Predicting adsorption of several greenhouse gases on granular activated carbon

Abstract

Introduction

**Data and methods**

Data collection:

All the data used in this research was collected from published literature. Around 500 paper were collection with around 4000 isotherms, all these studies were published after 2000 and only studies report the properties and the adsorption isotherms plot were collected. Also, all these papers reported some properties of GAC including surface areas, pore volumes, surface/ bulk element composition. As long as these properties being reported, these values were included in our datasets.

Algorithm and feature selection:

Different algorithms including tree based methods and support vector machine and neural networks were tested on all of these collected dataset. 1) randomly split each dataset into training and test by the ration of 8:2; 2) using grid search and cross validation on each training dataset to find the best hyper-parameters; 3) evaluate the model based on the best hyper-parameters on the reserved test datasets; 4) repeating the above processes several times, using the average of prediction errors from each round to select the best algorithm.

Since the total number of the input is relative small, we tried every possible combination of these features during the feature selection part. The focus for this part is to find the best feature set for describing the GAC. The best feature set is selected following 1) randomly split each dataset into training and test datasets by the ratio of 8:2 2) using different input feature and different algorithms to find the best model using grid search method; 2) evaluate the best model under every combination of feature set and algorithms on the reserved test dataset; 4) repeat these steps multiple times and using the averaged prediction error to find the best input feature sets.

In addition to the traditional modeling approaches that used some commonly available feature, we tried to included data point extracted from the reported BET curve. Since, some reported BET curve only covered limited numbers of data point, we first interpolate the collected BET curve to uni-length vectors with different numbers of elements such as 35, 50, and 80. The BET data points vector were then coupled with conditional parameters with/without commonly used physical properties.

Training size on the performance over the model’s performance on the test datasets.

It’s always an ongoing question for machine learning modeling that how many training sample are enough to ensure effectively train a machine learning model. Some study has tried to give an answer by monitoring the performance of models on the test datasets trained with different training sample, the general results is that the more training sample the better the performance of models, however, the improvement become negligible when the number of training sample exceed certain threshold. Our recently study also suggest that the quantity of the training dataset is very important, however, the diversity/quality of the training dataset may play more important role to achieve satisfactory modeling performance. Nevertheless, these results can only be obtained after the model being training, which makes them less helpful to address the original question before the model is trained. To address this question, method that without requiring training model or only using label information will be need. In this study, we developed a kernel-PCA based method to finish this task. PAC is a very useful algorithm that can reduce the dimension of input information but keep the most essential information, the major limitation for linear PCA is that it does not work well on information containing non-linear interaction. The kernel –PCA can effectively extract non-linear interaction by projecting the original information space to higher dimension thorough kernel technic. In the new method, we transpose the input matrix so that the variable changed from each feature to each data record, and then we used kernel PCA obtained from the combined data set of training and test dataset to reconstruct the test set, the reconstruction error for the test data set were monitored by varying the training dataset for a given test dataset. The same PAC- and reconstruction processes were repeated for multiple time.

Separate model and combined model

As suggested by previous studies that building a combined model for a few similar learning task will provide some improvements over one model per task. The four data set was combined into one big adsorption dataset, the above modeling process including model selection and hyper parameters optimization were repeated on the combined dataset.

**Results and discussion**

Comparison among different algorithms and input feature sets.



Figure 1. Comparison among different algorithms using two different input feature sets (left) using only BET; (right) using BET and total pore volume for GAC



Figure 2. Model performance after adding more features for GAC

In this section, all the possible combination between feature sets and algorithm were tried. To the end, several algorithms showed very close modeling performance. Considering both the model performance and the training speed, tree based algorithm-LGBM was selected for all the following exploration. As the different feature combination, the model that used all the available physical properties give the best performance, which also agree with previous studies. A clear trend is that the finer the pore volume being divided, the better the model performance (e.g. BET\_Vmic\_Vmeso > BET\_Vt). However, due to the availability of pore volume data, further test cannot be performed.

**Adding BET curve as the additional input features**

The adding of extracted data points from the reported BET curves provides some significant improvement over the model performance as compared with the best features sets that has been frequently used by existing studies. Specifically, when data points from the corresponding isotherms were added the best models mentioned above, the mean squared error on the test dataset dropped by around 18 percent, similar results were also achieved after adding data points from the BET curve to the feature of the methane models. The more interesting finding by feature importance analysis is that the data point belonging to the beginning and the ending part of the BET curve significantly have higher feature importance than the rest majority points in the middle regions. This interesting phenomena can be explained that the beginning region mostly can reflect the microporous properties of the adsorbent, while the ending part of the BET curve can mostly reflect the total pore volume. As suggested in the IUPCA document, the adsorption of small gas molecule is mostly controlled by the micro-pore region of adsorbents, while the adsorbents under higher pressure is mostly guarded by the total pore volume in the adsorbents. The finding also agrees with some existing studies, for example some studies has reported that the adsorption of CO2 and methane can be well correlated with the micro-pore volume under low pressure, under much lower pressure, theses adsorption can be better correlated with much finer region of the micro-pore volume. This finding well suggested that the insufficiency in modeling that only used the total pore volume and surface areas as well the complexity of the effects of micro-pore structure in controlling the adsorption processes. On the other hand, this finding may also suggest that more information about the adsorbents beyond the pore volume and BET surface areas should be cared in future studies.



Figure 3. Model performance on the reserved test datasets after adding different number (ranging from 35 to 80) of data points from the reported BET curve, in the legend, “Cuve\_Vt” means using the data points from the BET curve and calculated total volume as the features for describing the properties of GAC, “Curve\_BET” represents models that used the data points from the BET curve and calculated surface area as the feature for GAC; “Curve\_BET\_Vt” indicates that the built models used data points from the BET curve and BET surface area as well as the total pore volume as the feature for GAC, similarly, “Curve\_BET\_Vs” indicates that the built models used data points from the BET curve, surface area, pore volumes (including volume of total, micro-, and meso- pores).



Figure 3. The impact of training dataset size

To understand how the training sample size affecting the model’s performance on the test dataset, we fixed the test dataset and gradually increased the training sample size and monitoring the model’s performance the reserved test datasets. Two metrics were used to reflect effects of different training sample sizes; one is the prediction error on the test sets while the second one the reconstruction error for the test set input. The difference between this two metrics is that the first one can only be obtained after training models while the second can be calculate before the model being trained. With the increasing of training sample, both the MSE and the associated SD decreased quickly in the beginning and gradually reach plateau in the case of CO2 as well as of methane.



Figure 4. The impact of adding other datasets to the training sets of CFCs and E&E

The model based one the combined big data set showed negligible improvement for each individual dataset, however, the consistent performance of the model based on the combined dataset and the model based on individual data set suggested the promising capability of the employed Abraham descriptors for the representing various gases molecules. A more prominent results can be found in the E&E and the CFC data sets, each data set included more than more gas molecules such as ethane, ethylene, propane etc. Without using the Abraham descriptors, all these two dataset will be divided into several much smaller datasets, which will make the machine model training impossible. Given the Abraham descriptors, small and segmented data set can be combined into large dataset for model building. More importantly, our recent study has built a machine learning model for predicting Abraham descriptor with new chemical, which provide a promising solute for future modeling following the feature set used in the current study for new gas molecules.

Instead of using all the dataset to build a combined large dataset, we further tried to combined different dataset together to study whether some improvement can be achieved.