DEEP LEARNING OF INFRARED SPECTRA FOR BOILING POINT PREDICTION

Ching Yuhui Natalie[1], Alvin Liew[2], Chieu Hai Leong[2]
[1]River Valley High School, 6 Boon Lay Ave, Singapore 649961
[2]DSO National Laboratories, 12 Science Park Drive, Singapore 118225

Introduction

Boiling point estimation is of great importance in finding out a chemical's vapor pressure and hence toxicity, for modelling the distribution and fate of chemicals in the environment [1], and for the basis of design and simulation of chemical, biochemical and environmental systems [2]. Unfortunately, experimental boiling point data cannot always be found in literature. Since measurement is expensive and time consuming, estimation methods are generally of great value. However, there are three main issues with current boiling point estimation: 1. Current boiling point estimation requires knowledge of the structure of a compound. 2. It requires a pure sample of the compound, which is not always available. 3. To derive the structure of a compound, specialized apparatus are required to analyze the compound, and this can take up to half an hour. In situations where time is of utmost importance, such methods of boiling point estimation prove unfeasible. Computational techniques offer a realistic and advantageous alternative. Existing work has carried out research on predicting boiling points from molecular structures, but none directly from spectra, which can be taken quickly and effectively. Hence, a model that can predict boiling points from infrared (IR) spectra is able to mitigate the three issues listed above.

Related Work

Current models for the prediction of boiling points utilize chemical information such as molecular descriptors, which require knowledge of the structure of a compound. An example is the OPERA (OPEn (quantitative) structure-activity Relationship Application) model for boiling point proposed by Mansouri *et al*. [3]. It provides a suite of Quantitative Structure Activity Relationship (QSAR) models to predict physicochemical properties and environmental fate of organic chemicals based on molecular descriptors generated using the PaDEL software [4]. This method has achieved a remarkable accuracy of 22.08 RMSE on the test dataset, but knowledge on the exact structure of the compound is required, and it is not always available.

Hence, in this work, we focus on the construction of a model which predicts boiling points given only the IR spectra data from molecules and compare it to a baseline model which takes in actual structures, from which molecular descriptors are generated as input to the model.

Methodology

Data Preparation

The first dataset consists of molecular descriptors of 4077 chemicals as well as their boiling points obtained from the US Environmental Protection Agency. All null values in descriptors were replaced with the median value of that descriptor across all the chemicals. The molecular descriptors were then transformed such that each descriptor had a normal distribution with a mean of 0 and standard deviation of 1 across all molecules according to equation 1.

$$Z = \frac{x-u}{s} \quad (1)$$

Where:

- z is the standardized value.
- x is the original value.
- u is the mean of the feature.
- s is the standard deviation of the feature.

The second dataset contains the CAS, SMILES, name and boiling point of 11762 chemicals. The CAS, SMILES and boiling point were extracted from the dataset using regular expressions.

The third dataset consists of IR spectra data from the NIST chemistry webbook. There are 4211 files of IR spectra data, 398 of which are of solids, 1195 of which are liquids, and 2618 of which are of gases. In this model, we focus on predicting the boiling points using IR spectra of gaseous chemicals only, as the IR spectra of gaseous chemicals have the most well-defined peaks, which makes it easier for the model to predict, and there are far more gaseous chemical IR data than the other two states.

Since the boiling points are not specified in many of the files for IR spectra, chemicals present in both the third dataset with IR spectra and the second dataset with boiling points were extracted. There were 112 chemicals with more than one IR spectra file. Files that did not have any

contamination were preferentially chosen, and if either all the files had contamination or all the files had no contamination, files with a resolution of 2 were preferred, as most files for non-repeated chemicals used a resolution of 2, and a smaller resolution of 2 contains more detailed information as compared to other higher resolutions of 4 and 6. The total number of chemicals in the end was 1961.

In IR spectra data, the X values represent the wavelength or wavenumber at which the measurement is taken, and the Y values represent the intensity of frequencies absorbed or transmitted. In the dataset, the X values were either expressed as wavelengths in mm or expressed as wavenumbers in cm^{-1}while the Y values were in either absorbance or transmittance. The X factor and Y factors are scaling factors that are applied to the X and Y values respectively. The data was scaled such that the data is no longer scaled by the X and Y factors, and X units and Y units were standardized as cm^{-1} and absorbance respectively.

Y values in terms of transmittance were converted to absorbance according to equation 2.

$$A = log_{10}\left(\frac{1}{T}\right) \tag{2}$$

Where:

- *A is* the absorbance
- *T* is the transmittance

X values in wavelength (mm) were converted to wavenumbers (cm^{-1}) according to equation 3.

$$\bar{\upsilon} = \frac{1}{\lambda} \tag{3}$$

Where:

- $\bar{\upsilon}$ is the wavenumber in cm^{-1}
- λ is the wavelength in mm

The y values were then interpolated for every x in the range 700 – 3570, with an interval of 2cm-1, using cubic spline interpolation.

Suppose that distinct nodes $t0<t1<\cdots<tn$) and data $y0,\ldots,yn$ are given. . For any $k=1,\ldots,n$, the spline $S(x)$ on the interval $[tk-1,tk]$ is by definition a cubic polynomial $Sk(x)$, shown in equation 4.

$$Sk(x)=ak+bk(x-tk-1)+ck(x-tk-1)2+dk(x-tk-1)3,k=1,\ldots,n \quad (4)$$

Where:

- ak,bk,ck,dk are values to be determined, overall there are $4n$ such undetermined coefficients

Box cox transformation (refer to Appendix A) and the transformation specified in equation 1 were then applied to the data such that it was less skewed with a normal distribution and had a standard deviation of 1.

Opera model (k-nearest neighbours)

The OPERA model utilizes distance weighted k-nearest neighbors (kNN), which is a refinement of the classical k-NN classification algorithm where the contribution of each of the k neighbours is weighted according to their distance to the query point, giving greater weight to closer neighbours, and Euclidean distance.

Random Forest

The baseline model is constructed using an ensemble method known as random forest, that combines the output of multiple decision trees to reach a single result. Each decision tree is trained on a different subset of the data, and the predictions of all the trees are averaged to produce the final prediction.

I used sklearn's random forest regressor to predict the boiling points from molecular descriptors. First the baseline model takes in the structure of the chemical, which is in the form of SMILES. RDkit is then used to generate the molecular descriptors, which serve as inputs for the model, which then generates boiling points. Another random forest model is trained on IR spectra and is used to generate boiling points, to serve as a comparison with the baseline model, and show the difference molecular descriptors and IR spectra have on model accuracy.

XGBoost

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. The XG boosting algorithm creates a sequential ensemble of tree models, all of which work to improve each other and determine the final output. Random search was used for hyperparameter tuning. (Refer to Appendix B for results)

Feed Forward Neural Network (FFNN)

A simple FFNN was created to generate the boiling points based on IR spectra data. We postulate that FFNNs will provide better performance compared to the random forest and XGboost models. Decision trees and tree ensembles perform well on structured data, while neural networks are better able to handle unstructured data, such as images. Neural networks may also be able to better capture links between different peaks and the substructures within a molecule due to their ability to capture non-linear relationships with non linear activation functions. The FFNN model architecture used is shown in figure 2.

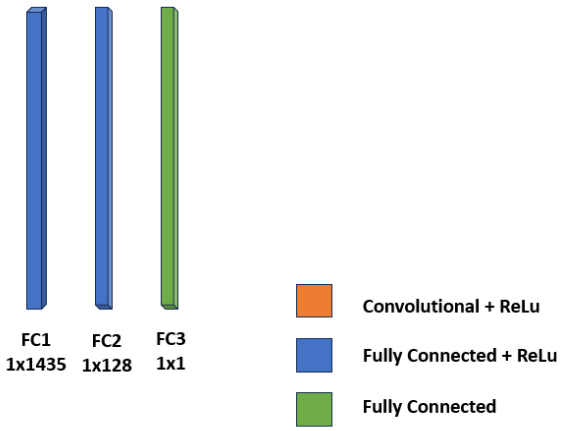

Figure 2: FFNN architecture

Convolutional Neural Network (CNN)

CNNs are widely used for image data, as a filter can be convolved over the input, effectively extracting features from images and learning to recognize patterns. This makes them well-suited for tasks such as object detection, image segmentation, and classification. In this work, we propose CNNs to predict boiling points as differently sized filters can be convolved over the spectral input and can identify differently sized peaks in the data. For example, larger filters can

identify broader peaks in the data, while smaller filters identify narrower peaks in the data. This can help the model to better understand spectral data and better identify links between different peaks and the substructures within a molecule, which directly affect boiling point. The basic model architecture used for the CNN model is shown in Figure 4.

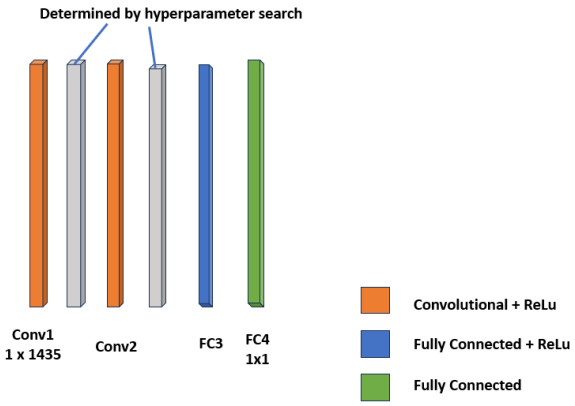

Figure 4: CNN model architecture

Layers such as pooling and dropout, along with model parameters such as output channels of layers, were decided through hyperparameter search using 2 rounds of random search. Based on the top 3 results from the first random parameter search, a secondary range of values for the next random search was defined. (Refer to Appendix C for the best set of hyperparameters)

Results and Discussion

Model	OPERA (molecular descriptors)	Baseline model (molecular descriptors)	Random Forest (IR spectra)	XGBoost (IR spectra)	FFNN (IR spectra)	CNN (IR spectra)
Test RMSE	22.08	20.79	58.99	56.02	47.60	42.41

Through this project, we hope to answer 2 questions:

1. **Is it feasible to explore the use of IR spectra for prediction of boiling points?**

   As seen from the results table, the CNN model has a higher RMSE compared to both models using molecular descriptors instead of IR spectra. The discrepancy in the accuracy of the models could be due to a combination of reasons. IR spectra data is more prone to noise due to experimental conditions and instrument variations, as compared to molecular

descriptors which are explicit and defined for every molecule. For example, IR spectra data is prone to contamination, and some files have resolutions higher than 2, which negatively impact the interpolation accuracy. Generally, molecular structure is better for predicting boiling point, as boiling points are dependent on types of functional groups within a molecule.

Although it has a lower accuracy, the CNN's ability to utilize IR spectra data that can promptly be acquired enables its use in cases with lack of specialized laboratory equipment, access to pure compounds and lack of information. With further research and experimentation, the model's results can be further improved, and is a feasible option to explore to mitigate the current challenges associated with boiling point estimation.

2. **Do CNNs outperform FFNNs in handling spectral input?**

In the context of spectral analysis for predicting chemical properties, this study shows that CNNs outperform FFNNs in handling spectral input, due to the CNN model's ability to leverage differently sized filters to learn patterns during the convolution process, a feature absent in FFNNs. Our research shows that beyond image recognition and classification, CNNs have potential to be used for spectral data with good results, and can be explored for use in models with spectral inputs.

Future work

Contrastive learning can be used to help encode implicit molecular information into the model's weights to increase accuracy. Applying contrastive loss between a spectral encoding from our CNN with a molecular encoding from a GNN which predicts the boiling point of chemicals from its molecular structure can achieve this. Contrastive loss between positive pairs, or molecular and spectral embeddings of the same chemical, is minimized, and vice versa, causing embeddings of the same chemicals to be as similar as possible, and vice versa. By learning these patterns, the CNN's weights will be updated to include this molecular information, which will help increase its accuracy in the prediction of boiling points.

The model can be improved to accommodate spectral data from various states, instead of just gas. Variation of spectral characteristics across different states requires enhancing the model's robustness to these variations. Since the dataset has significantly fewer spectra on solids and liquids, more data and data augmentation can enable the model to learn from an equal

representation of data. More layers can also be used to capture state specific patterns that may be unique to different states.

Citations

1. Dearden JC. Quantitative structure-property relationships for prediction of boiling point, vapor pressure, and melting point. Environ Toxicol Chem. 2003 Aug;22(8):1696-709. doi: 10.1897/01-363. PMID: 12924571.
2. Yash Nannoolal, Jürgen Rarey, Deresh Ramjugernath, Wilfried Cordes. Estimation of pure component properties: Part 1. Estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions, Fluid Phase Equilibria,Volume 226. 2004 https://doi.org/10.1016/j.fluid.2004.09.001.
3. Mansouri, K., Grulke, C.M., Judson, R.S. *et al*. OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminform 10, 10 (2018). https://doi.org/10.1186/s13321-018-0263-1
4. Yap, C.W. (2011-05). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry 32 (7) : 1466-1474. ScholarBank@NUS Repository. https://doi.org/10.1002/jcc.21707
5. Quiza, Ramon & Davim, J. Paulo. (2011). Computational Methods and Optimization. 10.1007/978-1-84996-450-0
6. Phung, & Rhee,. (2019). A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets. Applied Sciences. 9. 4500. 10.3390/app9214500.

Appendix

Appendix A: Box Cox Transformation

The Box Cox Transformation is a popular method of transforming non-normal dependent variables into a normal shape. This technique helps to stabilize variance and can improve the accuracy of any subsequent statistical tests or models.

Before applying Box Cox. other methods of data transformation, namely the log transformation, square root transformation, cube root transformation were tested on the left skewed data, but box cox was found to reduce the skewness of the data the most, and brought it closest to a normal distribution.

The formula of the box-cox transformation is shown in equation 5.

$$X^{(\lambda)} = \begin{cases} \frac{X^{\lambda}-1}{\lambda} for \lambda \neq 0 \\ logX for \lambda = 0. \end{cases} \quad (5)$$

Where:

- X is a random variable on the positive half line
- λ is a parameter chosen, in this case the value that maximizes the log-likelihood function

Appendix B: XGBoost hyperparameters

Subsample	0.9
Min child weight	0.1
Max depth	6
Gamma	4
Eta	0.1
Colsample by tree	0.2

Appendix C: CNN hyperparameters

Hyperparameter name	Value
Learning Rate	0.001
Out channels (first convolutional layer)	48
Kernel size (first convolutional layer)	3
Stride (first convolutional layer)	1
Pooling (after first convolutional layer)	4
Batch norm (after first convolutional layer)	Present
Padding (first convolutional layer)	3
Dropout (after first convolutional layer)	0.2
Kernel size (second convolutional layer)	27
Stride (second convolutional layer)	2
Pooling (after second convolutional layer)	1
Batch norm (after second convolutional layer)	None
Padding (second convolutional layer)	1
Dropout (after second convolutional layer)	0.0
Out Features (first linear layer)	672

The Out Channels of the second convolutional layer and the out features of the second convolutional layer are both 1 for a multivariate regression problem.