CMPE 255 - Data Mining

Instructor: Dr. Gheorgi Ghuzun
Department of Computer Engineering

Analysis Of Gas Sensor Array Under Dynamic Gas Mixtures

Team 13

Vaibhav Dnyandeo Ingale (016131167)
Lalitha Ramya Vemuri (016697317)
Hardwin Bui (011007156)

1. Introduction

This report describes design, implementation, and evaluation of a data mining system for chemical sensor time series data. The system is designed to identify patterns in data that can be used to predict concentration of different gasses in air.

1.1. Motivation

Chemical sensors are used in a wide variety of applications, including environmental monitoring, industrial process control, and medical diagnostics. In order to use chemical sensors effectively, it is important to be able to predict the concentration of gas that they are sensing. This can be done by analyzing time series data that is generated by sensors.

1.2. Objective

The objective of this project is to develop a data mining system that can be used to identify patterns in chemical sensor time series data. The system should be able to classify and predict concentration of different gasses in air with high accuracy.

1.3. Literature/Market review

There has been a significant amount of research on data mining for chemical sensor time series data. One of the most common approaches is to use machine learning algorithms to identify patterns in data. Machine learning algorithms can be used to learn the relationship between sensor readings and concentration of gas that they are sensing. Another approach to data mining for chemical sensor time series data is to use statistical methods. Statistical methods can be used to identify trends and patterns in data. Trends and patterns can be used to predict concentration of gas that sensors are sensing.

2. System Design & Implementation details

2.1. Algorithms considered/selected

1. Logistic Regression

Logistic Regression is an extension of linear regression to predict qualitative response for an observation. It defines the probability of an observation belonging to a category or group. Logistic regression is used to predict absence or presence of a specific gas or gas mixture based on readings from sensors. In dynamic gas mixtures, composition of gas mixture may change over time, making it challenging to accurately classify gas using traditional methods.

Logistic regression is well-suited for this type of problem because it can handle non-linear relationships between independent variables and dependent variables, which is often the case in gas sensor array data sets. Additionally, logistic regression can provide a probabilistic interpretation of classification, which can be useful in determining classification confidence.

2. LDA and QDA

These algorithms are based on Bayes theorem and are different in their approach for classification from Logistic Regression. On two different sets of training and test data and prints accuracy score, classification report, and confusion matrix for the first set of data. Performing Linear Discriminant Analysis (LDA) on two different sets of training and test data will give us an accuracy score, classification report, and confusion matrix for the first set of data.

3. SVM

Used SVM because they are capable of handling non-linear relationships and can effectively classify complex data with high accuracy. Additionally, SVM can handle high-dimensional data sets and are not easily affected by noise or outliers.

4. K-fold

Cross-validation can help evaluate performance of a machine learning model in a more reliable way. In gas sensor array analysis, K-fold cross-validation can be used to train and test models on different subsets of data, which can help prevent overfitting and improve generalization of models to new data.

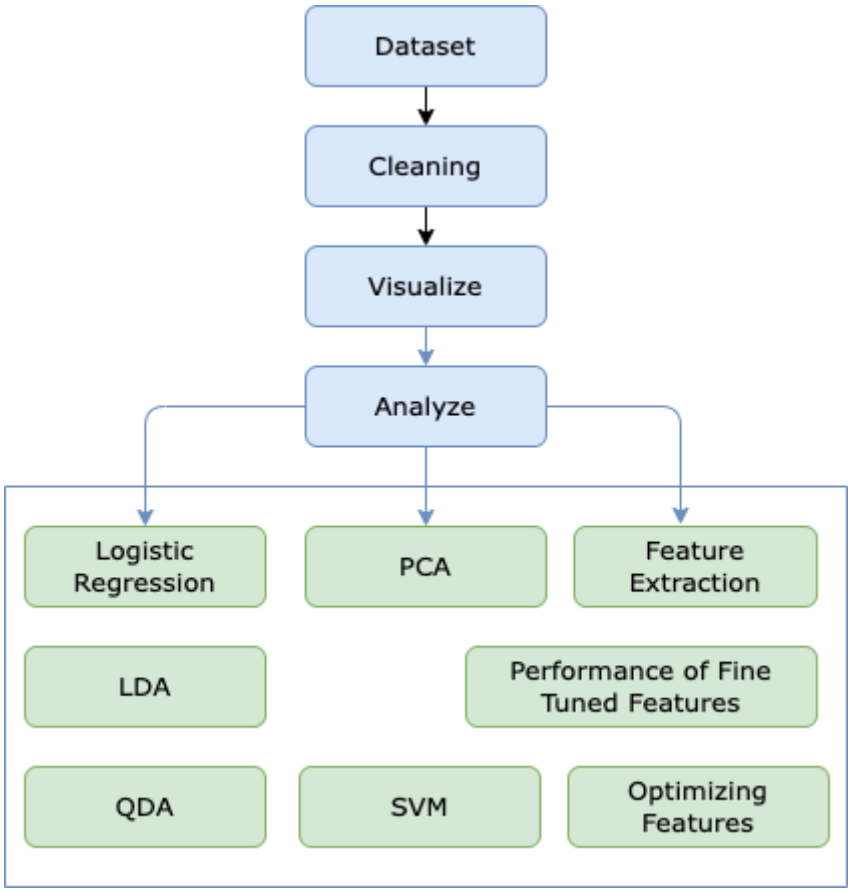
5. PCA

Can be used to reduce dimensionality of data and capture most important features that explain variability in gas sensor data. In gas sensor array analysis, PCA can help identify most significant sensor responses to different gasses, which can n be used to classify and identify unknown gas mixtures.

2.2. Technologies & Tools used

Python libraries - pandas, matplotlib, numpy, sklearn and statsmodels.
These tools help us as Pandas for data manipulation, matplotlib for visualization, numpy for numerical computing, sklearn for machine learning, and statsmodels for statistical modeling.
IDE - Jupyter notebook, Google colab, VS Code.

2.3. System design and architecture



2.4. Use cases

1. Industrial Safety: This system can be used to monitor levels of ethylene and methane gasses in industrial environments, such as chemical plants, oil refineries, and mines. This can help prevent accidents caused by gas leaks and ensure safety of workers.

2. Agriculture: This system can be used to monitor levels of ethylene gas in greenhouses, which is naturally produced by ripening fruits and vegetables. This can help farmers optimize the harvesting process and improve crop yields.

3. Environmental Monitoring: This system can be used to detect and monitor levels of methane gas emissions in landfills and wastewater treatment facilities. This can help identify and mitigate sources of greenhouse gas emissions.

4. Food Quality Control: This system can be used to detect presence of ethylene gas in food storage and transportation environments. This can help maintain the quality and freshness of perishable foods such as fruits and vegetables.

5. Medical Applications: This system can be used in medical applications to detect presence of certain gasses in a patient's breath, which can be used to diagnose various medical conditions such as asthma, lung cancer, and gastrointestinal disorders.

3. Experiments / Proof of concept evaluation

3.1. Dataset details:

In this project, we will be using the Gas Sensor Array Under Dynamic Gas Mixtures Data Set, which was pulled from UCI Machine Learning repository. This data provides Ethylene concentration level readings of 16 different sensors in 2 different chemical environments. This data set consists of 4178504 instances and 19 attributes.

Repo Link: https://archive.ics.uci.edu/ml/datasets/gas+sensor+array+under+dynamic+gas+mixtures

3.2. Methodology followed:

- Data Collection: The dataset used for this project is "Gas sensor array under dynamic gas mixtures Data Set" from UCI Machine Learning Repository. This dataset contains data collected from an array of 16 sensors that detect different types of gasses, including Ethylene and Methane.

- Data Cleaning and Preprocessing: The dataset is preprocessed to remove missing values and normalize sensor readings. The data is split into two separate datasets, one for Ethylene and one for Methane.

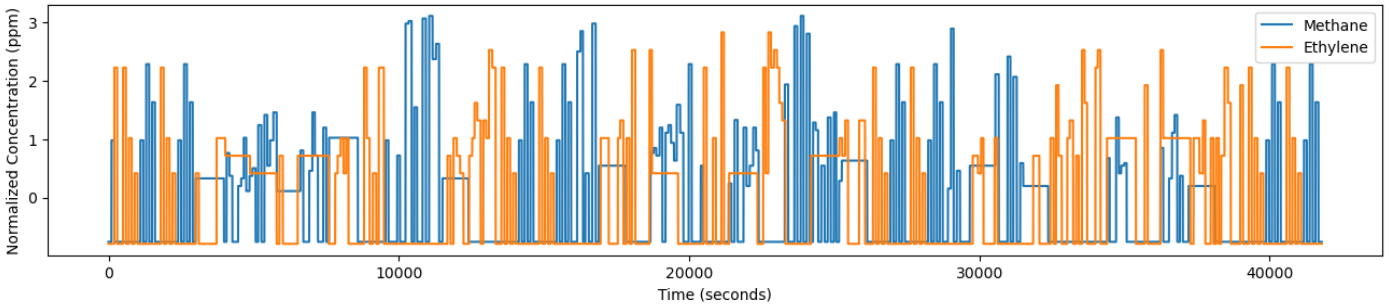
- Feature Selection: Principal Component Analysis (PCA) is used for feature selection to reduce dimensionality of data and select most important features that contribute to detection of Ethylene and Methane.

- Model Selection: Four different classification models were developed to predict presence of each gas type - Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR). For each gas type, these models were developed using cleaned and normalized input variables and were trained using a training dataset.

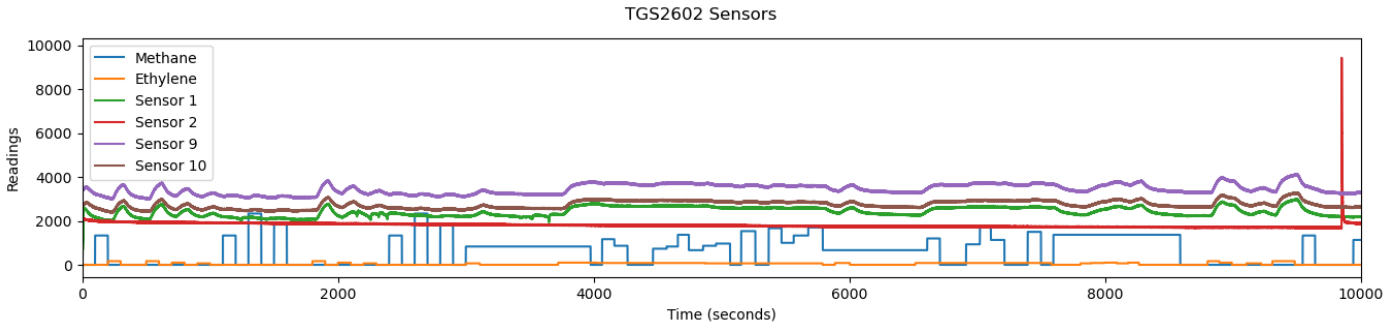
- Model Training and Testing: For LDA, QDA, and LR, data is split into half for training and half for testing, and accuracy is calculated using testing data. The accuracy is calculated using K-fold cross-validation with K=3.

- Model Evaluation: The accuracy of each model is evaluated using various metrics such as accuracy score, confusion matrix, precision, recall, and F1 score. To evaluate accuracy of developed models, a testing dataset was used. For LDA, QDA, and LR models, accuracy was obtained by comparing predicted labels to true labels. For RF models, accuracy was obtained using a K-fold cross-validation method with K=3.

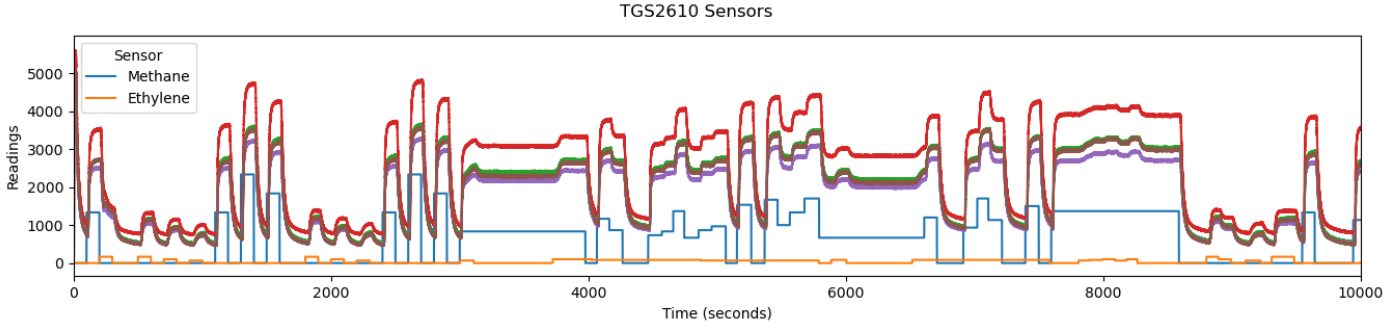
3.3. Graphs showing different parameters/algorithms evaluated in a comparative manner



This code is performing feature scaling on methane and ethylene concentration data. Feature scaling is a technique used to normalize a range of features or variables of data. The resulting scaled data is n plotted against time. This plot shows normalized concentrations of methane and ethylene over time. This plot is useful to compare trends and patterns in concentrations of both gasses over time, without being influenced by difference in range of concentrations of two gasses.

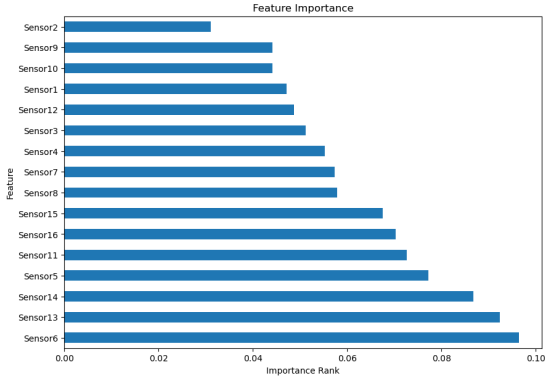


This plot shows readings from sensors 1, 2, 9, and 10 over time.The purpose of this is to compare data from multiple sensors and see if there are any patterns or correlations between different sensors.

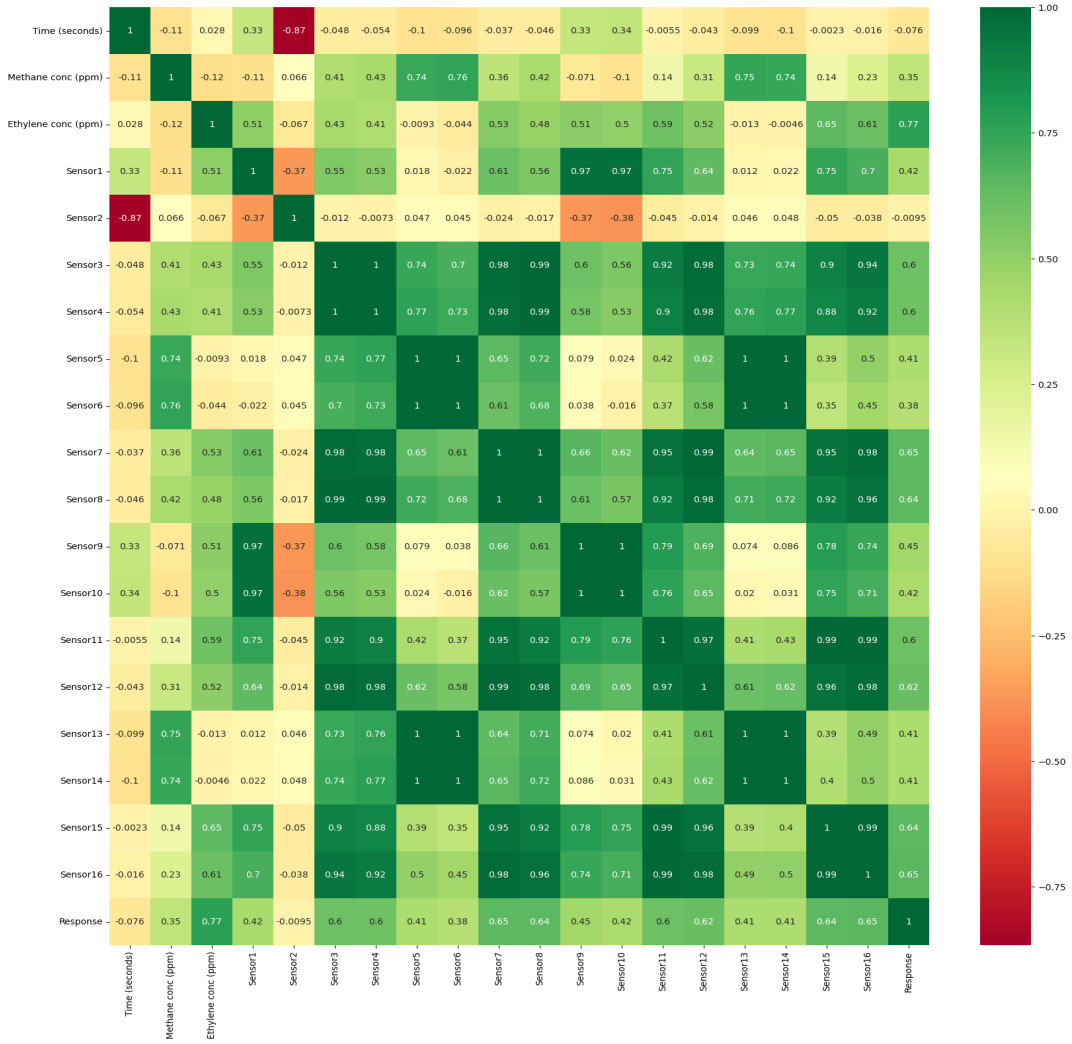


This plot shows readings from sensors 5,6, 13, and 14 over time.The purpose of this is to compare data from multiple sensors and see if there are any patterns or correlations between different sensors and the concentrations. Here we

can see the concentrations similar to the sensor readings, so we can say that the sensor 2610 is more responsive to the Methane and not to the Ethylene.



Feature importance analysis was helpful to identify most significant sensor responses that contribute to classification and identification of gasses, which can help improve accuracy and reliability of gas identification and quantification.



Helps to visualize correlation between different sensor responses to different gasses. This can provide insights into underlying patterns and relationships between different sensors and gasses.

3.4. Analysis of results

Accuracy Before Preprocessing:

Model	Methane	Ethylene
LDA	0.9330525948999929	0.9049099869235496
QDA	0.9222793612259316	0.8740788569306144
LR	0.9403240968538022	0.9130837256587525

Accuracy After Preprocessing:

Model	Methane	Ethylene
LDA	0.9390203784822619	0.9044925771383171
QDA	0.9295770786802177	0.8857002422024843
LR	0.9436619515094852	0.9124920885010571

Performance of fine-tuned features

Model	Methane	Ethylene
LDA	0.9411863157955488	0.9060610491120811
QDA	0.9327039058427005	0.8911695789431159
LR	0.9451661276887177	0.9164210645792413

Cross validation analysis of :

LDA :

For Methane :

Model accuracy running cross validation with two k-folds:
[0.94563814 0.93257768]

Model accuracy running cross validation with five k-folds:
[0.94564393 0.9347061 0.92965741 0.94249977 0.94102681]

For Ethylene :

Model accuracy running cross validation with two k-folds:
[0.90503748 0.90421459]

Model accuracy running cross validation with five k-folds:
[0.90582914 0.90650246 0.91870035 0.8920754 0.8983684]

QDA:

For Methane :

Model accuracy running cross validation with two k-folds:
[0.93080781 0.92206765]
Model accuracy running cross validation with five k-folds:
[0.9250043 0.9265373 0.90784669 0.93417827 0.93482265]
For Ethylene :
Model accuracy running cross validation with two k-folds:
[0.87452747 0.8740018]
Model accuracy running cross validation with five k-folds:
[0.88751486 0.89903583 0.8883493 0.87571856 0.85713358]

LR :
For Methane :
Model accuracy running cross validation with two k-folds:
[0.94361241 0.93898046]
Model accuracy running cross validation with five k-folds:
[0.94757852 0.93323081 0.93675972 0.94696171 0.94625947]
For Ethylene :
Model accuracy running cross validation with two k-folds:
[0.91173749 0.91108197]
Model accuracy running cross validation with five k-folds:
[0.915662 0.91186857 0.9191873 0.90256595 0.90720933]

4. Discussion & Conclusions

4.1. Decisions made

- The team had to determine what data set would be a good fit for the project based on requirements on data size.
- Throughout the duration of the project, a team had to make specific decisions together to determine how to move forward.
- Our team also had to decide between various pre-processing techniques in order to find better results from data mining.

4.2. Difficulties faced

- The type of data we were working with was something our team wasn't very familiar with. As such, it took some time for our team to find a good pre-processing technique for gas sensor data.
- Our team had issues with arranging times to discuss projects for longer periods of time at first. We found that we worked better having a more adaptive schedule for meetings.

4.3. Things that worked

- We found that the overall result improved when we pre-processed data. We were hesitant at first, but removing the initial period of time where sensors were actualizing proved to have better results.
- The second sensor appeared to not have proper functionality, so we found that removing it from data gave improved outcomes.

4.4. Things that didn't work well

- We had initially considered using the K-Fold algorithm as part of this project, but we ran into issues on getting it to work with our dataset. We found that the project was more successful without an algorithm altogether.
- One of the challenges encountered during this project was large memory demand of code, which caused frequent crashes of Jupyter notebook. This was especially problematic when working with

large datasets or complex models.

4.5. Conclusion

In retrospect, our team found that trying to work with a gas sensor dataset proved to be a difficult but enjoyable challenge. It was interesting having to consider which pre-processing, regression, and classification techniques would work best for a data set.

In this project, we have explored the potential of gas sensor array analysis for dynamic gas mixtures using machine learning and data mining techniques. We have shown that gas sensor array data contains valuable information that can be used to identify and classify different gasses with high accuracy and reliability. We have also demonstrated the importance of preprocessing and feature selection techniques, such as K-fold cross-validation, PCA, and correlation matrix heatmap, in improving performance of models and interpreting results. Despite challenges of dealing with noisy and complex gas sensor data, we have achieved promising results in identifying and classifying various gas mixtures.

Code Repository : https://github.com/vaibhavingale-sjsu/cmpe255.git

5. Project Plan / Task Distribution

Task	Assigned To	Completed by
Picking a Dataset	Everyone	Everyone
Breaking Down Dataset Formatting	Everyone	Everyone
Preprocessing	Everyone	Everyone
Logistic Regression	Lalitha	Lalitha
LDA / QDA	Vaibhav	Vaibhav
PCA	Hardwin	Hardwin
Project Documentation	Everyone	Everyone
Powerpoint Presentation	Everyone	Everyone

6. References

[1] https://www.researchgate.net/publication/273478538_Chemical_Gas_sensor_array_dataset
[2] https://archive.ics.uci.edu/ml/datasets/gas+sensor+array+under+dynamic+gas+mixtures#
[3] J. Fonollosa, et al., Chemical gas sensor array dataset, Data in Brief (2015)
[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6780764/
[5] https://www.mdpi.com/1424-8220/21/14/4826
[6] https://onlinelibrary.wiley.com/doi/full/10.1002/aisy.202200169