Machine Learning for Metal Oxide Gas Sensor Analysis

by Nathan Cao, Bigyan Kandel, and Oomman K. Varghese, Ph.D.
Department of Physics

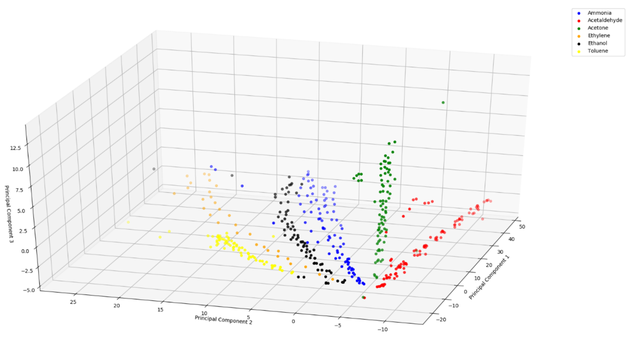
UNIVERSITY of HOUSTON

Background

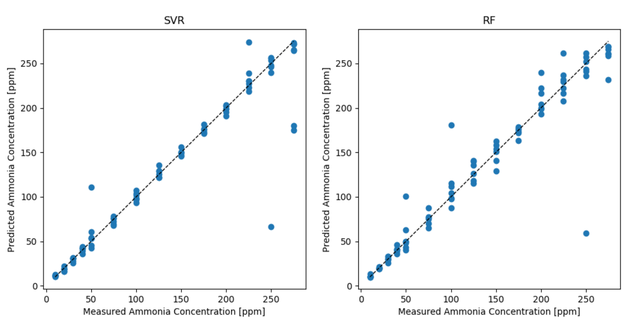
Metal oxide semiconductors have picked up interest in the science community due to their application as a sensor. Since metal oxide semiconductors are cheaper to produce and easier to carry compared to gas chromatography-mass spectrometry, many researchers have focused on expanding their ability to distinguish VOCs. It has been shown that an array of metal oxide sensors can be used to test a patient's breath to detect if they have cancer. However, metal oxide sensors are still unable to distinguish different cancers from one another, such as lung cancer and breast cancer [1,2,3]. A proposed method is to quantify the concentrations of the gases in a patient's breath using a metal oxide sensor array, since it has been shown that each cancer has unique biomarkers. Pattern recognition such as machine learning can prove to be effective in converting a raw signal to exact gas concentration [4]. We investigate the effectiveness of various machine learning models such as principal component analysis (PCA), support vector regression (SVR), and random forest (RF).

Results

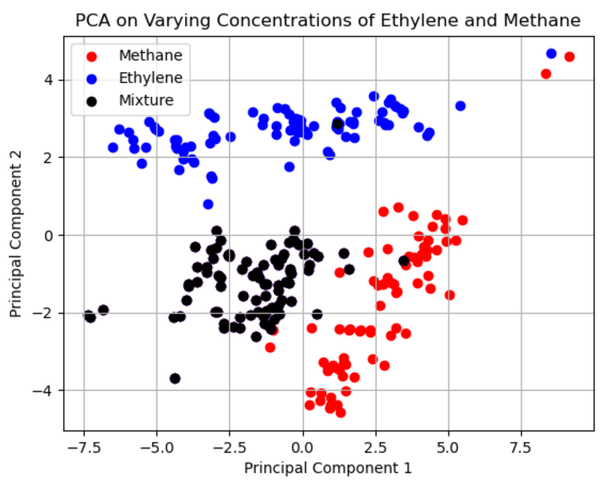
For the first dataset of pure gas mixtures the PCA was able to differentiate the gas by type.



The SVR and RF models were successful if limited in scope on predicting only concentrations of ammonia gas. Using k-fold cross validation (k=5), SVR resulted in a mean root mean squared error (RMSE) of 22.591841ppm and standard deviation of 12.754246ppm while RF resulted in a mean RMSE of 18.993506ppm and standard deviation of 14.255348ppm



When attempting to predict concentrations of both ammonia and toluene, the models would predict that pure gas mixtures of ammonia contained some toluene and vice versa. The overall accuracy however was high with a mean RMSE of 14.413ppm and a standard deviation of 15.472ppm for SVR and a mean RMSE of 10.118ppm and standard deviation of 7.548ppm for RF. Another issue that came up during both models was predicting values outside of training data. When trained on only low concentration sensitivities, both models were unable to predict the concentrations of higher concentration samples.
However, to further expand on the success of PCA, I applied it to another dataset with mixed gases as mentioned before. The model was able to differentiate between mixed and pure quite well for the mixture of methane and ethylene.



Conclusions

Principal component analysis is a valid model to use when analyzing the efficacy of sensor data and in identifying the qualitative identity of both pure and mixed gases. The pattern recognition portion of the project in identifying the quantitative amount of gases fell short, however more work can be done by expanding the algorithms used to a machine learning model that could effectively extrapolate data better. Perhaps a simpler regression model could be used (linear regression) or artificial neural networks could be explored to support a better quantitative result. The issue could also be solved by generating more data, and a more time conservative method of producing the data could be to construct a computer-generated model to predict sensor readings.

Methodology

- Due to COVID-19 restrictions, two datasets were taken from the University of California Irvine, one pure gas and one mixed gas
- Machine learning algorithms used were from the Python library Sci-Kit Learn alongside Numpy and Pandas.
- The dataset consisting of pure gases provided sensitivities (R_{gas}/R_{air}) of 16 metal oxide sensors exposed to a gas. Only the first and second batches were used to avoid issues in sensor recovery causing different values.
- PCA, SVR, and RF were applied to the first dataset. PCA was conducted to visualize the qualitative ability to identify the gas identity. SVR and RF were applied only to ammonia samples to determine the quantitative ability to identify the gas concentration. SVR and RF were then applied to predict ammonia and toluene concentrations.
- A grid search was conducted to determine hyper parameters of each algorithm for minimal root mean squared error.
- Effectiveness of each algorithm was determined using a k-fold cross validation (k=5).
- The dataset consisting of mixed gases consisted of an ethylene and methane dataset and an ethylene and carbon monoxide dataset. Concentrations of each were varied randomly for a 12-hour time series.
- Feature extraction was conducted to take both the sensitivity and resistance of each concentration variation.
- PCA was applied to determine if pure gas can be distinguished from gas mixture.

References

1. Jia, Z., Patra, A., Kutty, V. K., & Venkatesan, T. (2019). Critical review of volatile organic compound analysis in breath and in vitro cell culture for detection of lung cancer. *Metabolites*, *9*(3), 52.
2. Nasiri, N., & Clarke, C. (2019). Nanostructured chemiresistive gas sensors for medical applications. *Sensors*, *19*(3), 462.
3. Katwal, G., Paulose, M., Rusakova, I. A., Martinez, J. E., & Varghese, O. K. (2016). Rapid growth of zinc oxide nanotube–nanowire hybrid architectures and their use in breast cancer-related volatile organics detection. *Nano letters*, *16*(5), 3014-3021.
4. Xu, Y., Zhao, X., Chen, Y., & Zhao, W. (2018). Research on a mixed gas recognition and concentration detection algorithm based on a metal oxide semiconductor olfactory system sensor array. *Sensors*, *18*(10), 3264.