A GAUSSIAN PROCESS EMBEDDED FEATURE SELECTION METHOD BASED ON AUTOMATIC RELEVANCE DETERMINATION

Yushi Deng[a], Mario Eden[a], Shuxing Cheng[b], Haijing Gao[b], Selen Cremaschi[a, *]
[a] Chemical Engineering Department, Auburn University, Auburn, AL 36849, USA
[b] Chevron Technical Center, Houston, TX, USA

Abstract

In Gaussian Process (GP), when the Automatic Relevance Determination (ARD) structured kernel function is applied, each input feature is assigned a corresponding length scale. The feature importance is inversely proportional to the corresponding length scale, and the feature selection can be performed based on the feature importance ranking result. Among the ARD-based feature selection methods, no uniform score exists for quantifying the output variation explained by feature subsets. This study proposes a feature selection approach based on the GP mean function derivative decomposition. A cumulative feature importance score titled derivative decomposition ratio (DDR) measuring the cumulative feature importance of feature subsets is introduced. The DDR is used to determine the optimal feature subset, which is the most relevant feature subset with good predictive performance. The approach is applied to identify relevant dimensionless inputs for predicting liquid entrainment fraction in two-phase flow in a hybrid model. The feature selection result from DDR is compared with the feature importance results obtained by normalizing the average partial derivatives of the output over the inputs. By iteratively adding features following a descending order feature importance, the relation between the cumulative feature importance of feature subsets and the performance of the model built using the corresponding feature subset is investigated. The results reveal that the proposed feature selection approach can identify the optimal feature subset for the case study. The hybrid model built using the optimal subset has an identical Root Mean Squared Error (RMSE) as the model built with full feature space.

Keywords

Gaussian process, Machine learning, Feature selection, Entrainment fraction, Automatic relevance determination.

Introduction

To simulate and predict the system behavior, data-driven models are developed based on the relationship inferred from process data. The feature space of the process data is composed of measurable properties of the system being observed (Chandrashekar and Sahin, 2014). When a model is built to map the input to the output space, the presence of redundant input variables could result in superfluous computational time, model overfitting, and poor model performance (Jović et al., 2015). The extra model complexity due to irrelevant input features also impedes visualization. Building a model with relevant input features reduces model complexity and computational resources required to build and execute the model, helps to understand the data, and improves the model performance (Carlos Molina et al., 2002). Irrelevant feature removal is

[*] selen-cremaschi@auburn.edu

considered a data preprocessing step to ensure the model's efficiency and effectiveness (Liu, 2010).

Feature selection is defined as reducing the dimensionality of data to improve machine learning performance (Chandrashekar and Sahin, 2014; Liu, 2010). Generally, feature selection methods can be classified into three categories, filter, wrapper, and embedded methods (Guyon and Elisseeff, 2003). The filter methods employ a score measuring the statistical relevance between input and output features directly from data without building models (Kumar and Minz, 2014). The input features with an importance score below a predefined threshold are identified as irrelevant and pruned (Chandrashekar and Sahin, 2014). The filter methods are computationally efficient because they do not require building a model to link the inputs and the outputs (Guyon and Elisseeff, 2003).

Wrapper methods pick the optimal feature subset by evaluating the performance of a particular machine learning strategy. A specific subset generation strategy is employed to construct subsets of input features, and the optimal subset is determined by assessing the performance of the model built using the subsets (El Aboudi and Benhlima, 2016). Because the model-building process is time-consuming, the wrapper methods are computationally more expensive than filter methods (Carlos Molina et al., 2002).

The feature selection is incorporated into the model training process for the embedded methods. The feature importance is inferred from the model parameters. Then, the relevant features are picked according to the feature importance ranking results. Embedded methods reduce the computational expense compared to wrapper methods by eliminating the need for generating many subsets and building the corresponding models (Chandrashekar and Sahin, 2014). Some embedded feature selection approaches are developed for specific machine learning techniques. For example, the Lasso regression assigns zero weights to irrelevant features by introducing regularization terms (Tibshirani, 1996). In Random Forests (RFs), the feature importance is estimated from the Mean Decrease in Impurity (MDI) (Breiman, 2001). For Neural Networks (NNs), the feature importance is quantitatively evaluated using the output gradient over the inputs (Varma, 2020).

Gaussian Process (GP) is a non-parametric Bayesian regression characterized by its mean and covariance (Williams and Rasmussen, 2004). For GP models, a commonly used feature selection method is sensitivity analysis. Blix and Eltoft (2018) evaluate the feature importance by integrating the squared partial derivative of the GP mean function over inputs. Piironen and Vehtari (2016) measure the feature importance using the Kullback-Leibler divergence (KLD) change of the posterior distribution when adding a feature to train the model. Features contributing more to the KLD change are considered more important. Paananen et al. (2019) assess the feature importance by observing the KLD variability of the posterior distribution and the posterior mean prediction when changing the input value. Another feature selection approach is the automatic relevance determination (ARD), which infers the relative input feature importance from the inverse of input-dependent length-scale in the kernel function (Paananen et al., 2019). Williams and Rasmussen (1995) detected the irrelevant inputs by ranking the features using the relative feature importance inferred from ARD.

One limitation of the feature selection approaches for GP models is that the selection is based on the ranking results without a standard that classifies the relevant and irrelevant features. Ghoshal and Roberts (2016) proposed adding an irrelevant feature as a baseline to the GP model input space to overcome this limitation. They recommended that features with importance values two orders of magnitude greater than the baseline feature be considered relevant to the output. However, none of the existing GP-based feature selection methods quantifies the cumulative feature importance. If the cumulative feature importance were quantified, the contributions of different feature subsets to the cumulative could be assessed, enabling the identification of the optimal feature subset.

This study proposes a new feature selection approach based on the cumulative feature importance evaluation. The change in the output is represented using the squared magnitude of its total derivative. By decomposing this derivative into partial derivatives, the change in the output caused by varying each input variable is quantified by the squared magnitude of the partial derivative over each input feature. As the squared partial derivative magnitude is cumulative, the percentage of output change explained by each input feature is represented by the ratio of squared magnitude of partial derivative over the total derivative, which we define as derivative decomposition ratio (DDR). The DDR provides information for ranking the input features and enables selecting features at a specified cumulative feature importance value. The proposed feature selection approach prunes the irrelevant variables and reduces the computational burden without reducing the model performance. We applied the proposed feature selection approach to a parallel structured hybrid model developed for estimating the liquid entrainment fraction in two-phase flow. The proposed feature selection method was compared to normalized sensitivity (NS) derived from the sensitivity analysis introduced by Blix and Eltoft (2018). The next section describes the proposed feature selection methodology. Section 3 presents the application of the methodology on a hybrid model that predicts liquid entrainment fraction in two-phase flow. The results and discussion are summarized in Section 4, followed by conclusions and future directions in Section 5.

Methodology

Gaussian Process with ARD kernel

The Gaussian Process (GP) modeling is a supervised learning method with a theoretical basis in statistics. A GP is characterized by its mean and covariance functions (Williams and Rasmussen, 2006). Each input variable is assigned a length scale l_h in the kernel function with ARD structure to infer the feature importance. As an example, the squared exponential (SE) kernel with ARD structure of two points $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$ are shown in Eq. (1).

$$k(\boldsymbol{x}_p, \boldsymbol{x}_q) = \sigma_f^2 \, exp\left(-\sum_{h=1}^{d} \frac{(x_p^h - x_q^h)^2}{2{l_h}^2}\right) + \sigma_n^2 \Delta_{pq} \qquad (1)$$

In Eq. (1), d is the dimension of input $\boldsymbol{x}$, l_h is the characteristic length scale corresponding to the *h^th^* dimension of input $\boldsymbol{x}$, σ_n^2 is the output variance, the parameter σ_f^2 is the output-scale amplitude, and Δ_{pq} is Kronecker delta, which is one if $p = q$ and zero otherwise. The hyperparameters for the GP in Eq. (1) are $\theta = \{\sigma_f, \sigma_n, l_1, l_2, \ldots, l_d\}$, and they are estimated by the Maximum Likelihood Estimation (MLE) (Rasmussen, 2003). This paper used a Python-based package, GPy (GPy, 2012), to estimate the hyperparameters.

The values of the mean function and its variance for the unseen test data X_* are obtained by calculating the posterior distribution using the training data set (X, δ) via Eq. (2).

$$\begin{gathered} f_*|X, \delta, X_* \sim N\left(\overline{f_*}, cov(f_*)\right) \\ \overline{f_*} = K(X, X_*)^T [K(X, X) + \sigma_n^2 I]^{-1} \delta \qquad (2) \\ cov(f_*) = K(X_*, X_*) - K(X, X_*)^T [K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*) \end{gathered}$$

where f_* is the outputs for the test data given the inputs X and outputs δ of the training data and the inputs of the test data X_*. The outputs follow a normal distribution with $\overline{f_*}$ as mean and $cov(f_*)$ as variance. In Eq. (2), K is the covariance matrix, and I is the identity matrix.

Derivative decomposition ratio (DDR)

Blix and Eltoft (2018) analyzed the feature importance by evaluating the variation of the GP mean function in the h^{th} direction. They defined the sensitivity of the h^{th} input s_h as the integral of the squared partial derivative over N number of training samples. The empirical estimate of the sensitivity is shown in Eq. (3).

$$s_h = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\partial \phi(x_n)}{\partial x_n^h}\right)^2 \qquad (3)$$

where $\phi(\boldsymbol{x})$ is the predicted mean function, $\boldsymbol{x}_n$ is the n^{th} input vector, and x_n^h is the h^{th} input in $\boldsymbol{x}_n$. If the mean function $\overline{f_*}$ is substituted into Eq. (3), the resulting empirical estimate of the GP mean sensitivity is obtained and shown in Eq. (4).

$$s_h = \frac{1}{N} \sum_{q=1}^{N} \left(\sum_{p=1}^{N} \frac{\alpha_p (x_p^h - x_q^h)}{l_h^2} k(\boldsymbol{x}_p, \boldsymbol{x}_q)\right)^2 \qquad (4)$$

where $\alpha_p = (K(X, X) + \sigma_n^2 I)^{-1} \delta_p$ is the weight for the p^{th} sample of the GP mean function, δ_p is the p^{th} sample output, and x_p^h and x_q^h are the h^{th} input of $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$. Although the feature selection can be performed by ranking the mean sensitivity values defined by Eq. (4), they only provide a relative feature importance score and cannot be directly used to assess the total contribution of the selected (or individual) features to the changes in the output.

Here, we adopt the differential form of the total derivative to evaluate the individual contribution of each feature to the changes in the output. For example, in Figure 1, the total derivative $\frac{d\phi(x)}{dx}$ is decomposed into partial derivatives from three inputs. Given a mean function $\phi(\boldsymbol{x})$ with H inputs (i.e., $\boldsymbol{x}$ has H elements), the total derivative is the sum of the H partial derivatives, as shown in Eq. (5). The relationship between magnitudes of the total derivative and partial derivatives is given in Eq. (6). Based on Eq. (6), if the output change caused by each input is represented by the magnitude of each squared partial derivative, total importance from the total feature space sums up to the squared magnitude of the total derivative. The ratio of each squared magnitude of partial derivative to the total derivative is proportional to the contribution of a unit change in each input to the change in the output. We define this ratio as the derivative decomposition ratio (DDR) and estimate it using Eq. (7). We then use the sum of corresponding DDRs to assess the cumulative contribution of a feature subset.

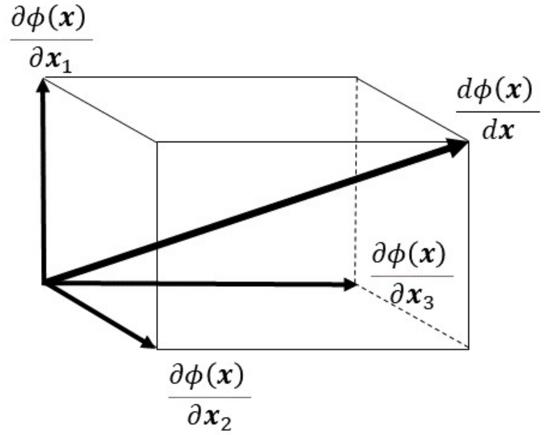
$\frac{\partial \phi(\boldsymbol{x})}{\partial \boldsymbol{x}_1}$
$\frac{d\phi(\boldsymbol{x})}{d\boldsymbol{x}}$
$\frac{\partial \phi(\boldsymbol{x})}{\partial \boldsymbol{x}_3}$
$\frac{\partial \phi(\boldsymbol{x})}{\partial \boldsymbol{x}_2}$

Figure 1. Example of total derivative decomposition

$$\frac{d\phi(x)}{dx} = \sum_{h=1}^{H} \frac{\partial \phi(x)}{\partial x_h} \qquad (5)$$

$$\left|\frac{d\phi(x)}{dx}\right|^2 = \sum_{h=1}^{H} \left|\frac{\partial \phi(x)}{\partial x_h}\right|^2 \qquad (6)$$

$$DDR_h = \frac{1}{N} \sum_{i=1}^{N} \frac{\left|\frac{\partial \phi(x)}{\partial x_h^i}\right|^2}{\left|\frac{d\phi(x)}{dx_h}\right|^2} \qquad (7)$$

The DDR_h calculates the fraction of change in $\phi(\boldsymbol{x})$ caused by a unit change in the h^{th} feature at each sample and averages this fraction over N samples. In Eqns. (5)-(7), $\boldsymbol{x}_h$ represents the h^{th} feature, and x_h^i represents the h^{th} feature of the i^{th} sample.

Another approach to estimating the contribution of each feature to the output variation is normalizing the sensitivity (Blix and Eltoft, 2018). For the h^{th} feature, we define normalized sensitivity (NS) in Eq. (8).

$$NS_h = \frac{s_h}{\sum_{h=1}^{H} s_h} \tag{8}$$

The DDR is calculated by estimating the feature contribution ratio at each sample and then averaging this ratio over all samples. In contrast, the NS is calculated by first averaging the contribution of each feature and then calculating a ratio of this average effect. The DDR and the NS are cumulative and sum up to one when the GP model is trained using the full feature space. Therefore, the total contribution of a feature subset can be estimated by the score calculated by summing the individual DDR or NS values of the features within the subset. Subsets with high cumulative feature importance scores are expected to have predictive importance and performance similar to the full feature set. The cumulative feature importance scores enable the determination of the optimal (i.e., smallest) feature subset given a predefined threshold.

Case Study

The proposed DDR-based feature selection approach is applied to determine the optimal feature subset for a hybrid model estimating the liquid entrainment fraction (Deng et al., 2022). A hybrid model comprises a first principle and a data-driven model (Von Stosch et al., 2014). The data-driven model is a GP regression model in this study. The GP regression model is trained using a dataset of liquid entrainment measurements given pipe diameter (*ID*), inclination angle (θ), gas density (ρ_G), liquid density (ρ_L), gas viscosity (μ_G), liquid viscosity (μ_L), gas-liquid surface tension (σ), superficial gas velocity (v_{SG}), and superficial liquid velocity (v_{SL}). The dataset is compiled from open literature, and the complete list of its sources can be found in Deng et al. (2022). It contains 1,662 liquid entrainment measurements in small-scale laboratory settings. To extend the model from laboratory to field scale, dimensional analysis is introduced to update the inputs from dimensional variables to dimensionless numbers (DNs). The proposed DDR is then applied to prune the irrelevant DNs.

Hybrid model for liquid entrainment fraction

Deng et al. (2022) developed a hybrid modeling approach to estimate the entrainment fraction and its uncertainty. A set of semi-mechanistic and empirical models developed for multiple flow orientations were combined with a data-driven model. A Gaussian Process (GP) model (Williams and Rasmussen, 2006) was built as the data-driven model to estimate the model discrepancy, defined as the difference between the experimental measurement and the semi-mechanistic model prediction. In this study, one of the hybrid submodels from Deng et al. (2022) (which employs Zhang et al. (2003) model as the semi-mechanistic model) is adopted to investigate the effectiveness of the proposed feature selection approach.

Dimensional analysis

Dimensional analysis (DA) is introduced to extend the model to regions where the experimental data is scarce or not available. DA is a common approach for grouping and reducing the number of phenomena relevant input variables in chemical engineering (Cheng and Cheng, 2004). Dimensional analysis develops DNs and updates the dimensional inputs with the DNs. Inference from the process response and DNs may provide a deeper understanding of the process mechanism. Furthermore, using DNs as inputs may extend the applicability of a model from laboratory to field scale (Ruzicka, 2008). In this study, the GP model inputs are replaced by DNs identified using the method proposed by Dai et al. (2022). The method yielded 49 unique DNs identified from the dimensional variables. We filtered the DNs to remove highly correlated ones using the Pearson correlation coefficient. From a set of DNs with correlation coefficients greater than ±0.95, only one is kept in the filtered set. The filtered set included 35 DNs as potential inputs to build the GP model.

Computational experiment details for evaluating DDR-based feature selection approach

The DDR is applied to remove irrelevant inputs given the complete input feature set, 35 DNs, for predicting model discrepancy. After building a GP, the percentage of output variation explained by subsets of features is estimated using the corresponding cumulative DDR. Subsets with a high cumulative DDR are expected to result in a better GP model accuracy. A computational experiment is designed by training a series of models with iteratively expanded feature subsets to study the relationship between the cumulative feature importance estimated as the sum of DDRs and the resulting model performance. The steps of the experiment are outlined in Figure 2.

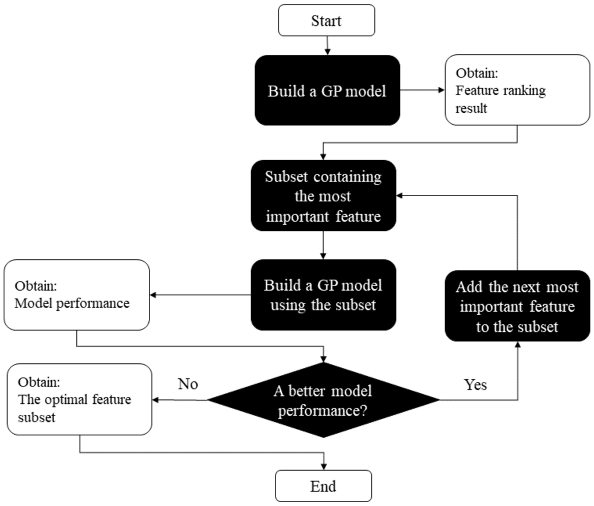
Start
Build a GP model
Obtain:
Feature ranking result
Subset containing the most important feature
Obtain:
Model performance
Build a GP model using the subset
Add the next most important feature to the subset
Obtain:
The optimal feature subset
No
A better model performance?
Yes
End

Figure 2. The experimental procedure used to train GP models with forward feature selection

A GP model trained with the full feature set is used to estimate the feature importance for each input. The computational experiments employed a wrapper method with forward selection (Rodriguez-Galiano et al., 2018) (Figure 2). Within each iteration, one feature is added to the subset following the descending order of the feature importance. The data is split into 70% and 30% as training and test sets. The GP model performances on both sets are measured using the Root Mean Squared Error (RMSE) defined in Eq. (9), where y_j^e and $\hat{y}_j^e$ represent the j^{th} experimental measurement and prediction, respectively. J is the total number of samples. Monte Carlo Cross Validation (MCCV) with 30 replicates is used to calculate an average RMSE for each feature set.

$$RMSE = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left\|y_j^e - \hat{y}_j^e\right\|^2} \qquad (9)$$

Experiments are performed twice, once using DDR (calculated using Eq. (7)) and once using NS (calculated using Eq. (8)) as the metric to assess feature importance. The first experiment set investigated if adding a feature with higher DDR contributes more to the model performance enhancement. By tracking the model performance in subsequent iterations, the experiment also explores if models trained using feature subsets with higher cumulative feature importance give lower RMSEs. The optimal subset of features is determined when the average model RMSE reaches a steady low value or is the lowest. The second experiment set investigates using NS as the metric.

Results and Discussion

The feature selection results using DDR and NS are summarized in Figures 3 and 4. The red lines show the change in cumulative DDR or NS as the number of DNs used as inputs for training the GP model increases. The blue and green lines are the RMSEs for training and test data sets.

Figure 3 shows that as the number of DNs increases in the feature subset, the average RMSEs for training and test data decrease while the average cumulative DDR approaches one. When the subset includes nine or more features, the average RMSEs and the cumulative DDR reach a steady value, suggesting that adding more features does not significantly improve the RMSE. In fact, the RMSE of the model trained using the full feature set is almost equal to the RMSE of the GP model trained using the selected nine features. These results suggest that feature subsets with a cumulative importance score close to one (the selected nine features) yield GP models with very similar performance as the full feature set. Features with small DDRs do not contribute to the model performance. These results suggest that the feature subsets with desired model performance can be constructed using the cumulative DDR.

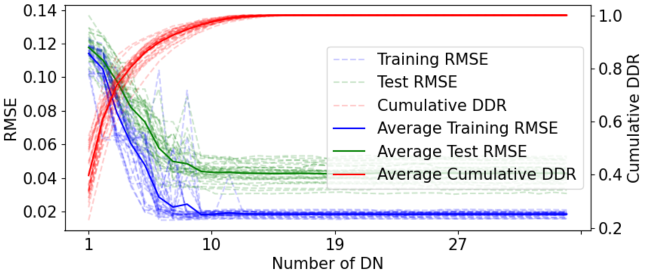

Figure 3. Feature selection using DDR and resulting model performance

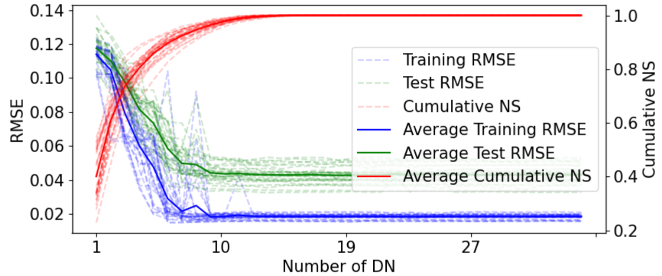

Figure 4. Feature selection using NS and resulting model performance

Figure 4 shows the model RMSEs for training and test data as the number of features increases following the descending order feature importance provided by NS. When the number of features reaches nine, the addition of a new feature does not further enhance the model performance. A comparison of Figures 3 and 4 reveals that although the MCCV replicate results of DDR and NS are not identical, the average RMSEs and the optimal feature subset size chosen using NS are the same as DDR. The difference between NS and DDR is that the NS calculates the ratio of the mean partial derivative while the DDR calculates the mean of the partial derivative ratio for the partial derivative of each feature at each sample. Theoretically, the NS would provide a more robust feature importance score when outliers are present in the data. In this case study, there is no significant difference between the DDR and the NS results.

In Figures 3 and figure 4, it can be seen that the MCCV RMSEs fluctuate for subsets containing six to eight features. This is because DDR and NS are developed from derivatives, which depend on sample locations. The randomness in data splits results in differences in sample location-dependent feature importance, which drives the feature ranking volatility and feature addition order within each cross-validation replicate. As the number of features increases, both the cumulative feature importance curve and the RMSE curves become flattened, which suggests that features with higher DDR and NS reduce the RMSE of the model more than features with lower feature importance score. When the optimal feature subset is obtained, the cumulative DDR and NS are both above 0.99, and the optimal subsets determined from the two methods are close to each other. We consider the smallest subset with a cumulative score higher than 0.99 the optimal subset.

Conclusions and Future Directions

This study introduced a Gaussian Process embedded feature selection approach. The approach focuses on the decomposition of the feature importance and develops a ratio, DDR, which represents the percentage of output change explained by the corresponding feature. The DDR is cumulative, and the total contribution from a subset of features can be obtained by summing up the DDRs of features in the subset. The cumulative feature importance score, the sum of DDRs, enables determining the optimal feature subset for any predefined cumulative feature importance score. The approach is compared with another cumulative feature importance score defined as normalized sensitivity (NS). Both approaches were applied to detect relevant inputs for a liquid entrainment fraction prediction model. The DDR-based feature selection approach reduced the feature set size by 70% without reducing the model RMSE. The optimal feature subset can be determined when the cumulative DDR approaches a high value, which was 0.99 in the case study. The feature selection results were close for DDR and NS-based approaches. Future work will investigate the effectiveness of the DDR and NS-based feature selection methods more extensively.

Acknowledgments

Chevron Technical Center provided financial support for this work.

References

Blix, K., Eltoft, T., 2018. Evaluation of Feature Ranking and Regression Methods for Oceanic Chlorophyll-a Estimation; Evaluation of Feature Ranking and Regression Methods for Oceanic Chlorophyll-a Estimation. IEEE J. Sel. Top. Appl. EARTH Obs. Remote Sens. 11, 1403. https://doi.org/10.1109/JSTARS.2018.2810704

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Carlos Molina, L., Belanche, L., Nebot, À., 2002. Feature selection algorithms: A survey and experimental evaluation. Proc. - IEEE Int. Conf. Data Mining, ICDM 306–313. https://doi.org/10.1109/icdm.2002.1183917

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Comput. Electr. Eng. 40, 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Cheng, Y.T., Cheng, C.M., 2004. Scaling, dimensional analysis, and indentation measurements. Mater. Sci. Eng. R Reports 44, 91–149. https://doi.org/10.1016/J.MSER.2004.05.001

Dai, W., Mohammadi, S., Cremaschi, S., 2022. A hybrid modeling framework using dimensional analysis for erosion predictions. Comput. Chem. Eng. 156, 107577. https://doi.org/https://doi.org/10.1016/j.compchemeng.2021.107577

Deng, Y., Avila, C., Gao, H., Mantilla, I., Eden, M.R., Cremaschi, S., 2022. A Hybrid Modeling Approach to Estimate Liquid Entrainment Fraction and its Uncertainty. Comput. Chem. Eng.107796. https://doi.org/https://doi.org/10.1016/j.compchemeng.2022.107796

El Aboudi, N., Benhlima, L., 2016. Review on wrapper feature selection approaches. Proc. - 2016 Int. Conf. Eng. MIS, ICEMIS, 2016. https://doi.org/10.1109/ICEMIS.2016.7745366

Ghoshal, S., Roberts, S., 2016. Extracting predictive information from heterogeneous data streams using Gaussian Processes. Algorithmic Financ. 5, 21–30. https://doi.org/10.3233/AF-160055

GPy, 2012. GPy: A Gaussian process framework in python.

Guyon, I., Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3, 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

Kumar, V., Minz, S., 2014. Feature selection: a literature review. SmartCR 4, 211–229.

Liu, H., 2010. Feature Selection, in: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning. Springer US, Boston, MA, pp. 402–406. https://doi.org/10.1007/978-0-387-30164-8_306

Paananen, T., Piironen, J., Andersen, M.R., Vehtari, A., 2019. Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution. AISTATS 2019 - 22nd Int. Conf. Artif. Intell. Stat. 89.

Piironen, J., Vehtari, A., 2016. PROJECTION PREDICTIVE MODEL SELECTION FOR GAUSSIAN PROCESSES, in: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing. pp. 1–6.

Rasmussen, C.E., 2003. Gaussian processes in machine learning, in: Summer School on Machine Learning. Springer, pp. 63–71.

Rodriguez-Galiano, V.F., Luque-Espinar, J.A., Chica-Olmo, M., Mendes, M.P., 2018. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. Sci. Total Environ. 624, 661–672. https://doi.org/10.1016/j.scitotenv.2017.12.152

Ruzicka, M.C., 2008. On dimensionless numbers. Chem. Eng. Res. Des. 86, 835–868. https://doi.org/https://doi.org/10.1016/j.cherd.2008.03.007

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288.

Varma, V., 2020. Embedded methods for feature selection in neural networks. arXiv Prepr. arXiv2010.05834.

von Stosch, M., Oliveira, R., Peres, J., Feyo de Azevedo, S., 2014. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Comput. Chem. Eng. 60, 86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008

Williams, C.K.I., Rasmussen, C.E., 2006. Gaussian Processes for Machine Learning. MIT press Cambridge, MA.

Williams, C.K.I., Rasmussen, C.E., 2004. Gaussian Processes for Machine Learning, International Journal of Neural Systems. https://doi.org/10.1142/S0129065704001899

Williams, C.K.I., Rasmussen, C.E., 1995. Gaussian Processes for Regression. Adv. Neural Inf. Process. Syst. https://doi.org/10.1016/0165-4896(94)90008-6