Internship report

Hironori Yasuda
Graduate School of Engineering, Department of Applied Physics, D1
(Motome Lab.)

Training destination: Zeon Corporation

In this internship, I conducted searching for hydrocarbons showing potential high growth of Carbon nanotube (CNT) using Machine learning.

【Content】

・ Period: 2020.10.5.-2020.12.10

・ The work was mainly carried out online, but I went to work several times at R&D Center (Kawasaki) for discussions and tours of the research institute. A result report meeting was held on the final day.

【Outline of the theme & the background】

Carbon nanotubes, which are attracting industrial attention, are synthesized by adsorbing hydrocarbons on a catalyst, but their growth rate varies greatly depending on the adsorbed hydrocarbons. The purpose of this research is to search for the optimum hydrocarbon structure where it is expected to obtain a higher yield of CNT by numerical calculation.

In concrete, we simulated the growth of various hydrocarbons on the catalyst, and the optimum hydrocarbon is searched from the obtained data.

First, the data on hydrocarbon reactions are added by first-principles calculations. Then, training data we calculated combined with the data obtained so far, hydrocarbons are estimated using regression analysis, etc., with the aim of improving the accuracy of the estimation.

However, since the structural variation of hydrocarbons increases exponentially with the number of carbon atoms, it is practically difficult to conduct the calculations for all structural patterns in hydrocarbons having a large number of carbon atoms in terms of calculation cost. Therefore, we used Bayesian optimization to focus on searching for hydrocarbon structures that are likely to be the candidates, and try to reduce the calculation cost of training data.

【Method】

First, using the open-source package Phase 0, we confirmed and verified the data obtained by first-principles calculations so far. In Bayesian estimation, analysis by the kernel method was implemented using the Bayesian estimation package Stan using the Monte Carlo method (FIG.1). First, cross-validation was performed with changing the number of Monte Carlo samples for some types of kernel functions so that

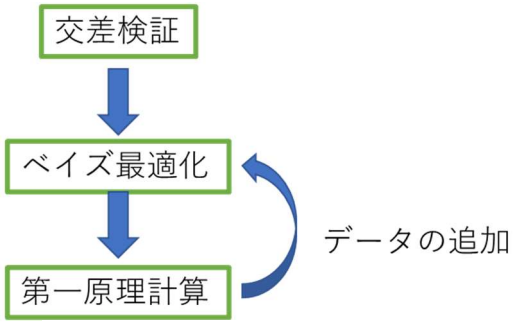

Figure 1：Problem solving flow

the probability model suitable for the data was optimized with respect to the R2 values. At this time, the two-dimensional feature values of hydrocarbon reduced by principal component analysis was used as an input. We also excluded data that are considered statistical outliers. Then, Bayesian optimization was performed using it, and a hydrocarbon having good absorptivity was estimated. Furthermore, first-principles calculations were performed on those hydrocarbons in Phase 0, and the results were added to the training data. This process was repeated several times in this internship.

【Result】

Here, we will introduce the estimation results for housing price data near Boston, which was used as a benchmark. As a kernel function, the 3rd-order Matern function

$$K\left(x, x^{'}\right) = p_1^2\left(1 + \sqrt{3}\frac{d\left(x,x^{'}\right)}{p_2}\right)e^{-\frac{d\left(x,x^{'}\right)}{p_2}}$$

was used and the number of Monte Carlo samples was 2000. The R2 value is as large as 0.89. Figure 2 shows a comparison of the predicted values at this time with the actual data, and it is true that the two results are in good agreement. In addition, figure 3 shows a plot of these predicted values and actual data with a 95% confidence interval, and the actual data was generally within the 95% confidence interval. In the training, this code was used to perform similar calculations on hydrocarbon data.

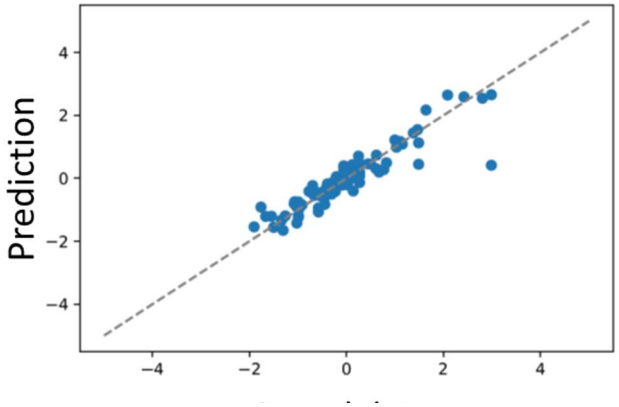

Figure 2：Comparison of Bayesian estimated and measured values for home price data near Boston. The dotted line is the straight line on the vertical axis = horizontal axis.

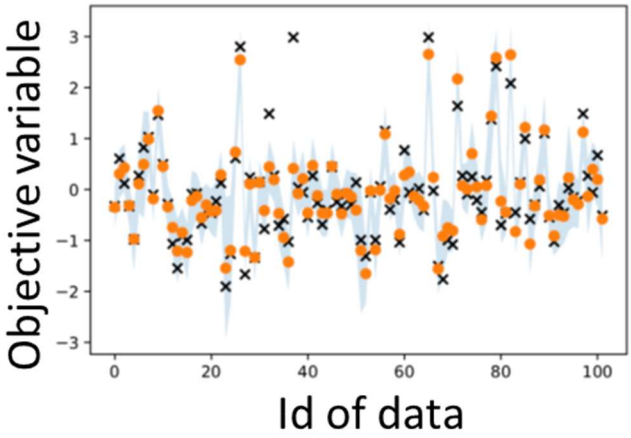

Figure 3：Plot of predicted (x), measured (o) and 95% confidence intervals for each data

【Feedback】

In this internship, I experienced the field where computer science is enthusiastically tackled as a prediction system for experimental research and improvement of manufacturing process in companies using machine learning. In addition, I was able to deepen my understanding of machine learning by studying and applying the knowledge of machine learning that I had never

experienced before. I also had the opportunity to interact with various research teams, and realized the contribution of the analysis I was doing for the actual process. Besides, it was impressive that the members of the computational-analysis team were discussing with the cooperated teams frequently and making consensus. The field of organic chemistry is different from my specialty that is close to the field of inorganic chemistry, but I realized the importance of the communication with those who have different background while maintaining consensus. To sum up, it was a very fulfilling internship.

【Acknowledgment】

First of all, I would like to thank all the members of R&D Center for accepting me while it was difficult to accept due to the coronavirus. While the calculation team was busy, they spent a lot of time setting themes, discussing, and processing data. In addition, the experimental team guided me to the institute and gave opinions from the standpoint of the experimentalists in the discussions.