Efficient Sensing of PAH Concentrations Using Fluorescence Imaging

Gaurav Mahamuni, Anamol Pundle
Mechanical Engineering Department
University of Washington
Seattle, WA 98195
gauravsm@uw.edu, apundle@uw.edu

Abstract

Polycylic Aromatic Hydrocarbons (PAHs) are a widespread class of environmental chemical pollutant known to have carcinogenic and mutagenic effects in humans and other living beings. In this work, we develop two machine learning models to predict concentrations of 16 PAHs in soot samples obtained from combustion of different fuels at different conditions. We use 1. PCA along with Multiple Linear Regression (MLR) and 2. Neural Networks to develop prediction models using training data. We use test data to calculate percentage error for each PAH. The average test error using PCA - MLR is 34% and for Neural Networks is 84%.

1 Introduction

Combustion generated particulate matter (PM), or soot, is a major pollutant, and is responsible for millions of deaths and illnesses every year [1]. It also has an adverse effect on the climate, and is thought to be the second most important contributing species to climate change after CO_2 [2]. PM is generated in several commonly encountered situations, such as cooking using biomass, smoking, and vehicular and industrial exhaust. PAHs adsorbed on the soot surface are often carcinogenic, and pose a significant health hazard to individuals exposed to $PM_{2.5}$ [3].

The presence and concentration (in $^{\mu g}/_{m^3}$) of these PAHs is usually determined using Gas Chromatography/Mass Spectroscopy (GC/MS), an expensive and cumbersome technique [4]. Novosselov lab is currently investigating an alternative technique for the efficient sensing of these compounds: fluorescence imaging, or Excitation Emission Matrix(EEM). As per this technique, the sample is inundated with light of different wavelengths, causing the sample to fluoresce. Each compound has a unique fluorescence trace for every incident wavelength of light. The output of this method is a spectrogram-like image of Emitted vs Excitation wavelength for each sample. Our hypothesis is that using machine learning, this data can be used to quantify the concentration of sixteen PAH's commonly observed in soot, out of which nine are carcinogenic. If successful, this technique would be much less time consuming and relatively inexpensive compared to GC/MS. Fig. 1 shows example an EEM image and its corresponding PAH concentrations determined from GC/MS.

2 Methods

2.1 Creating Images

Our dataset consists of 20 EEM image matrices of size 1000×201 and their corresponding PAH concentrations. We reduce the size of these 20 images to 100×33 using $2D$ MaxPool with filter size $(10, 6)$. We partition the data into a training and test set: 15 images for training and 5 for testing, and their corresponding known PAH concentrations. In order to generate more data to train our

CSE: 546 Machine Learning (2018), University of Washington, Seattle.

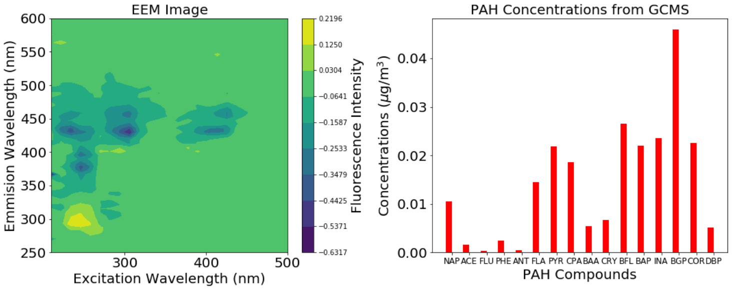

Figure 1: Example image (left) and corresponding PAH concentrations (right)

machine learning algorithms, we compute linear combinations of the existing images and their known concentrations. 585 images are generated from the training set and 145 from the test set. Each generated image is a linear combination of three randomly chosen images from its set, each image matrix multiplied by a random number between $^2/_{15}$ and $^8/_{15}$ ($^1/_3 \pm {}^1/_5$). This is equivalent to running an experiment with the physical samples mixed in the same proportion, since fluorescence in PAHs has been shown to be proportional to their concentrations based on (1),

$$F = KC \tag{1}$$

where F is image intensity and C is PAH concentration and K is a constant specific to each PAH [5]. Remaining 5 images and corresponding labels are used to create 150 test images and corresponding labels.

2.2 PCA and Multiple Linear Regression

Each EEM image in the training set is unfolded as a row array and stacked to create an nXp matrix where n is number of training samples (n=600) and p is number of pixels (p=3300). The resultant data matrix is demeaned columnwise. The 'PCA' function from class 'sklearn.decomposition' is used to decompose the data matrix. It uses the LAPACK implementation of the full SVD or a randomized truncated SVD [7], depending on the shape of the input data and the number of components to extract. We plot the percentage variance explained by each component vs number of components in Fig. 2. The scores are representative of each image after PCA.The dimension of the scores for each image is 600 which is reduced to d based on Fig.2. The reduced data matrix nXd represents the training data.

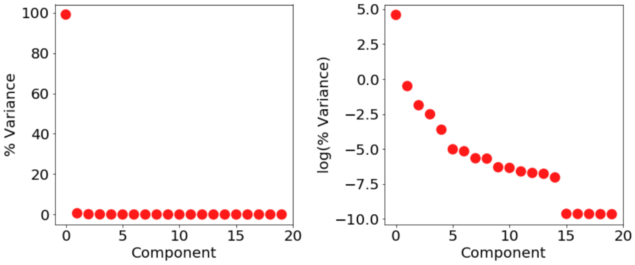

Figure 2: % Variance for each component (left). Since the first component accounts for most of the variance, a log plot gives more information on relative contribution of other components to the variance (right)

The 'linear_model' function from class 'sklearn' is used for multiple linear regression analysis. 'LinearRegression' fits a linear model with coefficients to minimize the residual sum of squares between the observed responses in the dataset. We use 'LinearRegression' to compute a linear model between reduced data matrix of size nXd and training labels of size $nX16$. The training error is calculated using the linear model.

Each EEM image in the test set is unfolded and demeaned the same way as images in training set to create a matrix of size mXp, where m ($m = 150$) is number of test samples.The function

’pca.transform()’ is used to obtain scores of test samples based on dimensionality reduction of training samples. First d scores out of 150 scores are chosen. Test PAH concentrations are predicted based on the d scores for each test sample using the linear model obtained. The test error is also calculated.

2.3 Neural Networks

The python library PyTorch is used to build and train the neural network model. For training the neural network, we further partition the 600 training images into 480 images for training and 120 images for validation. The neural network consists of one fully connected hidden layer and one fully connected output layer. We parcel the training data into batches of five images each. Therefore, the input to the neural network is a tensor of size $5 \times 1 \times 100 \times 33$. We choose the output of the hidden layer to be a vector of size 500, and the final output is a vector of size 16. We train the network via mean squared error loss.

The hyperparameters optimized for are the momentum and learning rate. We randomly choose 1000 values of momentum between 0 and 1, and 1000 values of learning rate between 1×10^{-6} and 1×10^{-1}. We train the neural network for each pair of hyperparameters for 5 epochs. Subsequently, the 20 pairs with the least mean squared error on the validation set are then trained further on 25 epochs of the data. Finally, the pair of hyperparameters with the lowest mean squared error on the validation set is chosen as the accepted set of hyperparameters. The model is then tested on the test set.

2.4 Error Calculation

We define the percentage error in the prediction of each PAH concentration as,

$$\%error = 100X \frac{\sum_{i=1}^{n} |y_{pred} - y|}{\sum_{i=1}^{n} |y|} \tag{2}$$

where n is the total number of samples in the test set, y_{pred} is the predicted concentration of the PAH and y is the actual concentration of the PAH.

3 Results and Discussion

3.1 PCA and Multiple Linear Regression

Fig.2 (right) shows that % variance buckles after 10 components, hence we choose $d = 10$. We reduce dimension of images from 3300 to 14 using PCA. Linear model obtained for scores and labels of training data is used to predict test data. %error for test data is shown in Fig 3

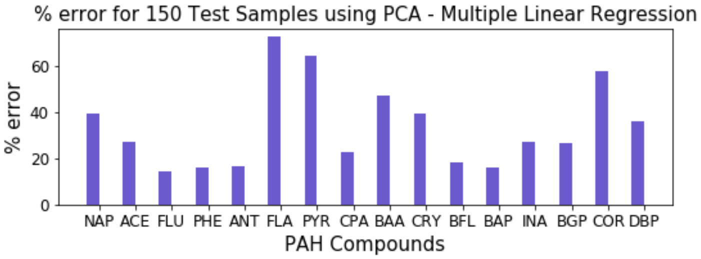

Figure 3: $\%error$ for all 16 PAH compounds. The average $\%error$ is 34%

The $\%error$ is low for FLU, PHE, ANT, BFL, BAP, high for FLA, PYR, COR, BAA, CRY and in between for NAP, ACE, CPA, INA, BGP and DBP. Fig. 4 shows that PAHs which show low $\%error$ are predicted well in random samples.

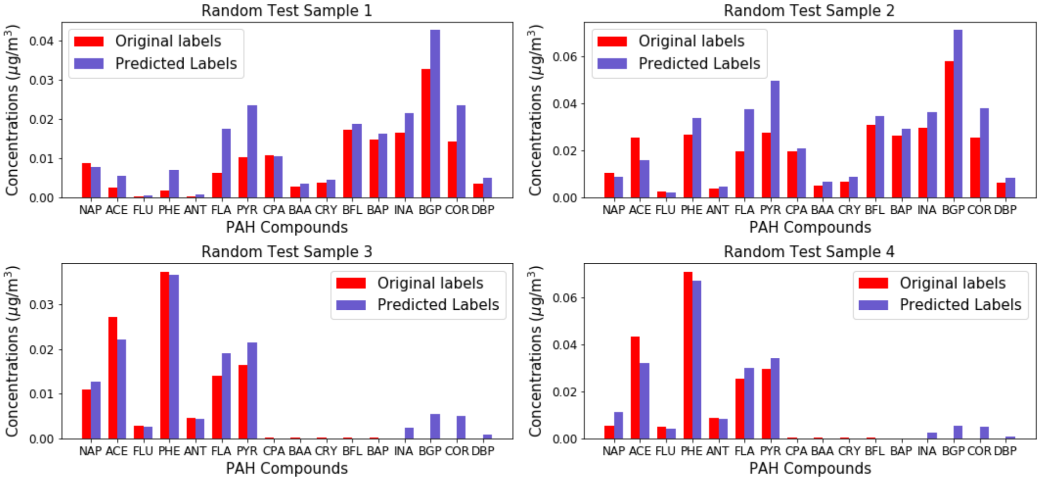

Figure 4: Actual and predicted PAH concentrations for 4 random samples from test set. PAH concentrations are somewhat overpredicted using PCA-MLR but gives an overall idea of which compounds are present

3.2 Neural Network

Fig.5 shows the $\%error$ of each PAH for the neural network on the test set. We observe that the model does not perform well; the maximum $\%error$ is 142.5% for PHE, while the minimum $\%error$ is 51.2% for NAP.

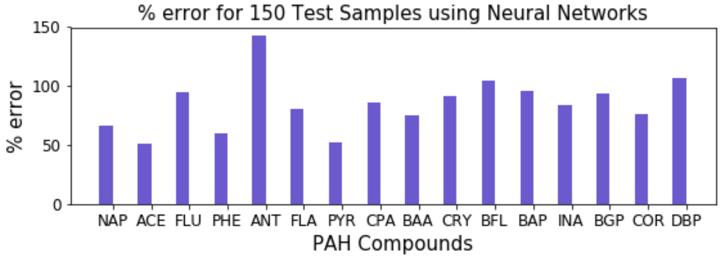

Figure 5: $\%error$ for all 16 PAH compounds. The average $\%error$ is 84%

4 Conclusions

In this work we have developed two machine learning models for predicting the concentrations of PAHs adsorbed on the surface of soot particles from EEM Images. The first model uses PCA to reduce the dimensions of images and correlates scores along principal components for each sample to its PAH concentrations using linear regression. The average test $\%error$ for PCA-MLR method is 34%. The second model uses a neural network with one fully connected hidden layer and a fully connected output. The average test $\%error$ for second model is 84%.

The PCA-MLR method predicts PAH concentrations with reasonable accuracy. This makes sense because the relationship between fluorescence and PAH concentration for any PAH is linear. However the fluorescence of PAH compounds overlap. This makes predicting PAH concentration from EEM images difficult, especially when number of PAHs in the sample increases. The 20 gold standard samples might have more than 16 PAHs. These 16 PAHs are dominant fluorophores for most of the samples in the gold standard samples, hence PCA-MLR gives better prediction for concentration of PAHs. For more complex samples, PCA-MLR might not perform well and we will need a well tuned convolutional neural network to predict PAH concentrations.

5 References

[1] Kampa, M., and Castanas, E., 2008, “Human Health Effects of Air Pollution,” Environ. Pollut., 151(2), pp. 362–367.

[2] Ramanathan, V., and Carmichael, G., 2008, “Global and Regional Climate Changes due to Black Carbon,” Nat. Geosci., 1(4), pp. 221–227.

[3] de Kok, T. M. C. M., Driece, H. A. L., Hogervorst, J. G. F., and Briedé, J. J., 2006, “Toxicological Assessment of Ambient and Traffic-Related Particulate Matter: A Review of Recent Studies,” Mutat. Res. Mutat. Res., 613(2–3), pp. 103–122.

[4] Mannino, Maria Rosaria, and Santino Orecchio. "Polycyclic aromatic hydrocarbons (PAHs) in indoor dust matter of Palermo (Italy) area: extraction, GC–MS analysis, distribution and sources." Atmospheric Environment 42.8 (2008): 1801-1817.

[5] Sun, Renhui, et al. "Analysis of gas-phase polycyclic aromatic hydrocarbon mixtures by laser-induced fluorescence." Optics and Lasers in Engineering 48.12 (2010): 1231-1237.

[6] Zepp, Richard G., Wade M. Sheldon, and Mary Ann Moran. "Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating Rayleigh and Raman scattering peaks in excitation–emission matrices." Marine chemistry 89.1-4 (2004): 15-36.

[7] Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp. "Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions." (2009).

Appendix A: PAH Data

Name	Molecular Weight	Chemical Formula
Naphthalene	128.17	$\mathrm{C}_{10}H_8$
Acenapthylene	152.2	$\mathrm{C}_{12}H_8$
Fluorene	166.23	$\mathrm{C}_{13}H_{10}$
Phenanthrene	178.23	$\mathrm{C}_{14}H_{10}$
Anthracene	178.23	$\mathrm{C}_{14}H_{10}$
Fluoranthene	202.26	$\mathrm{C}_{16}H_{10}$
Pyrene	202.26	$\mathrm{C}_{16}H_{10}$
Cyclopenta(cd)pyrene	226.27	$\mathrm{C}_{18}H_{10}$
Benzo(a)Anthracene	228.28	$\mathrm{C}_{18}H_{12}$
Chrysene	228.28	$\mathrm{C}_{18}H_{12}$
Benzofluoranthenes	253.31	$\mathrm{C}_{20}H_{12}$
Benzo(a)Pyrene	252.31	$\mathrm{C}_{20}H_{12}$
Indeno(123cd)Pyrene	276.39	$\mathrm{C}_{22}H_{12}$
Benzo(ghi)Perylene	276.39	$\mathrm{C}_{22}H_{12}$
Coronene	300.36	$\mathrm{C}_{24}H_{12}$
Dibenzopyrenes	302.38	$\mathrm{C}_{24}H_{14}$