2000 laser particle analyzer produced by the UK company Malvern; the measurement range was 0.02–2000 μm ; the repeated measuring error was less than 3%.

We calculate the *Phi* value of every sediment sample by using 51 sequences (Table 1), which represent the corresponding magnitude of various grain sizes. The data describe the percentage of the magnitude of each grain size accounting for the total magnitude of the grain size. Consequently, we constructed a dataset X with the 2141×51 matrix, where X_{ij} denoted the percentage composition of the j th grain size in the i th sample (Table 1).

4. Construction of Sample/Grain-Size Bipartite Network

In this paper, we construct the *Sample/Grain-Size* network based on the bipartite graph theory [18] in which the graph is denoted as $G = (V, E)$, where V is the node set and E represents the edge set. In the bipartite graph theory, V is divided into two disjoint subsets (A, B), where A is one class of nodes and B represents the other class; E denotes the association relationships between a node in the set A and a node in the set B .

According to this theory, the construction process of *Sample/Grain-Size* bipartite weighted network model is as follows.

In this process, one class, A , is the sample nodes and another class, B , denotes the grain-size nodes. As shown in Figure 2, the sample node numbered as Lz04-076 includes several grain-size nodes with the magnitude of 7.25–7.00. If the sample node includes a grain-size node, an edge will exist between this sample node and the corresponding grain-size

FIGURE 2: Construction process of the *Sample/Grain-Size* bipartite weighted network model.

FIGURE 3: The *Sample/Grain-Size* bipartite weighted network model.

node. The weight of edge denotes the number of grain sizes included in one sample. Based on this regular, we construct the final *Sample/Grain-Size* bipartite weighted network model as follows (Figure 3).

In this bipartite network, we identify the grain-size nodes as a green color, corresponding to the 51 class grain sizes with different magnitudes; we mark sample nodes as a pink color, corresponding to 2141 sets of samples. This model can clearly reflect the association relationships between the sample nodes and the grain-size nodes.

We construct a *Sample* network model projected from the *Sample/Grain-Size* bipartite network model (Figure 4). The *Sample* network model has 2141 nodes and 44,198 edges; a node denotes a sample and an edge represents that the two samples contain a grain size with the same magnitude. The weight of edge shows the frequency of the two samples having the same grain size.

5. Sediment Grain-Size Sample Analysis Based on Mini Batch K-Means

5.1. Idea of Sediment Grain-Size Data Analysis. In this paper, we cluster the *Sample* network model by the Mini Batch K-means algorithm. In the processing of every iteration time for the sediment samples, we randomly extract the mini batch subsamples from the total samples and update every mini batch sample by using the method of convex combination. At the same time, we use the per-center learning rates to increase the speed of the convergence rate. As the iteration times increased, we detect the convergence condition of this

FIGURE 4: The *Sample* network model.

algorithm when the clustering result has no change in successive iterations. In the end, we divide the sample nodes into several clusters.

5.2. Steps of Sediment Grain-Size Sample Data Analysis

- Step 1 Randomly extract b mini batch subsamples M from sediment sample dataset X with 2141 samples and 51 properties
- Step 2 Randomly select k samples as the initial clustering centers; save them into an array C storing k clustering centers which will be changed as the algorithm runs
- Step 3 Select a sample x from M ; calculate the clustering central sample node d_x having the nearest distance to the sample x by using Euclidean distance; save results in an array d . The Euclidean distance is as follows:
- Step 4 Acquire sample x and d_x ; update the per-center counter v :

$$f(C, x) = \min_{c \in C} \sqrt{\sum_{i=1}^{51} (x_i - c_i)^2}, \quad (7)$$

where $f(C, x)$ indicates the nearest Euclidean distance between the sample x and the central nodes in C . The i th property in the sample x is x_i

$$v[c] \leftarrow v[c] + 1. \quad (8)$$

- Step 5 Get the real-time per-center learning rates η , which speeds up the convergence of this algorithm

$$\eta \leftarrow \frac{1}{v[c]}. \quad (9)$$

- Step 6 Take the gradient step:

$$c \leftarrow (1 - \eta)c + \eta x. \quad (10)$$

- Step 7 If $M = \text{null}$, all the samples have been divided into a cluster, otherwise, return to step 4

Mini Batch K-Means

Input: the dataset of grain size is X ; the number of initial clusters k is 3; the iteration times is t ; the mini batch is b .

Output: the set of clustering labels is C ; the cluster label of every sample is c .
Initialize every sample label as $c \in C$.

$v \leftarrow 0$;

for $i = 1$ to t do

$M \leftarrow b$

//extract randomly b mini batch sub-samples M from X .

for $x \in M$ do

$d_x \leftarrow f(C, x)$;

//calculate and storage the clustering central sample d_x

end for

for $x \in M$ do

$c \leftarrow d_x$;

//acquire the central sample d_x

$v[c] \leftarrow v[c] + 1$;

//update the per-center counter v

$\eta \leftarrow \frac{1}{v[c]}$;

//get the real-time per-center learning rates η

$c \leftarrow (1 - \eta)c + \eta x$; //take gradient step

end for

end for

ALGORITHM 1: Pseudocode of Mini Batch K-means algorithm for sediment sample data processing.

$\epsilon - L1$: an ϵ - Accurate Projection to L1 Ball

Input: ϵ tolerance, L1-ball radius λ , vector $c \in \mathbb{R}^m$

if $\|c\|_1 \leq \lambda + \epsilon$ then exit

$upper \leftarrow \|c\|_\infty$; $lower \leftarrow 0$; $current \leftarrow \|c\|_1$

while $current > \lambda(1 + \epsilon)$ or $current < \lambda$ do

$\theta \leftarrow \frac{upper + lower}{2.0}$

$current \leftarrow \sum_{c_i \neq 0} \max(0, |c_i - \theta|)$

if $current \leq \lambda$ then $upper \leftarrow \theta$ else $lower \leftarrow \theta$

end while

for $i = 1$ to m do

$c_i \leftarrow \text{sign}(c_i) * \max(0, |c_i| - \theta)$

end for

ALGORITHM 2: Pseudocode of Mini Batch K-means algorithm for sediment sample data processing.

Step 8 If $iteration\ time \leq t$, return to step 1. The algorithm will stop when the convergence condition is satisfied or the $iteration\ time > t$

Algorithms 1 and 2 show the pseudocode of Mini Batch K-means algorithm for sediment sample data processing.

6. Simulations and Analysis

6.1. Multi-Index Analysis of Clustering Results. In this paper, we use the four indices AMI, NMI, completeness, and precision to verify the clustering result of sediment samples. We set the classical two parameters as *mini batch* and *reassignment ratio* in the Mini Batch K-means algorithm. After repeatedly testing 400 times, we acquire the corresponding results in Table 2. The maximum values in different

evaluation indices are marked as bold numbers: 0.40919072, 0.41485376, 0.44747697, and 0.92254367.

6.2. Heatmap Analysis of Clustering Results. The following heatmaps can objectively reflect the accuracy and efficiency of the clustering division of sediment sample data calculated by Mini Batch K-means algorithms. In each figure, every square represents the index score with different *mini batch* and *reassignment ratio* in a certain evaluation index. The various colors in the rightmost dashboard show the different scores, and the score range of every index is $[0, 1]$. The gradation of color in every square represents the size of the value.

As shown in Figures 5 and 6, the AMI can acquire the maximum value 0.40919072 when *reassignment ratio* $\epsilon = 0.025$ and *mini batch* = 25. The maximum value of NMI is 0.41485376 under the same parameters.

TABLE 2: Values of different parameters, *reassignment ratio* and *mini batch*, in different evaluation indices.

Evaluation indices	Mini batch Reassignment ratio	10	25	50	100	150	200	250	300	350	400
AMI	0	0.228121362	0.255117102	0.246499356	0.21748105	0.229442857	0.361703405	0.371481989	0.236873987	0.234965912	0.228301649
	0.001	0.373092772	0.324590503	0.234220515	0.239467411	0.22129395	0.229442857	0.233790795	0.220360264	0.375226487	0.218485929
	0.002	0.287940243	0.229004742	0.3374708	0.227533697	0.233574328	0.230282444	0.229905852	0.239052459	0.213798031	0.247221357
	0.005	0.220865936	0.379613278	0.295471433	0.382717495	0.244129445	0.227166056	0.257561299	0.223900056	0.229197683	0.224522543
	0.01	0.358589182	0.224356998	0.321456804	0.387328703	0.386584589	0.379123233	0.233574328	0.35333551	0.384838676	0.261293993
	0.015	0.210690796	0.257350101	0.368430163	0.233362852	0.372797418	0.231384	0.233574328	0.22848663	0.221189411	0.224619844
	0.02	0.370419215	0.384086717	0.225468744	0.286139496	0.233540603	0.232191714	0.22006892	0.385927006	0.365663322	0.230940806
	0.025	0.220486102	0.40919072	0.385226157	0.362571555	0.223848907	0.352126835	0.234905516	0.221068648	0.232103866	0.263270665
	0.03	0.363833758	0.404243026	0.398269621	0.365363502	0.235566424	0.232370173	0.361435437	0.353272529	0.298958527	0.2308378
	0.04	0.366158042	0.233655294	0.235924027	0.257480966	0.236366478	0.234551192	0.22955108	0.375532013	0.375173865	0.229442857
	0	0.275448684	0.315213955	0.2930531	0.263226806	0.277153665	0.389010747	0.392598757	0.285423463	0.284562758	0.275745815
	0.001	0.396829527	0.355098867	0.283576189	0.288149575	0.27021221	0.277153665	0.283336008	0.268524181	0.396568357	0.26721278
	0.002	0.29901197	0.277709568	0.368394245	0.274949128	0.282983141	0.278758694	0.27848551	0.287692232	0.260726088	0.301869173
	0.005	0.269156033	0.399910424	0.306001218	0.401001643	0.285609087	0.27094215	0.308190074	0.272768179	0.273623274	0.268337998
	0.01	0.385877697	0.267490458	0.351233084	0.407846087	0.40665482	0.399795843	0.282983141	0.378294211	0.405446733	0.312048714
	0.015	0.256394064	0.308650188	0.388822607	0.282637283	0.397536181	0.279861255	0.282983141	0.276049494	0.270618473	0.270277035
NMI	0.02	0.386885417	0.388617727	0.272823422	0.29787443	0.284060304	0.281414059	0.269993638	0.423281219	0.387117209	0.279373128
	0.025	0.266477018	0.41485376	0.407302793	0.389707983	0.269017064	0.377148115	0.284563458	0.261844075	0.280812201	0.314307297
	0.03	0.389384768	0.413418372	0.423393177	0.388052387	0.283876261	0.282750859	0.385267185	0.378678141	0.329722419	0.278498981
	0.04	0.391506815	0.283796705	0.283995333	0.268188818	0.286152928	0.284343865	0.277792995	0.39706664	0.399108033	0.277153665
	0	0.33010019	0.386892842	0.346038597	0.31607436	0.332290717	0.416626685	0.413233786	0.341441459	0.342128049	0.330553712
	0.001	0.420373957	0.386619802	0.340832226	0.344266003	0.327324404	0.332290717	0.340875352	0.324619118	0.417445629	0.324174705
	0.002	0.289428205	0.334248362	0.400318072	0.329746195	0.340339346	0.334935225	0.334812508	0.343764926	0.3153315	0.366071519
	0.005	0.325443853	0.419629785	0.296960563	0.433465142	0.331896127	0.320781419	0.366359566	0.329665976	0.324267515	0.318315311
Completeness	0.01	0.413484598	0.31654473	0.381913611	0.427796056	0.426115808	0.419928656	0.340339346	0.403249847	0.425500171	0.370265114
	0.015	0.309395011	0.367753577	0.408663479	0.339813063	0.422204195	0.33599622	0.340339346	0.331016253	0.328447547	0.322768397
	0.02	0.402440026	0.391672875	0.32760566	0.287621715	0.342984341	0.338563098	0.328582126	0.439773814	0.408136129	0.335462349
	0.025	0.319568317	0.44747697	0.428974907	0.417125266	0.320857984	0.402199137	0.342225591	0.30780224	0.337243024	0.372844734
	0.03	0.41499675	0.441883624	0.444262014	0.410444859	0.339618486	0.341517361	0.408948529	0.404160985	0.361737974	0.333513868
	0.04	0.416882784	0.342181432	0.339392468	0.258988328	0.343929226	0.342200924	0.333669453	0.418156627	0.422868505	0.332290717

TABLE 2: Continued.

Evaluation indices	Mini batch Reassignment ratio	10	25	50	100	150	200	250	300	350	400
Precision	0	0.656160673	0.320345633	0.450546474	0.218505371	0.43829986	0.420252219	0.202391406	0.446034563	0.297141523	0.43829986
	0.001	0.241709482	0.427986922	0.297786081	0.287473143	0.425408688	0.207547875	0.203035965	0.425408688	0.199813171	0.21463802
	0.002	0.205614199	0.660028024	0.890742644	0.437010743	0.203680523	0.205614199	0.297786081	0.286828585	0.290695936	0.431209715
	0.005	0.429276039	0.198524054	0.511779542	0.409294722	0.208836992	0.431854274	0.653582438	0.433143391	0.213993461	0.297786081
	0.01	0.366109295	0.210126109	0.230751985	0.921368519	0.253956095	0.369332088	0.297141523	0.599439514	0.366109295	0.273292854
	0.015	0.288762261	0.207547875	0.204325082	0.204325082	0.446034563	0.435721625	0.437010743	0.437655301	0.215282578	0.296496964
	0.02	0.355796357	0.750910789	0.209481551	0.510490425	0.299719757	0.204325082	0.422185894	0.427986922	0.206903316	0.436366184
	0.025	0.870798692	0.92254367	0.406071929	0.35901915	0.29843064	0.240420364	0.281027557	0.431854274	0.436366184	0.273937412
	0.03	0.211415227	0.422830453	0.191433909	0.20496964	0.288762261	0.299075199	0.210126109	0.370621205	0.340971509	0.441522653
	0.04	0.24235404	0.435077067	0.288117702	0.487930873	0.297141523	0.297141523	0.297786081	0.255889771	0.252022419	0.43829986

FIGURE 5: Heatmap of AMI scores.

FIGURE 6: Heatmap of NMI scores.

FIGURE 7: Heatmap of completeness scores.

FIGURE 8: Heatmap of precision scores.

TABLE 3: Clusters of sediment grain-size samples.

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

FIGURE 9: The clusters in the *Sample* network model.

Based on Figures 7 and 8, the completeness and precision can acquire the maximum values 0.44747697 and 0.92254367, respectively, when *reassignment ratio* $\epsilon = 0.025$ and *mini batch* = 25.

Through clustering analysis, we can assign these samples to their actual sediment clusters. Objectively, precision is the most significant index from these four performance evaluation indices. By analyzing the simulation results above, we can know that the clustering result of sediment grain-size samples calculated by Mini Batch K-means algorithm with appropriate parameters, $\epsilon = 0.025$ and *mini batch* = 25, has high precision: 0.92254367. The other three indices can also acquire maximum values: AMI = 0.40919072, NMI = 0.41485376, and completeness = 0.44747697.

6.3. Network Characteristic Analysis of Clustering Results and Comparison with Other Studies. We calculate the clustering result of sediment grain-size samples by using the Mini Batch K-means algorithm with the most appropriate parameters,

$\epsilon = 0.025$ and *mini batch* = 25. The simulation results are in Table 3 and Figure 9.

According to Table 3 and Figure 9, we divide the *Sample* network model into three clusters calculated by the Mini Batch K-means algorithm. Yi et al. divided the sedimentary environment of Lz908 through a variety of indices in the representative manuscripts [12, 16]. Compared with their results, the three clusters correspond three sedimentary environments: *marine*, *fluvial*, and *lacustrine*. The green cluster shows that these samples can be assigned to the *marine* sediment category; the orange cluster indicates that these samples can be split into *fluvial*; the blue cluster represents that these samples can be divided into *lacustrine*. This network can require a high precision, 0.92254367, of clustering division when the parameters of Mini Batch K-means algorithm are set as $\epsilon = 0.025$ and *mini batch* = 25. Furthermore, we find that most of the different points with precious studies are located at the junction of different sediment types (Figure 10). These results show that this method of analyzing sedimentary environment by using grain size is extremely effective and accurate.

7. Conclusions

During the last several decades, researchers have made significant advances in the environmental interpretation of grain-size analyses, but the definite criteria for environmental determination have not been given. Previous studies often overemphasized the subjective experience of the researcher and usually combined grain-size analysis with other methods and rarely used only grain size for sedimentary environment analysis. Recently, complex networks have been playing an increasingly significant role in data mining and knowledge discovery because they can reveal the potential relationship and concealed information between things. In this paper, we use complex networks and the bipartite graph theory to construct a *Sample/Grain-Size* network model and a *Sample* network model. Furthermore, we use the Mini Batch K-means algorithm to cluster the sediment grain-size samples.

FIGURE 10: Result comparison (profile of the Lz908 was reported in Yi et al.).

We use the representative evaluation indices AMI, NMI, completeness, and precision to verify the precision of the clustering results for the sample division. Simulation results show that this algorithm can divide the *Sample* network into three clusters—*marine*, *fluvial*, and *lacustrine*—a fact which is almost identical to the division in the classical manuscripts. At the same time, the evaluation indices can also acquire high values when we set the appropriate parameters as $\epsilon = 0.025$ and *mini batch* = 25. The results also denote that the

clustering results are efficient; for example, the samples that have the same classification with traditional method are up to 0.92254367, an excellent calculation result through a relatively convenient and low-priced way.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

We acknowledge the funding support from the China National Key Research Project (2016YFC0402801) and the National Natural Science Foundation (41406072).

References

- [1] W. Wei, J. Cai, X. Hu et al., "A numerical study on fractal dimensions of current streamlines in two-dimensional and three-dimensional pore fractal models of porous media," *Fractals*, vol. 23, no. 1, article 1540012, 2015.
- [2] T. Cordier, P. Esling, F. Lejzerowicz et al., "Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning," *Environmental Science & Technology*, vol. 51, no. 16, pp. 9118–9126, 2017.
- [3] Y. Jiang, Y. Gou, T. Zhang, K. Wang, and C. Hu, "A machine learning approach to Argo data analysis in a thermocline," *Sensors*, vol. 17, no. 10, p. 2225, 2017.
- [4] C. Zhou, K. Yin, Y. Cao et al., "Landslide susceptibility modeling applying machine learning methods: a case study from Longju in the Three Gorges Reservoir area, China," *Computers & Geosciences*, vol. 112, pp. 23–37, 2018.
- [5] J. Cai and B. Yu, "Prediction of maximum pore size of porous media based on fractal geometry," *Fractals*, vol. 18, no. 4, pp. 417–423, 2010.
- [6] P. Tahmasebi and A. Hezarkhani, "A hybrid neural networks-fuzzy logic-genetic algorithm for grade estimation," *Computers & Geosciences*, vol. 42, pp. 18–27, 2012.
- [7] M. Newman, "Community detection in networks: modularity optimization and maximum likelihood are equivalent," 2016, <http://arxiv.org/abs/1606.02319>.
- [8] J. Fu and J. Wu, "A deep stochastic model for detecting community in complex networks," *Journal of Statistical Physics*, vol. 166, no. 2, pp. 230–243, 2017.
- [9] F. Hu, Y. Zhu, Y. Shi, J. Cai, L. Chen, and S. Shen, "An algorithm Walktrap-SPM for detecting overlapping community structure," *International Journal of Modern Physics B*, vol. 31, no. 15, article 1750121, 2017.
- [10] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, article 026113, 2004.
- [11] F. Hu, M. Wang, Y. Wang, Z. Hong, and Y. Zhu, "An algorithm J-SC of detecting communities in complex networks," *Physics Letters A*, vol. 381, no. 42, pp. 3604–3612, 2017.
- [12] L. Yi, C. Deng, X. Xu et al., "Paleo-megalake termination in the Quaternary: paleomagnetic and water-level evidence from south Bohai Sea, China," *Sedimentary Geology*, vol. 319, pp. 1–12, 2015.
- [13] G. M. Friedman, "Differences in size distributions of populations of particles among sands of various origins: addendum to IAS Presidential Address," *Sedimentology*, vol. 26, no. 6, pp. 859–862, 1979.
- [14] G. M. Ashley, "Interpretation of polymodal sediments," *The Journal of Geology*, vol. 86, no. 4, pp. 411–421, 1978.
- [15] D. Sun, J. Bloemendal, D. K. Rea et al., "Grain-size distribution function of polymodal sediments in hydraulic and aeolian environments, and numerical partitioning of the sedimentary components," *Sedimentary Geology*, vol. 152, no. 3–4, pp. 263–277, 2002.
- [16] L. Yi, H.-J. Yu, J. D. Ortiz et al., "Late Quaternary linkage of sedimentary records to three astronomical rhythms and the Asian monsoon, inferred from a coastal borehole in the south Bohai Sea, China," *Palaeogeography, Palaeoclimatology, Palaeoecology*, vol. 329–330, pp. 101–117, 2012.
- [17] L. Yi, H. Yu, J. D. Ortiz et al., "A reconstruction of late Pleistocene relative sea level in the south Bohai Sea, China, based on sediment grain-size analysis," *Sedimentary Geology*, vol. 281, pp. 88–100, 2012.
- [18] K. Fukuda and T. Matsui, "Finding all the perfect matchings in bipartite graphs," *Applied Mathematics Letters*, vol. 7, no. 1, pp. 15–18, 1994.
- [19] M. Capó, A. Pérez, and J. A. Lozano, "An efficient approximation to the K-means clustering for massive data," *Knowledge-Based Systems*, vol. 117, pp. 56–69, 2017.
- [20] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: a k-means clustering algorithm," *Journal of the Royal Statistical Society Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.
- [21] K. P. Papadaki and W. B. Powell, "An adaptive dynamic programming algorithm for a stochastic multiproduct batch dispatch problem," *Naval Research Logistics*, vol. 50, no. 7, pp. 742–769, 2003.
- [22] L. Ana and A. K. Jain, "Robust data clustering," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*, Madison, WI, USA, June 2003.
- [23] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 9, article 09008, 2005.
- [24] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.
- [25] S. Romano, N. X. Vinh, J. Bailey, and K. Verspoor, "Adjusting for chance clustering comparison measures," *The Journal of Machine Learning Research*, vol. 17, pp. 4635–4666, 2016.
- [26] A. Rosenberg and J. Hirschberg, "V-measure: a conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 410–420, Prague, 2007.
- [27] T. Grinshpoun and A. Meisels, "Completeness and performance of the APO algorithm," *Journal of Artificial Intelligence Research*, vol. 33, pp. 223–258, 2008.
- [28] A. Biswas and B. Biswas, "Investigating community structure in perspective of ego network," *Expert Systems with Applications*, vol. 42, no. 20, pp. 6913–6934, 2015.