Tree-Classifier for Gaussian process regression (TCGPR)

In collecting material data, the conditions to generate these data including data accuracy and noise should be known to ensure the collected data are reproducible. Due to some unknown reasons, some experimental data may scatter greatly. In this circumstance, one has to increase the number of repeated tests, because only sufficiently repeated tests are able to provide the reliable means and variance of test results. However, some materials data reported in the literature might be inconsistent and incorrect. In addition to increasing the robustness of machine learning (ML) models to outliers continuously, outlier identification algorithms provide a shortcut for enhancing the regression/classification ability of ML models.

Tree-Classifier for Gaussian process regression (TCGPR) is a novel data preprocessing algorithm developed by Tong-yi Zhang and Bin Cao et. al. (2022) for identifying outliers and/or enhancing data connections. TCGPR is run from a specific initial dataset, and yields in the sequence forward ideology. By maximizing the expected decrease (ED) of the global Gaussian massy factor (GGMF), the initial dataset is continuously expanded until the fitting-goodness of the initial dataset saturates. Data processed by TCGPR are divided into two exclusive datasets as a stump-tree. One of which is the screened Gaussian correlation enhanced dataset and the other contains the remaining data. A new epoch of TCGPR can be done on the remaining dataset for partitioning data further, based on a new initial dataset initialized on it. A series of high-quality datasets are derived by iterative this process, until the data capacity of remaining dataset is small enough to be identified as outliers or the fitting goodness of it is good enough to attach the requirement for learning of machine learning (ML) models.

How the TCGPR model works

TCGPR is a data preprocessing method based on the Gaussian correlation between data. The ultimate goal is to maximize such Gaussian correlation by screening outliers and/or partitioning dataset. In TCGPR, the Gaussian radial kernel function embedded Gaussian process regression model (GPR) integrates the leave one out cross-validation (LOOCV), is applied for fitting datasets and evaluating them. A global Gaussian messy factor is proposed for evaluating the data messiness of a specific dataset, is defined as,

$$GGMF = \frac{\ln(NL+1)\times(1-R)}{\ln(ls+1)}, \quad (1)$$

where, $NL \in (0, +\infty)$ is the mean of the negative log-likelihood values on training data of LOOCV for measuring the fitting goodness of multivariate mixture Gaussian distribution on training data, the lower the better. R is the Pearson correlation coefficient of LOOCV for measuring the fitting goodness of Gprs on testing data, which is yield in -1 to 1, the higher the better. $ls \in (0.001, 1000)$ is the mean of optimized length scales of Gaussian radial kernel function on training data of LOOCV for reflecting the fitting surface. A low length scale corresponds to an unreasonably sharp fitting surface. Therefore, a lower GGMF is, the stronger the Gaussian connection between data. Gaussian radial kernel function is defined as Eq.(2),

$$K(X_i, X_j) = exp\left(-\frac{\|X_i - X_j\|^2}{2(ls)^2}\right), \quad (2)$$

where $\|\cdot\|$ represents the Euclidean distance of argument vector, X_i and X_j is datum argument vector.

TCGPR treats the GGMF as a normal distribution by taking prediction uncertainty into consideration at molding. The mean of prediction standard

deviation of dataset i on LOOCV, σ_i, is applied to measure the uncertainty of $GGMF_i$ of dataset i. Therefore, the GGMF of a specific dataset i is models as a normal distribution,

$$GGMF \sim N(GGMF_i, \sigma_i^2) \tag{3}$$

In TCGPR, a specific initial dataset is initialized firstly and expanded continuously by adding new data to it. As mentioned above, the GGMF is expected to decrease during this dynamic process. Intuitively, an effective strategy is to maximize the expected distance between current dataset's GGMF value, GGMF^*, and the expanded dataset's GGMF value, $GGMF_i$, which is yield in,

$$ED = E(max[\Delta,\ 0]) = \Delta\phi\left(\frac{\Delta}{\sigma_i}\right) + \sigma_i\varphi\left(\frac{\Delta}{\sigma_i}\right), \tag{4}$$

where $\Delta = \mathrm{GGMF}^* - GGMF_i$. $\varphi(\cdot)$ and $\phi(\cdot)$ are the standard normal density function and standard normal cumulative distribution function, respectively. σ_i is the mean of prediction standard deviation of dataset i on LOOCV.

Assume that there has an original dataset contains n discrete data, the initial dataset screened by TCGPR has capacity of m, and the addition path is q, which means add q data into the initial dataset at each time during the dynamic expansion process. There is a total of C_{n-m}^{q} kinds expanded candidates at the first-time. If $C_{n-m}^{q} \leq 10^6$, the brute force search strategy is applied for selecting the best candidate over the C_{n-m}^{q} situations with highest expected decrease, which is defined in Eq.(4). While if $C_{n-m}^{q} > 10^6$, 10^6 kinds of situations will be randomly sampled form all C_{n-m}^{q} cases. And the best candidate will be picked out over the 10^6 situations with highest expected decrease. After the first-time expansion, the initial dataset is bigger and contains $m+q$ discrete data. Again, the next q data will be added in it by considering the $C_{n-(m+q)}^{q}$ kinds of situations, until all remaining data are added in it or the algorithm stop criteria is attended, viz., the fitting-goodness is saturating.

Stop criteria

TCGPR starts from a specific initial dataset and dynamically expanded it follows the sequence forward ideology. During the dynamic expansion process, a series of datasets with different sizes will be produced, viz., a m-size initial dataset, a (m+p)-size first expanded dataset, …, a (m+kp)-size k-th expanded dataset. Pearson correlation coefficient, R, of these datasets on LOOCV are recorded along the expansion path, denotes as $R_0, R_1, \cdots, R_k$, respectively.

If $R_{k+1} < (1-\eta) \times max\{R_0, R_1, \cdots, R_k\}$, TCGPR will stop this path and output the datasets after k-th expansion. Otherwise, the initial dataset will be continuously expanded until all of the remaining data are used up. Where η is a tolerance coefficient, recommend value is 0-0.3, default value is 0.2.

Initial dataset

Initialize a better “initial dataset” is crucial to TCGPR. Initial dataset is the start part of a dynamic expansion chain, which will influence the ultimate result of data preprocessing. At the beginning, the capacity, m, of the initial dataset should be assigned by researchers according to the characteristics of the studied data. TCGPR provides two independent interfaces to confirm the initial dataset.

Interface 1：

If researchers have strong domain knowledge to their studies, and/or want to ensure that certain data must coexist for some specific intentions. The initial dataset can be assigned directly by the user.

Interface 2：

If researchers are not fully familiar with the data logic and have no additional artificial constraints on divided dataset. For a n-data dataset, TCGPR will brutely searches all C_n^m kinds of potential initial dataset, and selects the best one with highest expected decrease compared with an empty dataset. Where we define the GGMF of an empty dataset is $+\infty$. Similarly, if C_n^m is too large to calculate by the computer (default the upper limit is 10^6), 10^6 kinds of situations will be randomly sampled from all C_n^m cases. And the best candidate will be picked out over the 10^6 situations.

TCGPR employs the sequence forward ideology.