Journal of **Kim Il Sung** University (Natural Science) Vol. 3 No. 2 Juche103(2014)

Interpretation of the Comprehensive Geo-Anomalies by Geological Information Entropy

Pak Un Song

Abstract We select the geological information entropy as the unique criterion to predict comprehensive geo-anomaly, we have established the method to interpret the comprehensive geo-anomalies by combining both genetic algorithm and neural network.

Key words entropy, GNN, geo-anomaly

Introduction

Analyzing and determining the geo-anomalies are very important to predict the mineralization anomalous area. The geo-anomaly can not always indicate the ore deposit but mineralization are surely regarded as geological anomaly. The geo-anomaly will become the essential condition to predict new ore deposit [5, 8].

With the development of GIS technique, prediction of prospecting area have been developing in the way of interpreting comprehensive geological anomaly to combine geophysical and geochemical information based on the prediction of geo-anomaly in study area [2, 3, 7, 9].

These methods were used to analyze comprehensive geo-anomalies included linear regression model, exponential overlay model and weighted evidence model [1, 4]. In addition, in the branch of geology, researches for application of non-linear characteristics to predicted problems by combining neural network and genetic algorithm [4].

1. Information Entropy Calculation

For the sake of this, the study area is divided into the cells. In every cell, information entropy can be calculated as the following formula.[6]

$$H_r = -\sum_{i=1}^{N} P_i \ln P_i \Big/ H_m \tag{1}$$

where H_r is information entropy and N is the number of states(in here, the number of parts within every cell), P_i is the probability to be able to present i^{th} state of N states, H_m is maximum information entropy, $H_m = \ln N$.

In the above expression P_i could be calculated as this follows.

$$P_i = A_i \Big/ \sum_{j=1}^{N} A_j \tag{2}$$

where A_i is the value of i^{th} state and $A_j - j^{th}$.

A_j could be addressed as the area of the part separated within the cells.

When the area is divided into cells, it is very important to determine the size, shape of the cells and the number of the part in a cell.

First, information entropy depends on the size of cell. If the size of cell is converged to 0, the average information entropy will be 0 and if the size of cell is larger than the constant size, the average information entropy will be converged to the constant. In this case, the constant is related to *N*(the number of states separated within cells).

And then information entropy also depends on the shape of cell.

Therefore we did determine the suitable size, shape of cell and the number of state.

The curves of entropy according to the number of state are shown in Fig. 1.

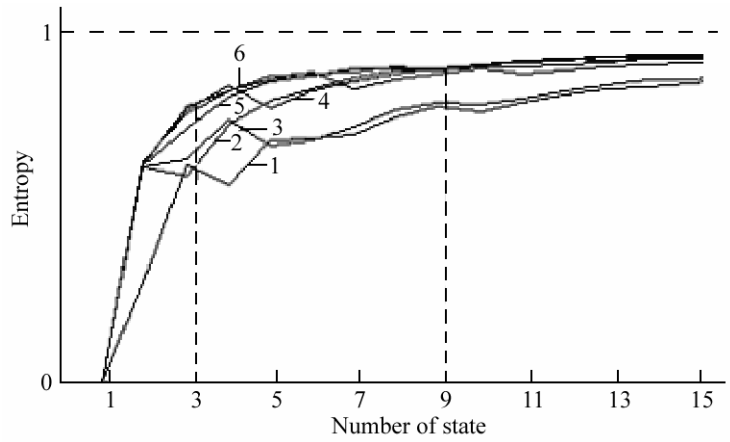

Fig. 1. Curves of entropy according to *N*
1－3, A_j is random from 0 to 1, 4－6, A_j is from 20 to 100

As shown in Fig. 1, the entropy is 0 if the number of the part in each cell is equal to 1 and when the number of the part in each cell is 1 to 3, the entropy will be gradually increased and also if the number of the part in each cell is larger than 3 to 9, it shows well the feature in the variation of entropy. Thus, it is suitable to set the number of the part in each cell as 3 to 9.

And then the shape and size of cell should be determined by regarding the distribution of the geological bodies because the value of information entropy is changed according to shape and size of cells.

2. Threshold to Predict the Information Entropy Anomaly

Threshold can be expressed as following.

$$B(i) = \begin{cases} 1 & E(i) > F \\ 0 & E(i) < F \end{cases}, \quad i = \overline{1, N} \tag{3}$$

where if $B(i) = 1$, it implies the anomaly, otherwise not. $E(i)$ is information entropy of i^{th}

cell and F is a threshold, N is the number of grids.

F could be calculated like this.

$$F = \overline{E} + E_s \tag{4}$$

$$\overline{E} = \frac{1}{N}\sum_{i=1}^{N} E(i) \tag{5}$$

$$E_s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[E(i) - \overline{E}]^2} \tag{6}$$

where $\overline{E}$ is mean, E_s is the standard deviation.

Feature of entropy distribution in the study area could be determined by both the mean and the standard deviation. So anomaly range can be determined by the mean adding the standard deviation.

3. Entropy Calculation of the Several Geological Information

3.1. Information entropy calculating method of geological map

First, geological map should be digitized.

By digitalizing information of layer and intrusive body become polygon and fault becomes line.

Then geological map is divided into cells and then every cell has its number.

Then entropy of layer, intrusive body and fault within every cell should be calculated by the expression (1).

3.1.1. Entropy calculating method of layer, intrusive body

① Overlay intersection

It must determine the boundaries of the area which should be separated, such as the coordinates of the bottom left-hand and top right-hand corner of the area. Then it has to determine the number of cells. And the boundary and the coordinates of the bottom left-hand cell should be calculated. And then the boundaries and coordinates of the next cell can be calculated from the first cell. Like this every cell should be determined. Then it should generate the polygons with the boundary and known-vertexes coordinates. And then it could be performed to overlay intersection. This process is as follows.

It should extract the polygons of layer and intrusive body intersected with every cell, included within every cell.

It should find the points that the boundary of cell is intersected or included with the layer or intrusive body within every cell.

New polygons of layer or intrusive body in every cell could be reconstructed by these points. The attributes of new polygons area are as following (table 1).

Table 1. Property of layer, intrusive body

Shape	No.	Code
polygon	1	1011
polygon	2	415
polygon	3	225
⋮	⋮	⋮
polygon	*N*	1325

As shown in this table, if the code can be expressed as 3 digits, it implies the layer and if the code can be expressed as 4 digits, it implies the intrusive bodies.

Every cell consists of the some different polygons. If separated parts in every cell have the same code, these are the same layer or intrusive body.

② It should calculate the area both of every cell and of separated parts in every cell.

③ Information entropy of layers and intrusive bodies in every cell should be calculated.

3.1.2. Information entropy calculating method of the fault

① Buffering fault

It must find the faults that controlled the formation and distribution of ore deposits. Then statistics of distance between known deposits and faults have to be done i.e. is the statistical histogram on the distance between known deposits and faults. The number of deposits tends to be increasing and then decreasing according to the distance away from the deposits. Buffer zone for the faults could be determined through the analysis on the statistical histogram on the distances. As you see in Fig. 2, buffer distance of the faults can become the distance responding to maximum sequence on this histogram.

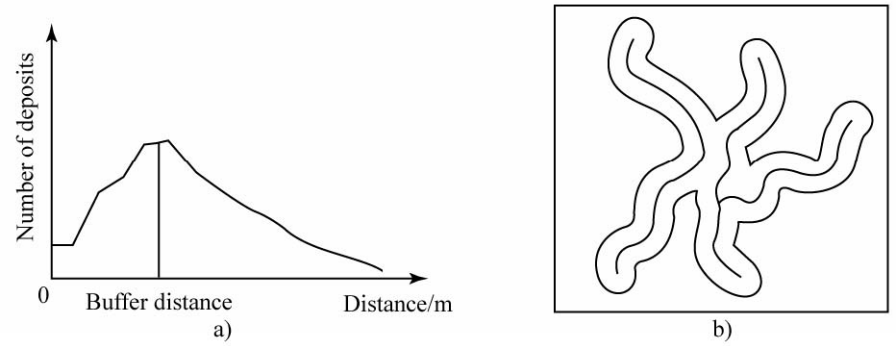

Fig. 2. Calculation of buffer distance of faults
a) buffer distance, b) buffering fault

② Overlay intersection

The features of faults are as table 2.

Table 2. Features of faults

Shape	No.	Code
polygon	1	0
polygon	2	1
polygon	3	0
⋮	⋮	⋮
polygon	N	0

As you see in the table 2, the code of buffer zone for the faults is 1 and otherwise is 0. It should calculate the area both of every cell and separated parts in every cell.

③ Information entropy of faults in every cell should be calculated.

Finally the information entropy of the geological map could be obtained through the integration of the information entropy of layer, intrusive body and fault using GNN.

3.2. Entropy calculating method of geochemical information

First of all the geochemical data could be processed by removing anomalous value according to rock unit, then it must determine standard value for geochemical anomaly.

The advantage of this method is that it considered the difference according to type of

rock. And then it can be able to determine accurately the standard value due to removing the anomalous data. And it could be determined standard value without regarding the distribution of element in the geological body.

Then study area is divided into cells and then every cell has its own number.

Next, contour map on the geochemical data should be drawn.

And then, the number of contour line in a cell could be counted and the areas of the parts separated within a cell were calculated.

Finally, entropy of the geochemical information in every cell should be calculated and then it should be separated the anomaly by the threshold.

3.3. Entropy calculating method of geophysical information

The region anomaly should be removed from the observed geophysical field.

Then study area is divided into cells and then every cell has its own number.

Next, contour map on the geophysical data should be drawn.

And then, the number of contour line in a cell could be counted and the areas of the parts separated within a cell could be calculated.

Finally, entropy of the geophysical information in every cell should be calculated and then it should be separated the anomaly by the threshold.

4. Construction of GNN and Analysis Algorithm of the Comprehensive Geo-Anomaly by Geological Information Entropy

It implies that geological information entropy could be nearly about 1 in the comprehensive geo-anomaly zone and if it may be converged to 0, not anomaly zone.

Since the distribution of initial population is not surely known in GNN, generally, it is hard to exhibit the global finding ability and searching ability for GNN. Also performing GNN may be limited to the local solution as an individual with the same fitness can be chosen due to randomize selection of GNN.

It is necessary to improve the training ability in order to interpret the comprehensive geo-anomaly efficiently. Because various geological factors could be influenced differently to form the anomaly and the relationship between these factors is so complex, has the complex non-linear relationship each other. So the propagated weights of each cell within neural network have to be modified so that the training ability of GNN is increasing.

4.1. GNN

This could be improved in 3 sides.

① Each weight for the initial population should be generated randomly by the probability distribution, $e^{-|2x|}$, so that one of this population have to be distributed uniformly by Hamming distance in the space for the solution.

② It improved the velocity of convergence and the earlier convergence for genetic algorithm by combining double selection, the protection of the best individual and the standard

proportional selection operator.

③ The individuals with the larger fitness than the others are selected and prevent the inbreeding in every generation.

By comparing the fitness between two individuals for crossover, then if one's fitness is the same as the other, the other can be selected. If the number of selections are exceeded in a threshold, an individual can be forced to select in previous generation.

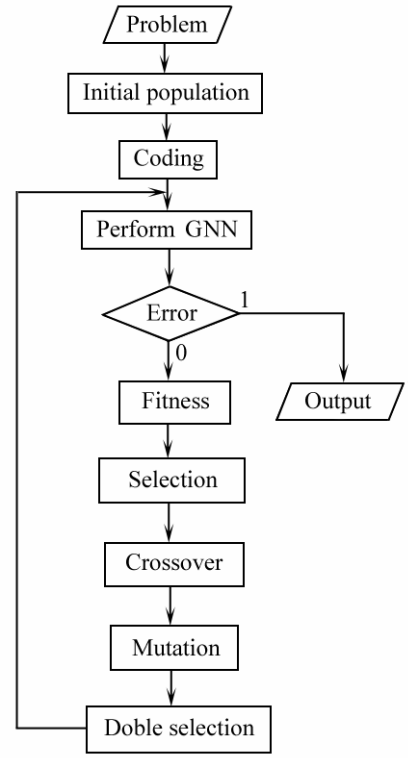

Fig. 3. Algorithm for performance of GNN

Performing diagram for GNN is as Fig. 3.

① Initial population

Hamming distance could be defined by

$$C_{ij} = \sqrt{\sum_{k=1}^{M} (X_{ik} - X_{jk})^2} \tag{7}$$

where C_{ij} is the Hamming distance between X_{ik} and X_{jk}. M is the length of chromosome, X_{ik} and X_{jk} are two series of digits respectively.

② Coding

An individual of GNN is just a single neural network and the parameters of it are the connective weight between the neural cells within neural network.

A chromosome could be constructed by arranging each weight of the neural network according to a certain order.

③ Performance of GNN

The structure of the neural network is related to the practical problems. So it has to determine based on the essential of problems and experience.

Error function could be applied to the mean square error deviation and the size of the initial population and the number of the training samples.

The convergence of the multilayer neural network based on genetic algorithm depends on the difference between the ideal output and the real one according to input of GNN.

④ Fitness function

$$\mathrm{MSE} = \frac{\sum_{k=1}^{M} [y(k) - \hat{y}(k)]^2}{m} \tag{8}$$

where $y(k)$, $\hat{y}(k)$ are respectively the ideal output and real one of GNN m is the number of input-output sample.

To prevent the early convergence, fitness function could be controlled linearly.

⑤ Selection

To improve the selection operator, the protection strategy of best individual is introduced.

That is, best individuals, 10% of the previous population are chosen by means of a "selection" operator and form a second population of offspring. The rest, 90% of the previous population, could be arranged in ascending order according to the magnitude of the fitness function. According to a prefixed probability, pairs of genotypes produce offspring via crossover, mutation.

⑥ Crossover, Mutation

When the population is going to converge to the local solution at the early step of the generation via crossover, mutation, we should disrupt the stability of the population so that the variety of the population is increased. On the other hand, when the population is going to diverge at the late step of the generation variety of the population should be decreased so that best individual is protected.

⑦ Double selection

At this step the population consists of the parent population and new population via selection, crossover and mutation.

Double selection occurs in that population to ensure that the fitter members of the population are selected.

⑧ Until the error could be acceptable, it should be repeated step ②－⑦.

⑨ It should store the parameters of GNN and then could use it to interpret the comprehensive geo-anomaly.

First advantages on this GNN are that this model has enough objecvity and the various geological information entropies could be integrated non-linearly by this one.

Verifying the effectiveness of GNN, we compared this GNN with linear regression model. Sample data used this comparison are shown at table 3. And the results from both GNN and linear regression model are as table 4.

Table 3. Sample data

No.	Layer	Fault	Intrusive body	Pb	Zn	Gravity	Magnetic force	Expected value
1	0.865 48	0.762 31	0.850 72	0.895 5	0.808 98	0.889 85	0.934 09	1
2	0.710 87	0.750 0	0.850 00	0.849 65	0.747 59	0.954 72	0.856 14	1
3	0.750 16	0.750 0	0.750 75	0.764 2	0.866 21	0.925 79	0.883 00	1
4	0.950 69	0.815 03	0.857 12	0.957 96	0.955 77	0.856 81	0.894 28	1
5	0	0	0	0	0	0	0	0
6	1	1	1	1	1	1	1	1

Table 4. Results from both GNN and linear regression model

No.	Output		Absolute error	
	Linear model	GNN	Linear model	GNN
1	0.83	0.99	0.17	5.02×10^{-3}
2	0.25	0.99	0.75	4.30×10^{-3}
3	0.66	0.99	0.34	5.28×10^{-3}
4	0.78	0.99	0.21	4.99×10^{-3}
5	0.63	0.003	0.36	1.39×10^{-2}
6	0.91	0.99	0.09	4.85×10^{-3}

As shown in table 4, GNN has the smaller absolute error more than linear regression model.

And then the convergence of this GNN could be improved more than normal one.

Because we determined the condition of distribution of the initial population so that the initial population have to be distributed uniformly by Hamming distance in the space for the solution. Also the population of each generation included some individuals with the fitter fitness so that best individual could be protected. And then comparing the fitness of the individual during crossover, then it prevents the inbreeding between individuals in every generation. So we compared this GNN with normal GNN in order to verify that. The results trained by both this GNN and normal GNN are as follows. The characteristics of convergence for it are shown in table 5.

Table 5. Characteristics of convergence with the number of geological factors

No.	Number of geological factor	Normal GNN			Improved GNN		
		Number of training	Error	Convergent rate/%	Number of training	Error	Convergent rate/%
2	4	23	5×10^{-7}	100	60	0	100
3	6	226	3.23×10^{-3}	98.7	160	2.39×10^{-11}	100
4	9	289	8.58×10^{-2}	95.3	180	1.38×10^{-1}	100
5	11	388	2.55×10^{-1}	88.0	230	1.11×10^{-3}	100
6	12	442	8.29×10^{-1}	80.1	350	2.30×10^{-5}	100

As shown in table 5, this GNN shows the increment clearly in both the convergent rates and errors rather than normal GNN.

4.2. Algorithm to analyze the comprehensive geo-anomaly

This algorithm for analysis of comprehensive by geological information entropy and GNN is as follows.

① Geological information used to analyze the comprehensive geo-anomaly should be selected.

② Then study area is divided into cells and then every cell has its own number.

③ Entropy of the various geo-information such as the information of the geological map, geophysical information and geochemical information within every cell should be calculated.

④ In GNN, the parameter of it should be initialized. The parameter include the number of the initial population, the number of chromosome, the number of cells of input layer in neural network, the number of hidden layers, the number of connective weight in neural network and the number of training sample. If the structure of neural network can be confirmed, each weight for the initial population and threshold should be generated by the probability distribution $e^{-|2x|}$.

⑤ GNN should be performed.

Training samples are trained by GNN. If the precision of GNN was approximate to the needed value, stop performing GNN and then the structure of GNN would be stored.

⑥ Given data should be entered into GNN then output is obtained.

Entered data consist of the information entropy of geological map, the entropies of the

geochemical information and of the geophysical information.

⑦ The average value and the standard deviation have to be calculate the whole output data.

⑧ It should draw the contour map by the whole output data and then interpret the comprehensive geo-anomaly.

Conclusion

We established the methods to calculate the information entropy of the geological map, geophysical and geochemical information.

We established the approach of the analysis of the comprehensive geo-anomaly by both geological information entropy and GNN.

Comprehensive geological information entropy is calculated by integrating the entropy of the different geological information by GNN. In conclusion, we can say that the greater the area has the comprehensive geological information entropy, the more the area has the probability to be a comprehensive geo-anomaly.

References

[1] L. Saro et al.; Engineering Geology, 3, 289, 2004.

[2] J. P. Lacassie et al.; Sedimentary Geology, 1, 175, 2004.

[3] L. Guangren et al.; Marine and Petroleum Geology, 3, 411, 2004.

[4] H. Gaamez et al.; Engineering Geology, 1, 11, 2005.

[5] P. Zhao et al.; Journal of China University of Geosciences, 12, 2, 108, 2001.

[6] F. Zhu et al.; Geoscience and Remote Sensing Letters, 7, 1, 151, 2010.

[7] 黄海峰 等; 甘肃地质学报, 11, 1, 89, 2002.

[8] 卜淘 等; 华东地质学院学报, 24, 3, 186, 2001.

[9] 赵鹏大 等, 地质科学情报, 19, 2, 99, 2000.