Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2020.010129
Article

Tech Science Press

An Apriori-Based Learning Scheme towards Intelligent Mining of Association Rules for Geological Big Data

Maojian Chen[1,2,3], Xiong Luo[1,2,3,*], Yueqin Zhu[4], Yan Li[1,2,3], Wenbing Zhao[5] and Jinsong Wu[6]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, 100083, China
[2]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, 100083, China
[3]Beijing Intelligent Logistics System Collaborative Innovation Center, Beijing, 101149, China
[4]Development and Research Center, China Geological Survey, Beijing, 100037, China
[5]Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, Ohio, 44115, USA
[6]Department of Electrical Engineering, Universidad de Chile, Santiago, 1058, Chile
[*]Corresponding Author: Xiong Luo. Email: xluo@ustb.edu.cn

Abstract: The past decade has witnessed the rapid advancements of geological data analysis techniques, which facilitates the development of modern agricultural systems. However, there remains some technical challenges that should be addressed to fully exploit the potential of those geological big data, while gathering massive amounts of data in this application field. Generally, a good representation of correlation in the geological big data is critical to making full use of multi-source geological data, while discovering the relationship in data and mining mineral prediction information. Then, in this article, a scheme is proposed towards intelligent mining of association rules for geological big data. Firstly, we achieve word embedding via word2vec technique in geological data. Secondly, through the use of self-organizing map (SOM) and K-means algorithm, the word embedding data is clustered to serve the purpose of improving the performance of analysis and mining. On the basis of it, the unsupervised Apriori learning algorithm is developed to analyze and mine these association rules in data. Finally, some experiments are conducted to verify that our scheme can effectively mine the potential relationships and rules in the mineral deposit data.

Keywords: Association rules; Self-organizing Map (SOM); K-means; apriori

1 Introduction

Data analysis method as an effective technique has been excessively employed in the era of big data [1–5]. During the last two decades, it drives many applications [6–12]. Specifically, while applying those methods to agricultural upgrading and reconstruction, it can not only accelerate the process of agricultural modernization, but also play an important role in realizing sustainable development [13]. Since the formation and development of agriculture is partly related to the evolution of surface geology, the geological data analysis approaches play an active role in supporting modern agriculture system [14,15]. While collecting and gathering a large amount of geological data in this field, it is challenging to analyze and utilize geological big data, and it also imposes an obstacle to the wide applications in modern agricultural system. In response to such limitation, we specifically conduct an exploration of using machine learning algorithms to achieve data-enabled intelligent analysis for geological big data in this article.

Generally, geological big data relates to the various layers of the Earth, the history about the formation and evolution of the Earth, the material composition and changes of the Earth, and many others [16,17]. Then, geological big data has the characteristics of "4V" of traditional big data, that is, volume, variety, velocity, and value. Moreover, it also has its own particularities, such as multi-source heterogeneity, spatial-

temporal correlation, and complexity fuzziness. Meanwhile, due to the huge space-time scope of geological object development and evolution, and the numerous factors affecting geological processes, the geological characteristics of high dimension, high complexity, and high uncertainty are more significant, which make the geological big data face unprecedented opportunities and challenges [18].

For geological big data, it is a challenging and important issue of extracting valuable information from these multi-source data in consideration of its complex characteristics, so as to analyze the mineral yield regularity, summarize the characteristics of a specific type of mineral deposit, and discover the attributions that may be included [19]. Hence, it is particularly difficult to mine association rules from geological big data.

In this field, more emphasis is being placed on using Apriori algorithm to mine association rules for geological big data [20–22]. An algorithm was proposed to convert the spatial data to non-spatial one in the geographic information systems (GIS) database, while using the Apriori algorithm to explore a multi-level multi-relational space association rule mining method based on inductive logic programming [23]. Based on multi-source geological spatial database and spatial data mining technology, considering the spatial characteristics and uncertainty of geological data, a regional metallogenic prediction method was developed on the basis of geological spatial data mining, where the experimental comparison results showed that the prediction results based on geological spatial data mining using Apriori algorithm were more accurate than the traditional evidence weight model method, and the method was effective for regional mineralization prediction [24]. Furthermore, based on the Apriori algorithm, the frequent itemsets of the associated ore and intrusive rocks of hydrothermal gold deposits were extracted, and it was found that the associated minerals were closely related to the acidity and alkalinity of the intrusive rocks [25].

However, for those work mentioned above, when the amount of data is too large, the developed algorithms may have the following disadvantages.

(1) If the degree of support or confidence is too low, Apriori algorithm will generate many useless rules that prevent users from quickly distinguishing and judging these rules, making it difficult for users to find truly useful knowledge.

(2) On the contrary, if the degree of support is too high, it will ignore some valid and strong association rules.

(3) When the number of database scans is too large or the frequent itemsets that need to be searched are too large, the algorithms will consume too much time or memory.

In order to avoid such limitations and design a more practical method, this article combines self-organizing map (SOM) and K-means with Apriori algorithm to effectively mine more valid and strong associated rules, especially for geological big data. The motivation of this scheme proposed here is shown as follows.

Currently, SOM as an effective unsupervised learning algorithm can be used in clustering. However, the number of clustering results obtained through SOM is particularly large, then the K-means algorithm could be used to perform secondary clustering on the data. K-means is a classic clustering algorithm [26]. Its core idea is to first determine the number of clusters, and then divide all data into clusters with the pre-defined number, according to the Euclidean distance. But the selection of the initial clustering center of the K-means algorithm has a great influence on the clustering results, and it is easy to fall into the local optimum. Specifically, it is sensitive to “noise” and isolated point data, and these defects greatly limit the clustering effect. By using SOM clustering as the input of K-means clustering, and using the K-means algorithm to perform secondary clustering on the results of SOM clustering, it is expected to overcome the above defects, and the data can be divided accurately according to the specified number of clusters [27,28]. Moreover, through the combination of SOM and K-means, it is able to improve Apriori algorithm [29,30].

Motivated by it, a SOM-K clustering-optimized unsupervised Apriori learning algorithm is developed to mine association rules in each category for geological big data. This algorithm is named as SOM-K-Apriori. In this article, we specifically combine “Support”, “Confidence”, and “Lift” of association rules in Apriori as the evaluation criteria to further facilitate this scheme to find more valid and strong association

rules. Furthermore, to effectively show the mining results, the association rules are clearly demonstrated by tables and parallel visualization method.

The contributions of this article are as follows:

(1) Aiming at the practical demand for geological big data analysis, a machine learning-based intelligent mining scheme for association rules is accordingly developed, which greatly improves computational performance of mining task.

(2) In the field of geological data analysis, through the use of unsupervised learning algorithm SOM-K-Apriori, the proposed scheme can find more valid and strong association rules with low support degree, while providing an intuitive visualized result of those association rules.

The rest of this article is arranged as follows. Section 2 will introduce the background, including SOM and Apriori algorithm. In Section 3, we detail the proposed scheme. In Section 4, the experiments of mining association rules for mineral deposit data are conducted to evaluate the performance of our proposed scheme. Finally, the conclusion is summarized in Section 5.

2 Background

In this section, we will simply introduce some key technologies in relation to our method.

2.1 Self-Organizing Map (SOM)

SOM was proposed by Kohonen, and it was also known as the Kohonen network [31]. It is one of the unsupervised learning methods. The main task of SOM is to convert input data of any dimension into one-dimensional or two-dimensional discrete data through computational mapping. Currently, it can be used for many applications, such as clustering, high-dimension visualization, data compression, and feature extraction [32,33].

Essentially, SOM is a neural network with only an input layer and an output layer. The input layer receives external input information and the output layer responds to them. The data of the input layer can be any dimension, and each node in the output layer represents a class that needs to be clustered. Competitive learning is implemented in training phase. Each input sample finds the node that is most similar to it in the output layer, called the active node or the winning neuron. Then, an appropriate method is executed to update the parameters of the active node. Meanwhile, the neighbor nodes have also updated the parameters appropriately based on the distance from them to the active node.

Therefore, there is a topological relationship between nodes in the output layer. This topological relationship needs to be determined manually. When their needs a one-dimensional model, the output nodes are connected in a line. Otherwise, when a two-dimensional topological relationship is needed, the output nodes are connected to form a plane. For example, the network structure of two-dimensional topological relationship is shown in Fig. 1, where the nodes of the output layer are fully connected to the nodes of the input layer.

2.2 Apriori Algorithm

The Apriori algorithm was proposed in [34]. It is now a classical algorithm in association analysis and mining, and it is used to find out the datasets that appear frequently in data values [35–37]. Finding patterns for these frequent sets can help us make some decisions [21,38].

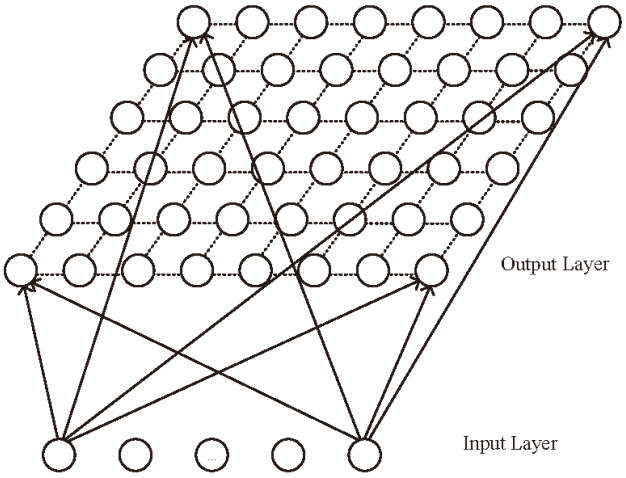

Figure 1: The network structure of SOM

Some basic concepts about Apriori are introduced as follows:

- Item and Itemset

Let itemset be a collection of all items:

$$\text{itemset} = \{\text{item}_1, \text{item}_2, \cdots, \text{item}_k\}, k = 1,2, ..., \quad (1)$$

where item_k is an item. The set of items is called an itemset, and the itemset containing k items is called a $k-$itemset.

- Association Rules

Association rules are implications of form $A \to B$, where A and B are both subsets of itemset and not empty sets, and $A \cap B = \emptyset$.

- Support

$$\text{support}(A \to B) = P(A \cap B), \quad (2)$$

where $P(A \cap B)$ represents the probability of an itemset containing set A and set B.

- Confidence

$$\text{confidence}(A \to B) = P(B|A) = \frac{\text{support}(A \cap B)}{\text{support}(A)} = \frac{\text{count}(A \cap B)}{\text{count}(A)}, \quad (3)$$

where $\text{count}(\cdot)$ is the count of the itemset.

- Lift

$$\text{Lift}(A \to B) = \frac{P(B|A)}{P(B)}. \quad (4)$$

If $\text{Lift}(A \to B) > 1$, it means that the rule $A \to B$ is a valid and strong association rule. Inversely, $\text{Lift}(A \to B) < 1$ means that the rule $A \to B$ is an invalid and strong association rule. Specifically, $\text{Lift}(A \to B) = 1$ represents that A and B are independent.

3 The Proposed Scheme

In this section, we will present each step of the scheme we proposed. The ultimate goal is to effectively mine the valid and strong association rules between mineral deposit attributions and visualize them.

The core processing process can be divided into four parts. Firstly, all geological data are expressed in the form of word embedding vector through word2vec technique. Secondly, the SOM algorithm maps the vector data in a two-dimensional space and puts similar data in adjacent locations. However, the number of clustering results of SOM is relatively large and the results may be unsatisfactory, we use the K-means method to further cluster. Thirdly, the Apriori algorithm is incorporated to analyze and mine the association

rules in each category of geological data. Finally, the evaluation criterion is designed through the combination of "Support", "Confidence", and "Lift" in Apriori, to evaluate the quality of results achieved by SOM-K-Apriori method. Especially, the whole model of SOM-K-Apriori is shown in Fig. 2.

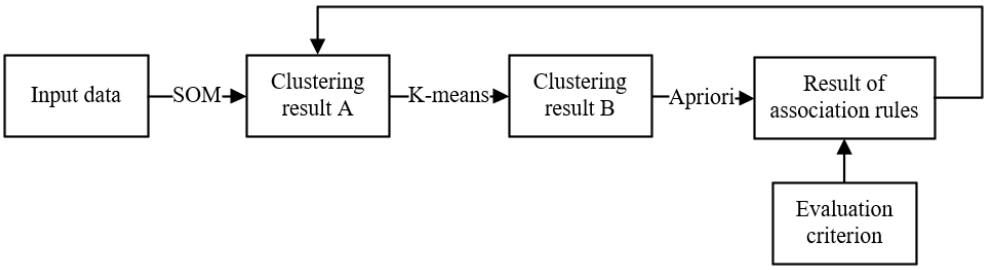

Figure 2: SOM-K-Apriori model

3.1 Preprocessing for Geological Data

First, the abnormal data, including checking data consistency, handling invalid values, and missing values, are cleaned up. Then, some key attributes of the data are analyzed through simple statistical methods, to find out the evolution trend of the attributes.

3.2 Word Embedding Using Word2vec

Since there are so many text data in geological data, in order to normalize those data, we use word2vec model to process dataset to achieve word embedding. Here, for those massive data, we convert each attribute value to a 50-dimensional word embedding vector before mining association rules.

3.3 Association Analysis via SOM-K-Apriori Algorithm

After getting vectors for all data, we use the mode SOM-K-Apriori to cluster data and mine association rules.

Firstly, the topological relationship of SOM network should be determined. Let the number of input data be N. Each input data vector is $\boldsymbol{x} = [x_1, x_2, \cdots, x_D]^{\mathrm{T}}$ with D-dimension. For each node in the output layer, its dimension is the same as the dimension of the input. Thus, the weight vector of each output node j is recorded as $\boldsymbol{w}_j = [w_{j1}, w_{j2}, \cdots, w_{jD}]^{\mathrm{T}} (j = 1,2, \cdots, L)$, where L is the number of output nodes.

1) Initialization

Each weight vector of output node is initialized with a smaller random value.

2) Competition

The competitive process is to find the weight vector $\boldsymbol{w}_j$ that is most similar to the vector $\boldsymbol{x}$. We can use the Euclidean distance as the discriminant function. Then, the smaller the Euclidean distance, the more similar the vector $\boldsymbol{x}$ is to the weight vector $\boldsymbol{w}_j$. The index $i(\boldsymbol{x})$ which symbolizes the output node that is most similar to the input vector $\boldsymbol{x}$ can be expressed as follows.

$$i(\boldsymbol{x}) = \arg\min |\boldsymbol{x} - \boldsymbol{w}_j|, j = 1,2, \ldots, L. \tag{5}$$

The output node j that satisfies (5) in the output layer is called the active node or winning neuron, which is most similar to the input vector $\boldsymbol{x}$.

3) Cooperation

Let $h_{j,i}$ be a set of excited nodes in the output layer that are affected by the active node i, where j represents the node number in the output layer. Moreover, $d_{j,i}$ represents the distance between the active node i and the excited node j, then $h_{j,i}$ is a unimodal function, which is related to the distance $d_{j,i}$. The

smaller the distance between the active node and the excited node, the greater the impact on the excited node. Thus, $h_{j,i}$ can also indicate the measure which the excited node is affected.

Generally, $h_{j,i}$ is a Gaussian function as follows:

$$h_{j,i(\boldsymbol{x})} = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2}\right), \tag{6}$$

where $i(\boldsymbol{x})$ is the location of the active node, and σ is the effective width of the topology neighborhood. In addition, σ is expressed as follows.

$$\sigma(n) = \sigma(0)\exp\left(-\frac{n}{\tau_1}\right), n = 1,2,3, ..., \tag{7}$$

where $\sigma(0)$ is the initial value of σ, and τ_1 is a time constant. Then, $h_{j,i}$ can be defined as follows:

$$h_{j,i(n)} = \exp\left(-\frac{d_{j,i}^2}{2\sigma(n)^2}\right),\ n = 1,2,3, ..., \tag{8}$$

Usually, the initial value $\sigma(0)$ is the radius of the output mesh, and the time constant is $\tau_1 = 1000\log(\sigma(0))$.

4) Weight update

The weight vectors of the active node and its surrounding excited nodes are adjusted by the gradient descent method.

$$\boldsymbol{w}_j(n+1) = \boldsymbol{w}_j(n) + \eta(n) \times h_{j,i(n)} \times \left(\boldsymbol{x} - \boldsymbol{w}_j(n)\right), \tag{9}$$

where $0 < \eta(n) \leq 1$ is learning rate defined by:

$$\eta(n) = \eta(0)\exp\left(-\frac{n}{\tau_2}\right), n = 1,2,3, ..., \tag{10}$$

where $\eta(0)$ is the initial value of η, and τ_2 is an another time constant.

5) Iteration

Steps 2), 3), and 4) are repeated until $\boldsymbol{w}_j$ no longer changes significantly.

Then, we use K-means algorithm to further cluster. Let the data vector mapped through SOM be x_i (i=1, 2, …, N), and it has the same dimension with the input data of SOM, where N is the total number of data. Moreover, $C = \{c_1, c_2, \cdots, c_K\}(K \leq N)$ is a set of K cluster. Subsequently, we can select K cluster center as follows.

1) The initial K cluster centers are randomly selected as $\{z_j | z_j \in \mathrm{R}^d, j = 1, \ldots, K\}$.

2) The distance between data x_i and each cluster center z_j is calculated, and then x_i is assigned to its nearest cluster center. The distance can be expressed as follows:

$$D = \arg\min_j \sum_{i=1}^{N}\sum_{j=1}^{K}\left\|x_i - z_j\right\|^2. \tag{11}$$

3) For each cluster j, the new cluster center is calculated again.

$$z_j = \left(\sum_{i=1}^{N} r_{ij}x_i\right) \cdot \left(\sum_{i=1}^{N} r_{ij}\right)^{-1}, \tag{12}$$

where $r_{ij} = 1$ when the data x_i belongs to the j-th cluster, otherwise $r_{ij} = 0$.

4) Steps 2) and 3) are repeated, until the cluster centers stay the same.

Thirdly, we use the Apriori algorithm to mine the association rules in data. The Apriori algorithm uses an iterative method called layer-by-layer search, where the k itemset is used to explore the $(k+1)$ itemset. In general, the process of mining association rules can be divided into two steps:

1) Find out all the frequent itemsets

First, after scanning the database, the count of each item is accumulated, and the frequent items whose appearance frequency are greater than or equal to the minimum support, are collected. We find the set of

frequent 1-itemsets, which is recorded as L_1. Then, L_1 is used to find the set L_2 of the frequent 2-itemset, L_2 is used to find L_3, and so on, until the frequent k-itemset can no longer be found. A complete scan of the database is required for each L_k found.

The flow chart of finding frequent itemsets in Apriori algorithm is shown in Fig. 3, where min_Sup means minimum support.

2) Generate strong association rules by frequent itemsets

Once we have found all the frequent itemsets, the next step is to find the association rules. If there is a frequent itemset S, A and B are both not empty subsets of it, and $A \cap B = \emptyset$.

$$\text{confidence}(A \rightarrow B) \geq \text{minConf}, \tag{13}$$

where minConf is the minimum confidence, then there is $A \rightarrow B$.

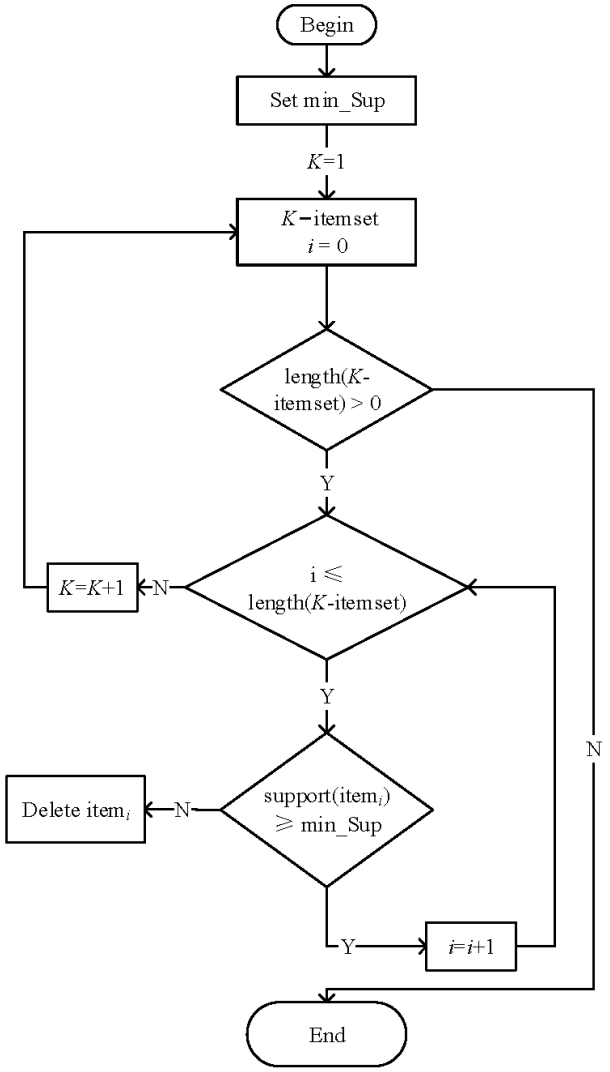

Figure 3: The flow chart of finding frequent itemsets in Apriori algorithm

3.4 Performance Evaluation

Currently, there is no clear way to show the quality of results obtained by Apriori. However, the value of "Lift" with each association rule in Apriori can reflect whether the rule is valuable, "Support" and "Confidence" can also show the importance of rules. Therefore, in the proposed scheme the evaluation

criterion is designed by combining "Support", "Confidence", and "Lift" of all association rules mined in Apriori. Let the number of association rules is M, then the evaluation criterion can be defined as follows.

$$\mathrm{EC} = \frac{1}{M}\sum_{i=0}^{M}(\mathrm{Support}_i \times \mathrm{Confidence}_i \times (\mathrm{Lift}_i - 1)), \quad (14)$$

where $\mathrm{Support}_i$, $\mathrm{Confidence}_i$, and Lift_i are the "Support", "Confidence", and "Lift" of the i-th association rule, respectively. With (14), we can adjust the cluster number of K-means according to the EC. When the best EC is found, the mined association rules are optimal.

4 Experimental Results and Discussion

4.1 Experimental Description

The dataset used here is from some hydrothermal copper deposits in China, and there are totally 252 hydrothermal copper deposits used in the experiment. We select six important attributes from each hydrothermal copper deposit, including metallogenic epoch, mineral composition, rock structure, rock fabric, alteration type, and alteration intensity.

Here, our experiments are conducted in the Python 3.6.4 environment running on the computer with an Intel(R) Core(TM) i7-6700 CPU and a 16 GB RAM.

4.2 Parameters Optimization

In our method, there are some parameters to be optimized, such as the dimension of each mineral deposit attribute, the size of SOM network, and the number of cluster centers in K-means.

Generally, the dimension of word embedding in Chinese words is variable in different scenarios. In most cases, it is 50, 100, 200, or 300. Considering the size of data, 50-dimensional vectors are enough to represent all deposit attributes. From Fig. 4, we can find that when the number of cluster centers is larger, more association rules are mined, however the value of EC difference is not large. Meanwhile, when there are more cluster centers, the association rules between categories are also lost more. Hence, through the combination of the size of data, EC, and the number of association rules, the number of clusters is set as 3, which means that the data is divided into 3 categories. Lastly, according to empirical equation, the size of SOM network is set as 21 × 12.

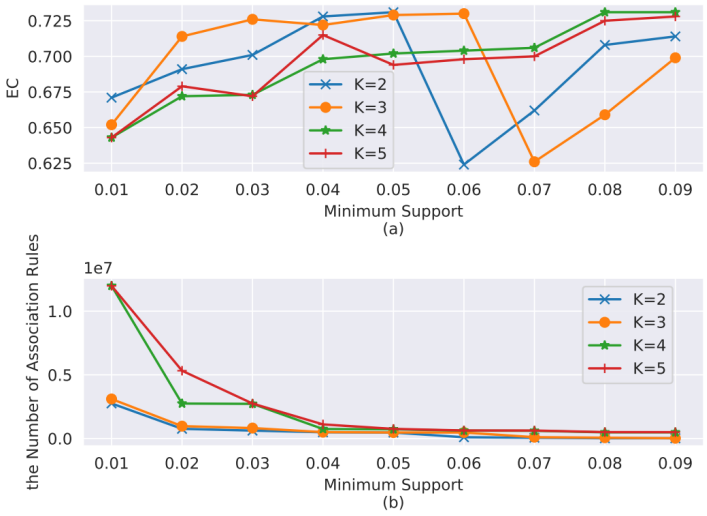

Figure 4: The relationship between EC, the number of association rules, and minimum support while using algorithm SOM-K-Apriori with different numbers of cluster centers

4.3 Performance Comparison

In order to verify the superiority of the algorithm SOM-K-Apriori, we compare it with Apriori algorithm, and the results are shown in Figs. 5 and 6.

In Fig. 5, when the minimum support is constant, the running time of SOM-K-Apriori is always shorter than that of Apriori. That is because the input/output cost of Apriori increases exponentially with the size of data, and once the range of database is reduced, the input/output cost is also greatly reduced.

In addition, from Fig. 6, we can easily observe that the method SOM-K-Apriori is able to find more valid and strong association rules. This is because some valid and strong association rules are with low support in the database. When the data are clustered with similar characteristics by SOM and K-means, they have high support in categories. This phenomenon becomes clearer as support increases.

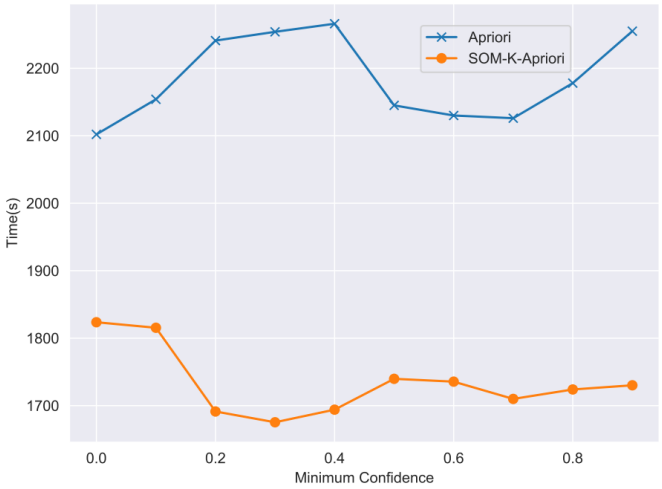

Figure 5: The relationship between running time and minimum confidence while using algorithms SOM-K-Apriori and Apriori

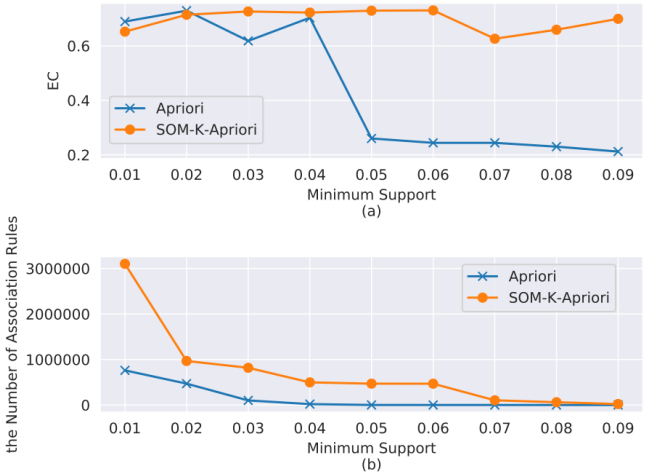

Figure 6: The relationship between EC, the number of association rules, and minimum support while using algorithms SOM-K-Apriori and Apriori

From the above experiment results, we find that, on the one hand, the algorithm SOM-K-Apriori can quickly find all association rules in the database. On the other hand, it can also find the association rules with low support but valid and strong.

4.4 Experimental Results

Firstly, for those selected six important properties in each hydrothermal copper deposit, we clean the abnormal data, and a simple statistical method is employed to analyze the regulation of mineral deposit key properties, including epochs, the rock assemblage composition, and rock structure. The results are shown as Figs. 7–10.

Fig. 7 shows the distribution of minerogenetic epochs in domestic hydrothermal copper deposits. We can see that the minerogenetic epochs of hydrothermal copper deposits are mainly concentrated in the Jurassic to Cretaceous, and its support is nearly 0.366.

Fig. 8 shows the statistical result of the rock assemblage composition in domestic hydrothermal copper deposits. The most main mineral composition of rocks is Andesite, and its support is nearly 0.360.

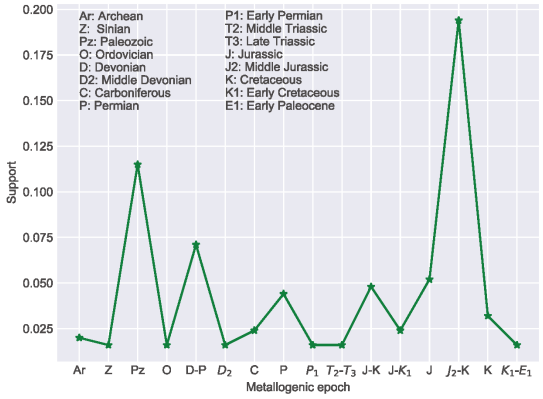

Figure 7: The distribution of minerogenetic epochs in domestic hydrothermal copper deposits

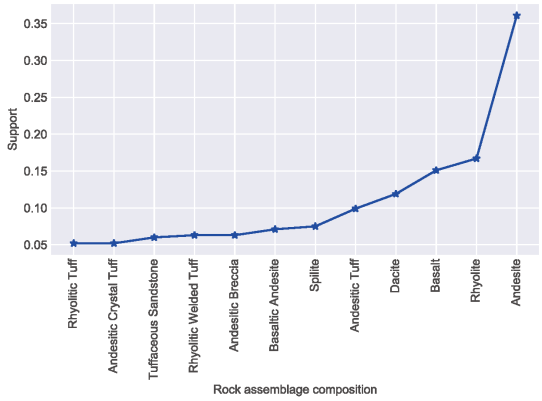

Figure 8: The statistical result of the rock assemblage composition in domestic hydrothermal copper deposits

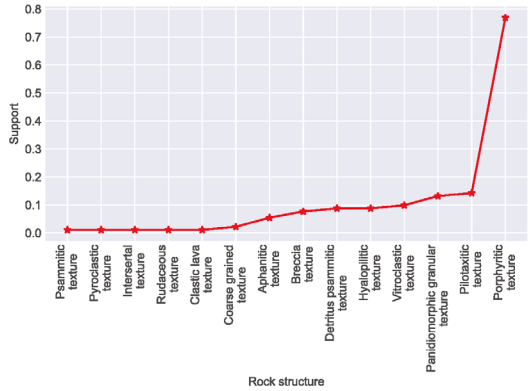

Figure 9: The distribution of the rock structure of Andesite in domestic hydrothermal copper deposits

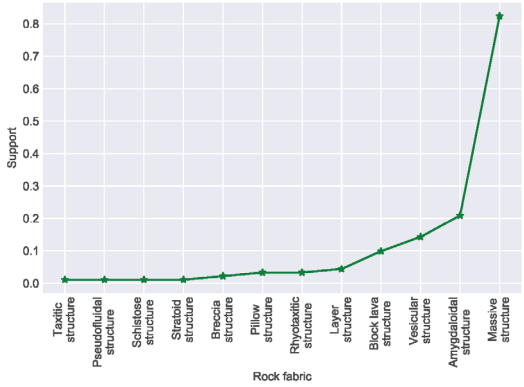

Figure 10: The distribution of the rock fabric of Andesite in domestic hydrothermal copper deposits

In order to get more information with the domestic hydrothermal copper deposits, we show the rock structure and rock fabrics of its main mineral composition Andesite. These two results are shown in Figs. 9 and 10.

Fig. 9 shows the distribution of the rock structure of Andesite in domestic hydrothermal copper deposits. Andesite has many kinds of rock structures, and the most important part of which is Porphyritic texture, whose support is as high as 0.769.

Similarly, Fig. 10 indicates that Andesite has many kinds of rock fabrics, but the main rock fabric is Massive structure, whose support is 0.824.

Secondly, we convert each attribute value to a 50-dimensional word embedding vector using word2vec model. Because some attributes contain multiple values, each mineral deposit data is represented by a 600-dimensional vector, in which the missing positions are supplemented with "0". The word embedding expressions of the mineral deposit attributes are shown in Tab. 1.

After getting 600-dimension data represented each mineral deposit, we cluster the data. Through 200 iterations, the result obtained only by SOM is shown in Fig. 11(a), and the final result obtained by SOM and K-means is shown in Fig. 11(b), where each color represents a big category. The numbers of mineral deposits contained in each big category are 72, 76, and 104, respectively, named as clu_A, clu_B, and clu_C.

Table 1: The word embedding expression of the deposit attribution

The value of deposit attributions	Embedding of deposit
J_2 Andesite ...Middle	0.0044779866 -0.0034446819 ...0 0

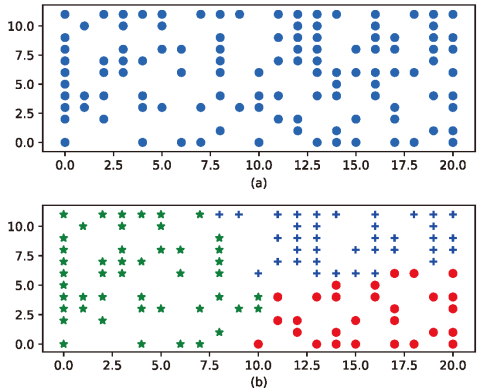

Figure 11: The clustering result: (a) Using SOM, and (b) Using SOM and K-means

Then, we use the Apriori method to mine the data, after running SOM and K-means. Through many experiments, we set the parameter of minimum support to 0.09, and set the minimum confidence to 0.6. The results regarding association rules of associated rocks, alteration type and their strength in different big categories with SOM algorithm are shown as Tabs. 2–4, respectively.

Tab. 2 indicates that Rhyolite is the highly associated rock of Rhyolitic Tuff, and Andesite is the highly associated mineral of Andesitic Volcanic Clastic Rock. The alteration type of Rhyolitic Tuff is strong Silicification. Andesite and Andesitic Volcanic Clastic Rock are mainly Massive Structure and Porphyritic Texture, while Rhyolite and Rhyolitic Tuff mainly are Layer Structure and Vitroclastic Texture.

In Tab. 3, we can easily find that Andesite is the highly associated rock of Basalt, Pyroxene Andesite, Dacitic Breccia Lava and Andesitic Tuff. Pyroxene Andesite and Dacitic Breccia Lava are associated rocks of each other. The alteration types of Pyroxene Andesite and Dacitic Breccia Lava are strong Hornfelsic. Andesite, Rhyolite, Basalt and Andesitic Tuff are all Massive structure and Porphyritic texture while Pyroxene Andesite and Dacitic Breccia Lava are Block Lava structure and Detritus psammitic texture.

Tab. 4 demonstrates that Andesite is the highly associated rock of Andesitic Crystal Tuff. The alteration types of Spilite is mainly medium Hornfelsic. Rhyolite, Andesite and Basaltic Andesite are Porphyritic texture and Massive structure, Dacite is Porphyritic texture, Massive structure and Vitroclastic texture. Andesitic Crystal Tuff has many structure types and fabric types, such as Porphyritic texture, Pilotaxitic texture, Vesicular structure, Panidiomorphic Granular texture, Amygdaloidal structure, and Massive structure.

Finally, to intuitively show those mined results, through the use of improved parallel method [39], we provide some visual descriptions of mining association rules, as Figs. 12, 13, and 14. In those figures, the antecedents are below the "––", and above the "––" are consequents. Those three figures responds to Tabs. 2–4.

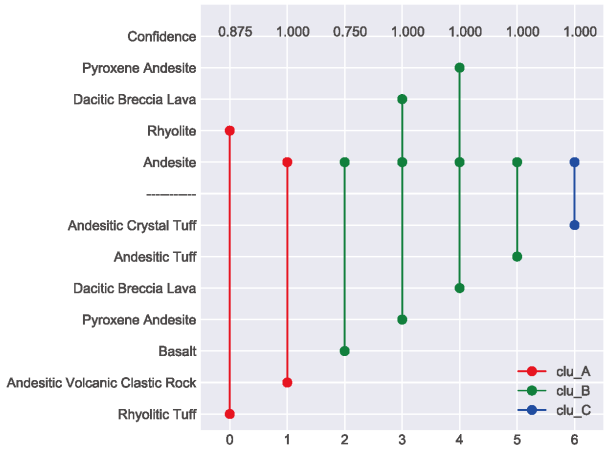

Figure 12: The association rules of associated minerals

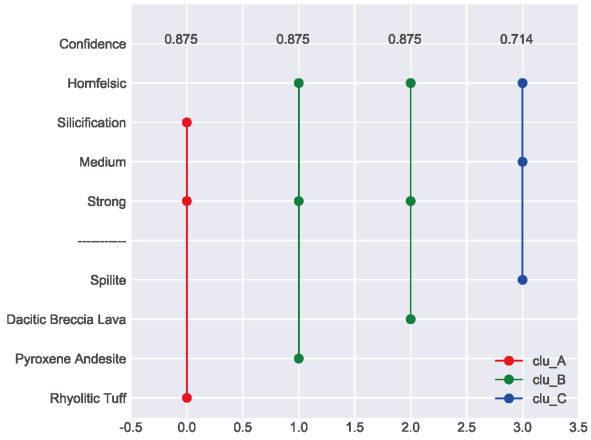

Figure 13: The association rules of minerals with alteration type and alteration intensity

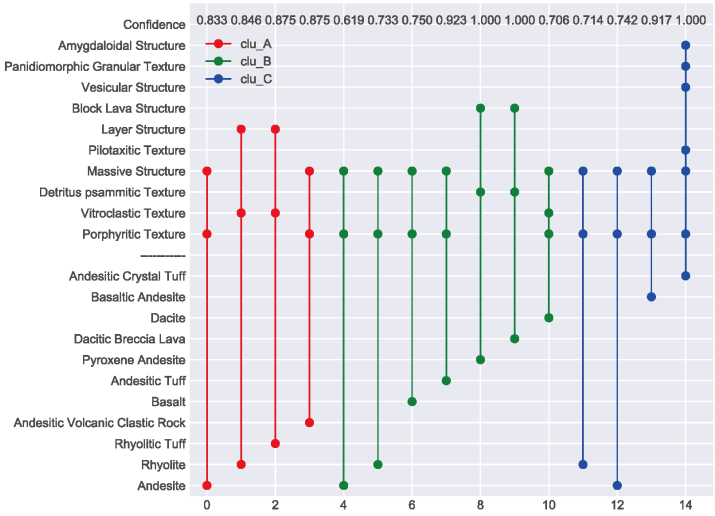

Figure 14: The association rules of minerals with structures and fabrics

Table 2: The association rule of mineral deposits in clu_A

Association rules	Confidence	Occurrences times	Lift
(Rhyolitic Tuff) ==> (Rhyolite)	0.875	7	9.00
(Andesitic Volcanic Clastic Rock) ==> (Andesite)	1.000	8	9.00
(Rhyolitic Tuff) ==> (Silicification, Strong)	0.875	7	9.00
(Andesite) ==> (Massive Structure, Porphyritic Texture)	0.833	15	4.00
(Rhyolite) ==> (Layer Structure, Vitroclastic Texture)	0.846	11	5.54
(Rhyolitic Tuff) ==> (Layer Structure, Vitroclastic Texture)	0.875	7	9.00
(Andesitic Volcanic Clastic Rock) ==> (Massive Structure, Porphyritic Texture)	0.875	7	9.00

Table 3: The association rule of mineral deposits in clu_B

Association rules	Confidence	Occurrences times	Lift
(Basalt) ==> (Andesite)	0.750	12	4.75
(Pyroxene Andesite) ==> (Dacitic Breccia Lava, Andesite)	1.000	8	9.50
(Dacitic Breccia Lava) ==> (Pyroxene Andesite, Andesite)	1.000	8	9.50
(Andesitic Tuff) ==> (Andesite)	1.000	13	5.85
(Pyroxene Andesite) ==> (Hornfelsic, Strong)	0.875	7	9.50
(Dacitic Breccia Lava) ==> (Hornfelsic, Strong)	0.875	7	9.50
(Andesite) ==> (Massive Structure, Porphyritic Texture)	0.619	26	1.81
(Rhyolite) ==> (Massive Structure, Porphyritic Texture)	0.733	11	4.75
(Basalt) ==> (Massive Structure, Porphyritic Texture)	0.750	12	4.75
(Andesitic Tuff) ==> (Massive Structure, Porphyritic Texture)	0.923	12	5.85
(Pyroxene Andesite) ==> (Block Lava Structure, Detritus psammitic Texture)	1.000	8	9.50
(Dacitic Breccia Lava) ==> (Block Lava Structure, Detritus psammitic Texture)	1.000	8	9.50

Table 4: The association rule of mineral deposits in clu_C

Association rules	Confidence	Occurrences times	Lift
(Andesitic Crystal Tuff) ==> (Andesite)	1.000	12	8.67
(Spilite) ==> (Hornfelsic, Medium)	0.714	10	7.43
(Dacite) ==> (Massive Structure, Porphyritic Texture, Vitroclastic Texture)	0.706	12	6.12
(Rhyolite) ==> (Porphyritic Texture, Massive Structure)	0.714	10	7.43
(Andesite) ==> (Porphyritic Texture, Massive Structure)	0.742	23	3.35
(Basaltic Andesite) ==> (Porphyritic Texture, Massive Structure)	0.017	11	8.67
(Andesitic Crystal Tuff) ==> (Porphyritic Texture, Pilotaxitic Texture, Vesicular Structure, Panidiomorphic Granular Texture, Amygdaloidal Structure, Massive Structure)	1.000	12	8.67

5 Conclusion

In this article, an intelligent scheme is proposed to analyze and mine the association rules of some hydrothermal copper deposits. The proposed scheme is mainly divided into three steps. First, the word embedding vectors are generated for geological data. Then, we use the SOM and K-means to cluster the similar attribution characteristics into one category. Finally, the evaluation criterion-guided Apriori algorithm is specifically used to mine the association relationship in every category. We can see from the experimental results that compared with the traditional Apriori, there are two main advantages in the proposed scheme, one is to quickly mine all the association rules by reducing the scope of scanning, and the other is to mine valid and strong association rules which are with low support. Hence, this intelligent scheme is more suitable for mining valid and strong association rules for geological data. However, in this model SOM-K-Apriori, as the data is clustered by SOM and K-means, the association rules among categories will be lost. Therefore, in the future, for those data with more key attributes, we will reduce the dimensionality of them, preventing the loss of association rules between categories.

Funding Statement: This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFC0600510, in part by the National Natural Science Foundation of China under Grant U1836106 and Grant 41872253, in part by the Beijing Natural Science Foundation under Grant 19L2029, in part by the Beijing Intelligent Logistics System Collaborative Innovation Center under

Grant BILSCIC-2019KF-08, in part by the Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB, under Grant BK19BF006, and in part by the Fundamental Research Funds for the University of Science and Technology Beijing under Grant FRF-BD-19-012A.

Conflicts of Interest: No potential conflict of interest was reported by the authors.

References

[1] Y. Liu, A. Liu, X. Liu and X. Huang, "A statistical approach to participant selection in location-based social networks for offline event marketing," *Information Sciences*, vol. 480, pp. 90–108, 2019.

[2] L. Wang, Z. Zhang and X. Luo, "A two-stage data-driven approach for image based wind turbine blade crack inspections," *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 3, pp. 1271–1281, 2019.

[3] X. Luo, Y. Xu, W. Wang, M. Yuan, X. Ban *et al.,* "Towards enhancing stacked extreme learning machine with sparse autoencoder by correntropy," *Journal of The Franklin Institute*, vol. 355, no. 4, pp. 1945–1966, 2018.

[4] X. Luo, C. Jiang, W. Wang, Y. Xu, J. H. Wang *et al.,* "User behavior prediction in social networks using weighted extreme learning machine with distribution optimization," *Future Generation Computer Systems*, vol. 93, pp. 1023–1035, 2019.

[5] X. Luo, Y. Li, W. Wang, X. Ban, J. H. Wang and W. Zhao, "A robust multilayer extreme learning machine using kernel risk-sensitive loss criterion," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 1, pp. 462–476, 2020.

[6] L. Feng, X. Xu, H. Yuan and Q. Zhang, "Detecting individual content-structure patterns in time series data," in *Proc. of the 13th Int. Conf. on Service Systems and Service Management*, Piscataway, NJ, USA, pp. 1–5, 2016.

[7] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," arXiv:1707.02919, 2017.

[8] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang *et al.,* "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[9] X. Luo, J. Sun, L. Wang, W. Wang, W. Zhao *et al.,* "Short-term wind speed forecasting via stacked extreme learning machine with generalized correntropy," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4963–4971, 2018.

[10] M. Abdel-Basset, M. Mohamed, F. Smarandache and V. Chang, "Neutrosophic association rule mining algorithm for big data analysis," *Symmetry*, vol. 10, no. 4, 2018.

[11] M. Chen, Y. Li, X. Luo, W. Wang, L. Wang *et al.,* "A novel human activity recognition scheme for smart health using multilayer extreme learning machine," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1410–1418, 2019.

[12] N. Xue, X. Luo, Y. Gao, W. Wang, L. Wang *et al.,* "Kernel mixture correntropy conjugate gradient algorithm for time series prediction," *Entropy*, vol. 21, no. 8, 2019.

[13] K. H. Coble, A. K. Mishra, S. Ferrell and T. Griffin, "Big data in agriculture: A challenge for the future," *Applied Economic Perspectives and Policy*, vol. 40, no. 1, pp. 79–96, 2018.

[14] J. C. Kwon, J. S. Lee and M. C. Jung, "Arsenic contamination in agricultural soils surrounding mining sites in relation to geology and mineralization types," *Applied Geochemistry*, vol. 27, no. 5, pp. 1020–1026, 2012.

[15] B. Y. Choi, S. T. Yun, K. H. Kim, K. Kim and S. J. Choh, "Geologically controlled agricultural contamination and water-rock interaction in an alluvial aquifer: Results from a hydrochemical study," *Environmental Earth Sciences*, vol. 68, no. 1, pp. 203–217, 2013.

[16] Y. Zhu, Y. Tan, X. Luo and Z. He, "Big data management for cloud-enabled geological information services," *Scientific Programming*, vol. 2018, pp. 1–13, 2018.

[17] Q. Zhang and X. Liu, "Big data: New methods and ideas in geological scientific research," *Big Earth Data*, vol. 3, no. 1, pp. 1–7, 2019.

[18] Y. Zhou, S. Chen, Q. Zhang, F. Xiao, S. Wang *et al.,* "Advances and prospects of big data and mathematical geoscience," *Acta Petrologica Sinica*, vol. 34, no. 2, pp. 255–263, 2018.

[19] J. Chen, J. Xiang, Q. Hu, W. Yang, Z. Lai *et al.,* "Quantitative geoscience and geological big data development: A review," *Acta Geologica Sinica*, vol. 90, no. 4, pp. 1490–1515, 2016.

[20] M. Zhai, "Granites: Leading study issue for continental evolution," *Acta Petrologica Sinica*, vol. 33, no. 5, pp. 1369–1380, 2017.

[21] J. Wang and Y. Zhu, "Application of incremental updating association mining algorithm in geological disasters system," in *Proc. of the Second Int. Conf. of Sensor Network and Computer Engineering*, Paris, France, pp. 45–53, 2018.

[22] D. Wedge, A. Lewan, M. Paine, E. J. Holden and T. Green, "A data mining approach to validating drill hole logging data in pilbara iron ore exploration," *Economic Geology*, vol. 113, no. 4, pp. 961–972, 2018.

[23] R. Ma, X. Ma and Y. Pu, "Spatial association rule mining from GIS database," *Journal of Remote Sensing*, vol. 9, no. 6, pp. 733–741, 2005.

[24] B. He, Y. Cui, C. Chen and J. Chen, "Geology spatial data mining method for regional metallogenic prediction," *Advances in Earth Science*, vol. 26, no. 6, pp. 615–623, 2011.

[25] L. Chang, Y. Zhu, G. Zhang, X. Zhang and B. Hu, "Spatial correlation analysis of mineral resources information," *Acta Petrologica Sinica*, vol. 34, no. 2, pp. 314–318, 2018.

[26] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[27] G. C. Tjhai, S. M. Furnell, M. Papadaki and N. L. Clarke, "A preliminary two-stage alarm correlation and filtering system using SOM neural network and K-means algorithm," *Computers & Security*, vol. 29, no. 6, pp. 712–723, 2010.

[28] J. Chen, R. Peng, S. Li and X. Chen, "Self-organizing feature map neural network and k-means algorithm as a data excavation tool for obtaining geological information from regional geochemical exploration data," *Geophysical & Geochemical Exploration*, vol. 41, no. 5, pp. 919–927, 2017.

[29] R. J. Kuo, L. M. Ho and C. M. Hu, "Integration of self-organizing feature map and k-means algorithm for market segmentation," *Computers & Operations Research*, vol. 29, no. 11, pp. 1475–1493, 2002.

[30] W. Niyagas, A. Srivihok and S. Kitish, "Clustering e-Banking customer using data mining and marketing segmentation," *ECTI Transaction on Computer and Information Technology*, vol. 2, no. 1, pp. 63–69, 2006.

[31] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[32] G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, M. Re and S. Spanó, "AW-SOM, an algorithm for high-speed learning in hardware self-organizing maps," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 2, pp. 380–384, 2020.

[33] J. Liu and L. Xu, "Improvement of SOM classification algorithm and application effect analysis in intrusion detection," in *Proc. of the Int. Conf. on Intelligent Computing, Communication and Devices*, Singapore, pp. 559–565, 2019.

[34] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. of the 20th Int. Conf. on Very Large Data Bases*, San Mateo, CA, USA, pp. 487–499, 1994.

[35] S. Tao, Y. Li, X. Xiao and L. Yao, "Load forecasting based on short-term correlation clustering," in *Proc. of the 7th IEEE Innovative Smart Grid Technologies–Asia*, Piscataway, NJ, USA, pp. 1–7, 2017.

[36] S. Wu, Y. Zhang and Y. Su, "The medium-voltage distribution detwork fault fiagnosis based on data association analysis," in *Proc. of the Int. Conf. on Green Energy and Applications*, Piscataway, NJ, USA, pp. 57–63, 2017.

[37] N. Shakhovska, R. Kaminskyy, E. Zasoba and M. Tsiutsiura, "Association rules mining in big data," *International Journal of Computing*, vol. 17, no. 1, pp. 25–32, 2018.

[38] G. Sheng, H. Hou, X. Jiang and Y. Chen, "A novel association rule mining method of big data for power transformers state parameters based on probabilistic graph model," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 695–702, 2016.

[39] J. Zhang, Y. Yang and W. Tan, "Visualization methods of association rules based on I-Miner," *Journal of Xihua University (Natural Science Edition)*, vol. 29, no. 6, pp. 55–58, 2010.