Instruction Guide

Deep autoencoder network connected to geographical random forest (DAN–GRF)

Zeinab Soltani[a], Saeid Esmaeiloghli[b]

December, 2023

Introduction

This program code was scripted over functions compiled in the R language environment to implement the DAN–GRF algorithm, initially developed in a paper titled: "A deep autoencoder network connected to geographical random forest for spatially aware geochemical anomaly detection." The DAN–GRF is an innovative algorithm connecting a deep learning architecture to a spatial machine learning processor to account for local neighborhood information and spatial non-stationarities in support of spatially aware anomaly detection.

According to the DAN–GRF algorithm, a deep autoencoder network (DAN) is trained to abstract deep feature codings (DFCs) of multi-element input data. A local version of the random forest algorithm, geographical random forest (GRF), is then connected to the input and code layers of the DAN processor to establish nonlinear and spatially aware regressions between original geochemical signals (dependent variables) and DFCs (independent variables). After contributions of the latter on the former are determined, residuals of GRF regressions are quantified and interpreted as spatially aware anomaly scores related to mineralization.

[a] Department of Mining Engineering, Amirkabir University of Technology, Tehran, Iran; e-mail: zs.soltani@aut.ac.ir.
[b] Center for Exploration Data Mining, National Iranian Copper Industries Company, Tehran, Iran; e-mail: esmaeiloghli@gmail.com.

Input data

The input data includes:

- data: a text file comprising an $n \times p$ matrix of the geochemical dataset, in which rows represent n geochemical samples. The first two columns are for spatial coordinates (x, y), while the remaining $p - 2$ columns express concentrations of geochemical elements. To perform a quick experiment, the user can call up a small-sized test dataset from the "Inputs (Test Data)" folder. This dataset contains 100 geochemical samples, in which the spatial coordinates (i.e., x and y) and concentrations of seven trace elements (i.e., Ag, As, Au, Cu, Pb, Sb, and Zn) are reported for each sample.

Outputs

The program returns the following outputs/results:

- A plot depicting pairwise scatter plots, probability density functions, and Pearson's product-moment correlation coefficients for original input data of elemental concentrations;
- ILR Transformed Data: a text file comprising an $n \times (p - 3)$ matrix of multi-element concentration data transformed by isometric log-ratio (ilr) transformation. In this matrix, rows and columns represent geochemical samples and ilr-transformed geochemical variables, respectively;
- A plot depicting pairwise scatter plots, probability density functions, and Pearson's product-moment correlation coefficients for ilr-transformed data of elemental concentrations;
- Deep Features: a text file comprising DFCs encoded by the DAN algorithm;
- Reconstruction: a text file comprising ilr-transformed variables reconstructed by the DAN algorithm;
- Anomaly Score: a text file comprising a vector of anomaly scores (reconstruction errors) generated using the DAN algorithm;

- Normalized Anomaly Score: a text file comprising a vector of anomaly scores normalized into the range of $[0,1]$;
- Bandwidths and R^2 of Local Model: a text file comprising a numeric table with two columns that represent the examined bandwidths and the corresponding local R^2 values, respectively, to adopt the optimal bandwidth value for GRF regression;
- Local Model Summary: a text file comprising a table of local model summary and goodness of fit statistics derived from the GRF model;
- Residuals and Local Goodness of Fit: a text file comprising a table of residuals and local goodness of fit statistics derived from the GRF model;
- Anomaly by DAN–GRF: a text file comprising a vector of anomaly scores (residuals) generated using the DAN–GRF algorithm.

It should be noted that the results obtained for the test dataset can be found in the "Results" folder.

Citation

To cite this program, this would be an appropriate format:

Soltani, Z., Hassani, H., & Esmaeiloghli, S. "A deep autoencoder network connected to geographical random forest for spatially aware geochemical anomaly detection." (Submitted Paper, 2023).