h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Gaussian Process Models for the Prediction of Chemical Experiment Outcomes

verfasst von: Tamara Döß   Referent: Prof. Dr. Horst Zisgen   Koreferent: Prof. Dr. Markus Döhring   Externer Betreuer: Dr. Christopher Geis

Ausgabedatum: 06.07.2022   Abgabedatum: 21.12.2022

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

Motivation and Research Topics

In pharmaceutical companies, lab experiments are conducted to develop a robust process to efficiently manufacture active ingredients. These experiments are expensive and time-consuming [3, 5] and should therefore be reduced to a minimal necessary number. This can be achieved by using Gaussian Processes (GPs), especially GP coregionalisation models, to predict experiments outcomes. The predictions are used by lab-heads to decide which experiments to conduct next.

The thesis investigates several GP model versions and the data preprocessing of the pharmaceutical use case in order to improve the predictive quality of the implemented GP model. The following to topics are investigated in detail:

1. Evaluate the results and training speed of several types of multiple-output (coregionalisation) GP models with the aim of using the model with best quality:
   - Select suitable models for the use case.
   - Implement the selected models with the GPflow python package.
   - Produce evaluation metric results to allow a proper comparison.
2. Evaluate the experiment clustering process and establish a similarity score to identify similar experiments. The aim is an increase of input data points for model training.

Related Work

No related work is found that handles the specifics of the use case of the thesis. However, several related topics are found:

1. GPs for predictions in a general chemical context and for chemical reactions.
2. GP coregionalisation models in a general context.
3. Machine Learning methods for the prediction of chemical reactions and reaction conditions.
4. Machine Learning methods for the optimisation of reaction yield.

Discussion and Conclusion

Separate discussions can be given for the two research topics:

1. Results are based on tests for two experiment clusters of different size and show that the implemented model versions do not lead to significant differences in prediction quality. Therefore, no recommendations about which model version to use in production are possible. A conclusion is that either the models exploit the little information contained in the training data and already result in the best possible predictive quality or the model type of GPs is not suitable to handle the provided data and infer knowledge.
2. Increasing the training data size by skipping the experiment clustering process or using the implemented similarity score led to a deterioration of the predictive quality. This leads to the conclusion that the experiment clustering based on defined business rules is useful to group the experiments in clusters that lead to better modelling results as well as shorter training times. Whether an increases data set of comparable experiments leads to better results can only be analysed once new comparable experiments are conducted to expand the existing clusters.

A further conclusion is that the data quality should be analysed in detail to decide whether too many features are removed during data preprocessing or the degree of randomness that may have an influence on the target values is to high.
Further research topics are recommended:

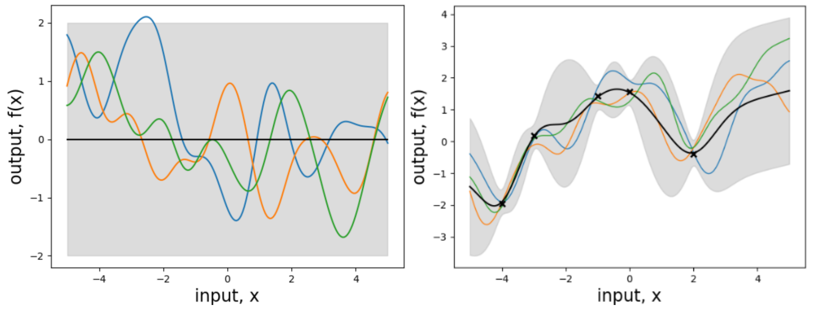
1. Benchmarking of predictions by conducting new experiments and comparing results to the predictions.
2. An in-depth analysis of further GP model settings.
3. An in-depth analysis of the available data to improve data preprocessing.

Gaussian Processes Regression Models

A Gaussian Process $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$ is a stochastic process and thus a collection of random variables where any finite set of random variables has a joint Gaussian distribution [6, 2]. The function value $f(x_i)$ is not a scalar but a Gaussian distribution so that predictions of GPs are also Gaussian distributions defined by mean and variance. $m(x)$ is the mean function, $k(x, x')$ the kernel function of a GP. The predictive function of a GP is given by the posterior distribution (fig. 1)

$$\begin{aligned} \mathbf{f}_* | X, \mathbf{y}, x_* &\sim \mathcal{N}(\bar{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*)) \\ \bar{\mathbf{f}}_* &= K(x_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y} \qquad (1) \\ cov(\mathbf{f}_*) &= K(x_*, x_*) - K(x_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, x_*). \end{aligned}$$

with $K(X, X)$ being the kernel matrix for training data and x_* being a new data point. The choice of kernel influences the shape of drawn GP samples [4]. The thesis uses the radial basis function kernel.



(a) (b)

Fig. 1. GP prior (a) and posterior with noisy data (b) and drawn samples. The shaded area displays the model uncertainty defined by the predicted variance [6].

Coregionalisation Models

GP regression models can be used to make predictions for a number of P target variables. One group of multi-output GP models are coregionalisation models that are able to detect correlations between targets [1]. This correlation is detected by a type of kernel called *Sum of Separable Kernels* which is defined as

$$K(x, x') = \sum_{q=1}^{Q} k_q(x, x')\mathbf{B}_q. \qquad (2)$$

with $\mathbf{B}_q$ being a $P \times P$ *coregionalisation matrix* encoding the target correlations.

This kernel is used by coregionalisation models. Three types of coregionalisation models are investigated that are distinguished by the choice of Q and R_q defining the rank of $\mathbf{B}_q$ [1]. Both parameters determine the model complexity. Table 1 shows the specifics of the coregionalisation models.

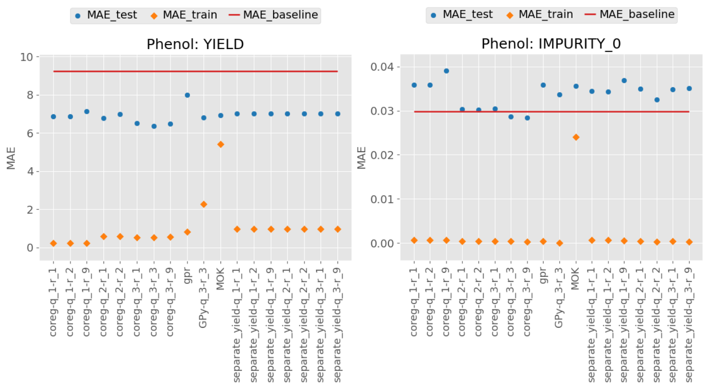
Model Name	Q	R_q
Linear Coregionalisation Model	> 1	> 1
Intrinsic Coregionalisation Model	$= 1$	> 1
Semiparametric Latent Factor Model	> 1	$= 1$

Table 1. Specifics of coregionalisation models

Gaussian Process Model Comparison

In total 15 coregionalisation model versions are implemented. Additionally, one general multi-output model without a coregionalisation matrix and separate single-output models for all targets are implemented. Therefore, 17 models are compared to the existing GPy coregionalisation model. The comparison is applied to two data sets of different size and investigates the standardised mean absolute error (MAE), non standardised MAE and Log Loss evaluation metric values. This summary shows MAE results for the data set of size 20.

The plots in figure 2 show a baseline error done by a model always predicting the training data mean and variance as a red horizontal line. It can be seen that the trained GP models lower the baseline error for the yield target by approx. 12% on average for test data but only two models slightly fall below the baseline error of the impurity target. This shows that the models do not result in a consistent quality for all targets. Furthermore, it can be seen that the MAE values of most models are similar considering the target value range. Especially for the yield target that has a value range from 50.8% to 101.2% the differences in predictive quality are not expressive enough to allow decisions as to which model to use in production.



(a) (b)

Fig. 2. MAE results for the yield target (a) and an impurity target (b).

Experiment Clustering Evaluation

Before training models, the data provided by conducted experiments is divided into clusters depending on the experiment setup. This process decreases the training data set size as models are trained for each cluster separately. Therefore, a new clustering process based on an similarity score is implemented. The similarity score measures the similarity of operation sequences between two experiments. Modelling results of the original clustering process (left), no clustering process (right) and the new clustering process (middle) are compared and the MAE results are displayed in figure 3. The original clustering process results in the best predictive quality regarding the MAE value.

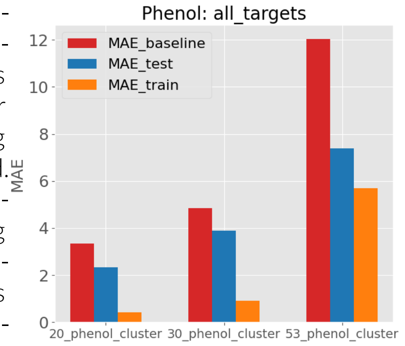


Fig. 3. MAE for different data sizes.

References

[1] Alvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for vector-valued functions: a review. *Found. Trends Mach. Learn.*, 4(3):195–266.

[2] Bishop, C. M. (2006). *Pattern recognition and Machine Learning*. Springer Science+Business Media, New York, NY, 1 edition.

[3] Dolas, R., Siddheshwar, S., Somwanshi, S., Merekar, A., Godge, R., and Pattan, S. R. (2013). Optimization techniques in designing of pharmaceutical dosage form. *International Journal of Current Trends in Pharmaceutical Research*, 1(2):137–143.

[4] Duvenaud, D. (2014). *Automatic model construction with Gaussian processes*. Dissertation, University of Cambridge, Cambridge.

[5] Eyke, N. S., Green, W. H., and Jensen, K. F. (2020). Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *Reaction Chemistry & Engineering*, 5(10):1963–1972.

[6] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Massachusetts.