ABSTRACT

Saving money and time is a ubiquitous industry goal. In the pharmaceutical industry laboratory experiments take up a lot of time and budget [29, 30, 49]. Therefore, it is desirable to reduce the number of executed experiments needed to reach a defined goal. This thesis investigates a pharmaceutical use case with the aim of predicting the outcome of chemical experiments by using Gaussian Process Regression models. These predictions enable experiment simulations and a reduction of conducted experiments. A research of related work shows that the specific use case in connection with the used models has not yet been represented in current literature.
Specifically, multi-output coregionalisation models are implemented using the GPflow python package and compared to an existing coregionalisation model based on the GPy python package. The aim is to analyse whether the predictive quality of the model can be increased by either using updated model settings or increasing the training data set by updating the use case specific data clustering process.
The results show that a GPflow model has superior predictive quality compared to a corresponding GPy model. However, the comparison of Gaussian Process models with different settings all trained using the GPflow package shows that none of the model versions leads to prominently better predictive quality. Additionally, an increase of training data size by using an updated data clustering process does not lead to better modelling results as well.

***Keywords*:** Gaussian Process, Coregionalisation, Kernel Methods, Chemical Experiment Outcomes, Sequence Similarity

ZUSAMMENFASSUNG

Das Einsparen von Geld und Zeit ist ein allgegenwärtiges Ziel von Industrieunternehmen. In der pharmazeutischen Industrie verbrauchen chemische Laborexperimente viel Zeit und Budget [29, 30, 49]. Daher ist es wünschenswert, die Anzahl an Experimenten, die benötigt werden, um ein definiertes Ziel zu erreichen, zu verringern. Die vorliegende Abschlussarbeit behandelt einen pharmazeutischen Anwendungsfall mit dem Ziel, die Ausgänge chemischer Experimente mit Hilfe von Gauß Prozess Regressions Modellen vorherzusagen. Die Vorhersagen erlauben Simulationen von Experimenten und damit eine Reduzierung der tatsächlich ausgeführten Experimente. Eine Recherche verwanter Arbeiten zeigt, dass die Themen des behandelten Anwendungsfalls in Verbindung mit den genutzten Modellen bisher nicht in der Literatur vertreten sind.
Diese Abschlussarbeit implementiert multi-output Coregionalisation Gauß Prozess Modelle mit dem GPflow Python Paket und vergleicht diese mit einem bereits vorhandenen Modell, das mit dem GPy Python Paket erstellt wurde. Das Ziel ist herauszufinden, ob die Vorhersagequalität verbessert werden kann. Die Verbesserung kann durch andere Modell Einstellungen oder eine Erhöhung der Anzahl an Trainingsdaten erreicht werden.
Die Ergebnisse zeigen, dass das GPflow Modell eine höhere Vorhersagequalität erzielt als das entsprechende GPy Modell. Jedoch zeigt der Vergleich zwischen verschiedenen Modell Versionen, die alle mit dem GPflow Paket erstellt wurde, keine auffälligen Unterschiede in der Vorhersagequalität. Auch die Implementierung eines neuen Clusterings der Ausgangsdaten, welches in einer höheren Trainingsdatenmenge resultiert, zeigt keine Verbesserung der Vorhersagequalität.

***Schlagworte*:** Gauß Prozess, Coregionalisierung, Kernel Methoden, chemische Experiment-Ausgänge, Folgen-Ähnlichkeit