UNIVERSITY OF
SOUTH CAROLINA

Stacked Modelling Framework

by

Kareem Abdelfatah

Bachelor of Engineering Fayoum University 2006

Master of Computer Engineering Hellwan University 2010

Master of Science University of South Carolina 2017

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Computer Science and Engineering

College of Engineering and Computing

University of South Carolina 2019

Accepted by:

Gabriel Terejanu, Major Professor

John Rose, Committee Member

Marco Valtorta, Committee Member

Jianjun Hu, Committee Member

Andreas Heyden, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Summary

In this work, stacked model of independently trained Gaussian processes, called StackedGP, is proposed as a modeling framework in the context of model compo- sition to address two challenges: component-based modeling under structural uncer- tainty and propagation of uncertainties to quantities of interest. One of the current challenges in component-based modeling arises from the fitted parametric nature of submodels with no information most of the time on the magnitude of the uncertainty of the parameters. Although sampling methods are used to propagate the parametric uncertainty to the quantities of interest, the uncertainty in model predictions may still be underestimated by ignoring model form uncertainty. The proposed model is based on a network of independently trained Gaussian processes accompanied by an approximate scheme to obtain expectations of quantities of interest that require model composition. The proposed model provides estimation of the quantities of interest with quantified uncertainties. This leverages the analytical moments of a Gaussian process with uncertain inputs when squared exponential and polynomial kernels are used. The StackedGP can be extended to any number of nodes and layers and has no restriction in selecting a suitable kernel for the input nodes. In the second chap- ter, the numerical results show the utility of using StackedGP to learn from multiple datasets and propagate the uncertainty to quantities of interest. StackedGP has be used to enhance the prediction of primary responses by creating an intermediate layer of predictions of secondary responses. This comes with a lower computational com- plexity as compared with multi-output methods - and can make use of off-the-shelves Gaussian processes.

The analytical stacked Gaussian process (StackedGP) framework is introduced to one of challenging environmental problems, predict aflatoxin concentrations with quantified uncertainties. Stochastic Gaussian processes are used to model tempera- ture and water-activity inputs to drive the aflatoxin Gaussian process. An analytical scheme is used to calculate the expectations of aflatoxin predictions by propagating the uncertainty through the stacked Gaussian process. Historical field data from three different states are used to validate the model predictions at the Gaussian process component level and at the system level for the entire stacked Gaussian process. A new derivation has been proposed to extend the analytical framework by calculating the covariance of StackedGP between different geographical locations. This helps with updating the prior predictions of StackedGP using field measurements, and pro- vides the opportunity to couple StackedGP with Bayesian optimization techniques to identify regions where to obtain new field measurements. Thus, the proposed StackedGP can be easily extended to integrate other environmental data such as pre- cipitations and humidity. The proposed predictive framework can also be expanded to address the real-

time monitoring of a wide variety of mycotoxins in crops. The information generated by the aflatoxin StackedGP can be used to design a sustainable crop management by reducing the amount of chemicals used to grow the crops.

In the third part of the thesis, we also introduce the stackedGP framework to chemi- cal application. First, we compared linear models versus advanced machine learning models to predict transition state energies (TS) using more than 1330 descriptor combinations by considering a database for adsorption and transition-state energies across metal surfaces (Ni, Pt, Pd, Ru, Rh, Cu) for the decarboxylation and decarbonylation of Propionic acid. The analysis shows that considering bond counts and metal descriptors can help to achieve lower MAE for both linear and non-linear models. Besides, the non-linear complex models cannot achieve statistically significant better results than the best linear ones. In addition, we discuss various elementary reaction grouping approaches which show that the one-model approach can perform similarly to the one-step approach which might be a benefit when reaction data on various metals are missing. The same conclusion has been reached in the missing data study which demonstrated that conventional descriptors (product or reactant energy) will give higher MAE compared to the ones that also use bond count in- formation and metal descriptors. This is shown in the Ridge/GP_RBF models for descriptors Er, Ep1, Ep2, Br, Bp2, E_{CH_3CH} that perform very well. In this work also, we show that the stackedGP can be used to build the pipeline of the prediction of the transition-state energies with very high performance compared to DFT calculations and stacked ridge models. This work can be extended by studying more descrip- tors combinations for predicting transition-state energies. Also, building the volcano plots to study the TOF is the next step for this work. In addition, this work can be extended for other metal surfaces.