Sparse feature selection for clustering and sample-wise distance, with application to geochemical data

*Kenta Ueki[1], Hideitsu Hino[2]

1. Earthquake Research Institute, The University of Tokyo, 2. Graduate School of Systems and Information Engineering, University of Tsukuba

Many geochemical data, such as multiple samples from the single volcano or rock suite, and the multipoint-local analysis in one sample, are the assemblage of many samples each of which has high dimensional composition data. In order to extract geochemical process hidden in the rock samples, it is necessary to conduct a data-driven multivariate analysis of high dimensional data consisting of multivariate-multi samples. However, most previous geochemical studies have been carried out at low dimensions, such as only the relationship between several elements, or focusing on only a few specific samples. Furthermore, data analyses were carried out with giving some specific assumptions such as chemical compositions of some end-member components. Furthermore, data distributions and its shape are not fully utilized in previous studies, although the rock formation process should be reflected in the shape of the data distribution made by many samples. It is expected that more geochemical information can be extracted from the geochemical data by analyzing the distribution of data in multidimensional space quantitatively.
In order to analyze such kind of data, we considered the "distribution" defined by the observation values of the series of samples or multiple analyses. We measured the distance between the distributions. The distance between distributions is derived by a nonparametric method which does not assume any specific probability distribution. The distance corresponding to each feature quantity is defined. The total distance is defined by the weighted sum of "element distances". By using clustering with this weight and further selecting features by imposing sparse constraints on the weights, we can calculate the distances between sets and the quantities characterizing distances (in the case of this study, the elemental species and the specific isotopic ratios).
The advantages of this method are,
1. It enables us to determine variables characterizing the distance,
and
2. It is unnecessary for all samples to have analytical values of all elements with this method.
Using rock chemical composition database "petdb" (http://www.petdb.org), compositional data of 3988 MORB samples, up to 49 elements (including 5 isotopes and 10 major elements) was compiled and used for analysis. Based on its spatial distribution, MORB was grouped into several clusters and chemical compositions and distances between the clusters are compared. Elements or isotope ratios that characterize the spatial variation and the distances between the clusters were obtained using this method. As a result, MORB is clustered into the east-west hemisphere. Sr isotope ratio was found to be most important as an amount characterizing the spatial variation of MORB. Clustering of this east-west hemisphere is consistent with the structure shown by Iwamori and Nakamura (2015). Since the Sr isotopic ratio is sensitive to the amount of recycled material in the source mantle (e.g., Hoffman, 1997; Albarede, 2009), it is suggested that the distribution of recycled material of the subducted slab is systematically different between the east and western hemispheres.

Keywords: MORB, Machine learning, Geochemical data

