**Supplementary File**

# **1. Theoretical Bases**

## **1.1 The Bayesian Theory**

The Bayesian formula is stated as follow:

 (1)

Where P(*Ai|B*) is the probability of event *Ai* when event *B* has occurred, , , . In the Bayesian theory, P(*A*) is called the prior probability of event *A*, and the is P(*Ai|B*) the posterior probability of event *A*. More details can be found in the publication of Kononenko (1993) and Li et al. (2014).

## **1.2. Naïve Bayes Algorithm**

The Naïve Bayes algorithm (NB) is based on the Bayesian theory (Kononenko 1993). Supposing that the samples are in an *n*-dimensional space, *X* = {*x1*, *x2*, …, *xn*}, and there are *m* categories represented by *C*1, *C*2, …, *Cm*, then an unknown sample *X’* can be assigned to *Ci* with formula (2):

 (2)

According to the Bayesian theory, the denominator, *P*(*X’*), is a constant. Therefore, the maximum *P*(*Ci*|*X’*) can be represented by the maximum *P*(*X’*|*Ci*)*P*(*Ci*). Because the *n*-dimensional space is large, the amount of computation needed for such calculations is huge. So the Naïve Bayes assumes that all of the *n* features are independent. In this way, for each unknown sample *X’*, the probability that *X’* belongs to *Ci* can be calculated as *P*(*X’*|*Ci*)*P*(*Ci*). Finally, the category of *X’* can be found out by the maximum of these probabilities.

## **1.3. Copula Function**

Sklar (1959) pointed out that an *n*-dimensional joint probability distribution function (PDF) can be decomposed into *n* marginal distribution functions and a copula function. Nelsen (2013) strictly defined that copula functions are “functions that join or ‘copula’ multivariate distribution functions to their one-dimensional marginal distribution functions”. A function  is a copula function if it satisfies: (1) for every , ***C***(***u***)=0 if at least one coordinate of *u*=0; ***C***(***u***)=***u****k* if all coordinates of ***u*** are 1 except ***u****k*; (2) for every  with ***a***≤***b***:

 (3)

where ***V****c*([***a****,****b***]) is the *n*th order difference of *C* on [***a***, ***b***]. The relationship between a copula function and a joint distribution function can be described as follow: if *H*(*X1*, *X2*, …, *Xn*) is an *n*-dimensional distribution function and *F1*, *F2*, …, *Fn*arethe marginal distributions of *X1*, *X2*, …, *Xn*, then:

 (4)

If *F1*, *F2*, …, *Fn* are all continuous, *C* is unique; otherwise, *C* is unique on Ran*F1*× Ran*F2*× … ×Ran*Fn.* Conversely, if *F1*, *F2*, …, *Fn* are distribution functions and *C* is a copula function, then the *H*(*X1*, *X2*, …, *Xn*) described by formula (4) must be a joint PDF.

The joint PDF, , can be described by formula (5):

 (5)

where the former  is the copula density function that can accurately describe the relationship between the *n* variables, the latter *f1*(*x1*)*f2*(*x2*)…*fn*(*xn*) is the product of the *n* marginal PDFs. Therefore, the joint PDF, *f* (*x1*, *x2*, …, *xn*) can be calculated by the two parts.

Many kinds of copula functions have been presented, such as Gaussian copulas and Archimedean copulas. In this study, the Gaussian copula is adopted because it is more convenient to situations where the number of variables is more than two. An *n*-dimensional Gaussian copula can be described as follows:

(6)

where *X*=(*X1*, *X2*, …, *Xn*) ~*Nn*(0, ),  is the symmetric covariance matrix and it can be described by formula (7):

 (7)

Then the copula density function is:

(8)

The parameters in copula density functions are usually estimated by the maximum likelihood estimation (MLE) and the moment estimation (ME). In this research, the MLE is adopted. The principle of MLE is as follows: by giving a family of PDFs, the parameters the PDFs can be determined by maximize the likelihood function of the PDFs. More details can be found in the publication of Choroś et al. (2010).

## **1.4. Genetic Algorithm (GA)**

The GA is a computing method that simulates the Natural Selection Theory of Evolution and mechanisms of genetics, and it is a method for finding the optimum solution (Thede 2004). A GA starts by randomly generating a population from a solution set, and every individual of the population is a feasible solution represented by a series of codes (usually binary codes). After that, according to the principle of “survival of the fittest”, the population will evolve to a new population generation by generation. In the algorithm, the evolution process is an iterative process. In an iteration, the individuals that map to poor solutions will be eliminated, and new individuals will be generated in a process of mutation and crossover. Therefore, individuals in a new population are closer to the optimal solution, and the last generation of the population can be regarded as a set of optimal solutions of the problem. Fig. 1 shows the flowchart of GA:

(a) Initialization: set the counter of iteration, *t*, to 0; supposing the maximum number of iteration is *T*; randomly generate a population, P(0).

(b) Evaluation of individuals: calculate the adaptations of every individual of P(*t*).

(c) Selection: eliminate the low adaptation individuals (lower than 70% of the individuals) and retain the high adaption individuals.

(d) Crossover: for the reserved individuals, use a kind of crossover operator to generate new individuals.

(e) Mutation: randomly change the codes of the individuals in a small probability to make a small probability of mutation. After selection, crossover and mutation, the old population, P(*t*), will evolve to P(*t*+1). Then return to step (b).

(f) Terminating condition: if *t* = *T* or the problem is satisfied with the new population, then take the individual that has the highest adaptation in the whole evolution process as the optimal individual and decode it as the optimal solution, and terminate the algorithm.

More details about the GA can be found in the publication of Thede (2004).



Fig. 1. Flowchart of GA

# **2. Mathematical Model of Discriminating Among Tectonic Settings**

## **2.1. Possibility Distribution Function of Elements**

The first step of deriving the mathematical model is to determine the possibility distribution functions of the elements of rocks.

The distributions of rock’s chemical elements are complex. Totally, they can be divided into two categories: symmetrical and asymmetrical (Ahrens, 1954; Huang et al., 2013). In this research, we suggest to use the normal distribution to fit the symmetrical distributions, and use lognormal and gamma distribution to fit the asymmetrical distributions, as shown in formula (9-14).

Normal distribution:

 (9)

 (10)

Lognormal distribution:

 (11)

 (12)

Gamma distribution:

 (13)

 (14)

Therefore, the PDF of an element can be fitted by formula (10), (12), or (14). The degrees of fitting are estimated by log likelihood. Fig. 2(a) shows an example of the fitting. The log likelihood of normal, lognormal and gamma distribution are -2416, -3393 and -2718 respectively. Therefore, the normal distribution is chosen for fitting.



Fig. 2. Examples of distribution of elements

However, as shown in Fig. 2(b), some elements may obey some strange distributions that can not be well fitted by any common distribution functions. The red circle in Fig. 2(b) indicates the probability density of zeros of the element. To solve this problem, we suggest treating the zeros and non-zero values separately. Therefore, the PDF of each element should be a set of piecewise function. Fig. 3 shows the cumulative probability function (CDF) of an element. The second section is the range of non-zero values of a set of data, and it can be fitted with the three distribution function mentioned above. The first section is the range from 0 to the minimum of non-zero value of an element, and the cumulative probability function of this part is defined as follow: (1) take the zeros as some small values, *ε*, that uniformly distributed in the small range, *r*; (2) for a specific element, assume *r* is the range that from 0 to 0.01 times of the minimum non-zero values; (3) assume *ε* is the average of 0 and 0.01 times of the minimum non-zero values, as shown in Fig. 3.



Fig. 3. The cumulative probability distribution of an element

Therefore, the CDF of an element can be described as formula (15):

 (15)

where, *Min* is the minimum of the non-zero values of the corresponding element; *p0* is the proportion of zero-data, as shown in Fig. 3; *F’* is the CDF of normal, lognormal or gamma distribution. Therefore, the PDF is as formula (16):

 (16)

Where *f’* is the PDF of normal, lognormal or gamma distribution.

## **2.2 Joint Probability Density Functions of Rocks**

This second step is to determine the PDF of a rock in a specific tectonic setting. For example, in the island arc environment, what is the probability of a rock whose geochemical components is {SiO2: 50.27 wt%, TiO2: 0.78 wt%, …, Pb: 12 ppm}. As is known to all, the contents of the elements are not mutually isolated but having some relation existed among them. Therefore, joint PDFs are needed to calculate the probability of rocks.

A joint PDF can be established based on Copula theory. The copula density function of the joint PDF can be derived from formula (17), and is as follow:

(17)

Where *ei* represents the *i*th element (e.g. *e1* may be SiO2, *e2* may be TiO2), *xei* is the content of the *ei*,  is the CDF of *ei*. And the  can be calculated by the formula (15). Then the joint PDF is derived as formula (18):

(18)

Where  is the PDF (or margin density function) of *ei*, and can be calculated by the formula (16).

In this way, the probability of a rock can be determined with its geochemical components. It should be noted that different tectonic settings correspond to different joint PDFs. In this research, three joint PDFs were derived, including (1) the joint PDF of island arc, , (2) the joint PDF of ocean island, , and (3) the joint PDF of mid-oceanic ridge, . In the there functions, *TS* represents tectonic setting, *TSia* is island arc, *TSoi* is ocean island, and *TSmor* is mid-oceanic ridge.

## **2.3. Mathematical Model for The Discrimination**

This section is about how to discriminate among different tectonic settings by combining the Bayesian theory and the joint PDFs derived in Section 2.2, that is how to establish the mathematical discrimination model.

According to the Bayesian theory (Pawlak 2003), formula (19) is set up:

 (19)

where, *TStype* is the type of a specific tectonic setting; *P*(*TS=TStype*) is the probability of event of *TStype*; *X*={*x1, x2, …, xn*} is a sample that contains *n* features and the features include major elements, trace elements, and isotopes; *P*(*X*={*xe1, xe2, …, xen*}) is the probability of the occurrence of *X* within the scope of all types of tectonic settings; *P*(*X*={*xe1, xe2, …, xen*}|*TS=TStype*) represents the probability of the occurrence of *X* when the tectonic setting has been determined as *TStype*; *P*(*TS=TStype* |*X*={*xe1, xe2, …, xen*}) represents the probability of the occurrence of *TStype* when the features of a sample are *X*; For simplicity, the event, *TS=TStype*, is directly represented by *TStype* in the rest of the paper.

Applying the total probability formula of the Bayesian theory, formula (19) can be converted to formula (20):

 (20)

Considering that the distribution of each *xei* (*i*=1, 2, …, *n*) is continuous, the *P*(*X*={*xe1, xe2, …, xen*}|*TStype*) can be calculated by the joint PDFs of formula (18). Then the formula (21) is derived:

 (21)

The formula (21) can be regarded as the general formula for calculating the probability of the tectonic settings according to the chemical components of rocks. In this research, it is assumed that *TS* is comprised of island arc (*TSia*), ocean island (*TSoi*) and mid-oceanic ridge (*TSmor*).When a sample *X*={*xe1, xe2, …, xen*} is given, by calculating the *P*(*TSia*|*X*={*xe1, xe2, …, xen*}), *P*(*TSoi*|*X*={*xe1, xe2, …, xen*}) and *P*(*TSmor*|*X*={*xe1, xe2, …, xen*}) and comparing the three probabilities, the type of tectonic setting can be determined by the maximum of these three probabilities.

As to the *P*(*TStype*) (the probability of occurrence of *TStype*), it is difficult to estimate the probability of the occurrence of island arc, ocean island and mid-oceanic ridge. In this research, we suggest to use the proportions of samples to represent them approximately. For example, if there are 10000 island arc samples, 20000 ocean island samples, and 30000 mid-oceanic ridge samples, then *P*(*TSia*): *P*(*TSoi*): *P*(*TSmor*) = 1: 2: 3, and *P*(*TSia*) =1/6, *P*(*TSoi*)=1/3 and *P*(*TSmor*)=1/2. Obviously, this assumption can be valid only if there are enough samples.

## **2.4. Optimization of The Mathematical Model**

### *2.4.1. Why Optimization Is Needed for The Mathematical Model*

Under ideal situations, all distributions of the elements from different tectonic settings can be fitted, and precise joint PDFs can be established to calculate the probabilities of the components of rocks. However, there is always error in measuring and fitting. The PDF of an element may be different from the reality because of the limitation of the data, for example, the number of data may be not large enough for a good fitting. On the other hand, even if the distribution of an element is well-fitted, the element is not helpful for discrimination if the distributions of the element in different tectonic settings are the same. Therefore, it needs to choose which elements should be used to build the mathematical model.

### *2.4.2. Selections of Parameters Based on The GA*

For one element, there are two decisions: use it and not use it. Therefore, for *n* elements, there are 2*n* solutions. In the geochemical analysis, there are usually more than 100 elements, thus the number of solutions is larger than 2100, which demands serious computational power. To this end, the GA is adopted to select the most effective elements.

The key problems of using GA are: (1) how to encode the solutions into individuals; (2) how to calculate the adaptions; (3) how to design the selection operator; and (4) how to design the crossover and mutation operators. The strategies taken in this research are as follows:

(1) Encoding method: the binary encoding method is adopted. For a specific variable, 1 represents selection and 0 represents rejection. For example, if an individual is 10011, it means that the 1st, the 4th, and the 5th variables are used. Therefore, the length of an individual depends on the number of variables, and the number of selected variables depends on the number of 1.

(2) Adaption calculation: in this research, the adaptions are represented by the discrimination percent success of the mathematical model. For example, when calculating the adaption of an individual, 10011, the 1st, the 4th, and the 5th elements should be used to build a mathematical model. And then the adaption of the individual is the percent success of the model.

(3) Selection operator: elitism reservation is adopted for selection. There are many kinds of selection operators, such as elitism reservation, roulette wheel selection, and tournament selection. Elitism reservation has been proven to be able to achieve global optima, and is better than the other two methods. Its main idea is to copy the best individual that occurred in the iteration to the next generation without any crossover and mutation process.

(4) Crossover and mutation operators: crossover and mutation is the main way to generate new individuals and helps the algorithm to escape from local optimization. The strategies taken in this research are the same as the classical GA (and not explained here).