

Automated Fish Classification Using Unprocessed Fatty Acid Chromatographic Data: A Machine Learning Approach

Jesse Wood, Bach Nguyen, Bing Xue, Mengjie Zhang, Daniel Killeen

INTRO

- Gas Chromatography is an analytical chemistry method that produces high-dimensional low-sample data.
- This study compares classification and feature selection when applied to gas chromatography data from fish oil.
- Classification to predict Fish Species and Body Parts, two datasets that share the same features.
- Feature selection to reduce the dimensionality, improve computational efficiency, and (even) improve classification performance.

METHODS

- Evaluation: average balanced classification accuracy over 10-fold cross-validation.
- 30 independent runs for each classification method.
- Each feature selection method for number of features in { 50, 100, ..., 4800 }. PSO was evaluated on 30 independent runs.

RESULTS

- Classification: Linear SVM performed best, with near-perfect accuracy for fish species, and best accuracy for body parts.
- Feature Selection: PSO has improves accuracy for fish species, using 25% of the features making it 4 times faster.

DISCUSSION

- Linear SVM provides an interpretable and accurate model.
- PSO/mRMR feature selection improve accuracy and efficiency.
- Body Parts is harder to classify than Fish Species

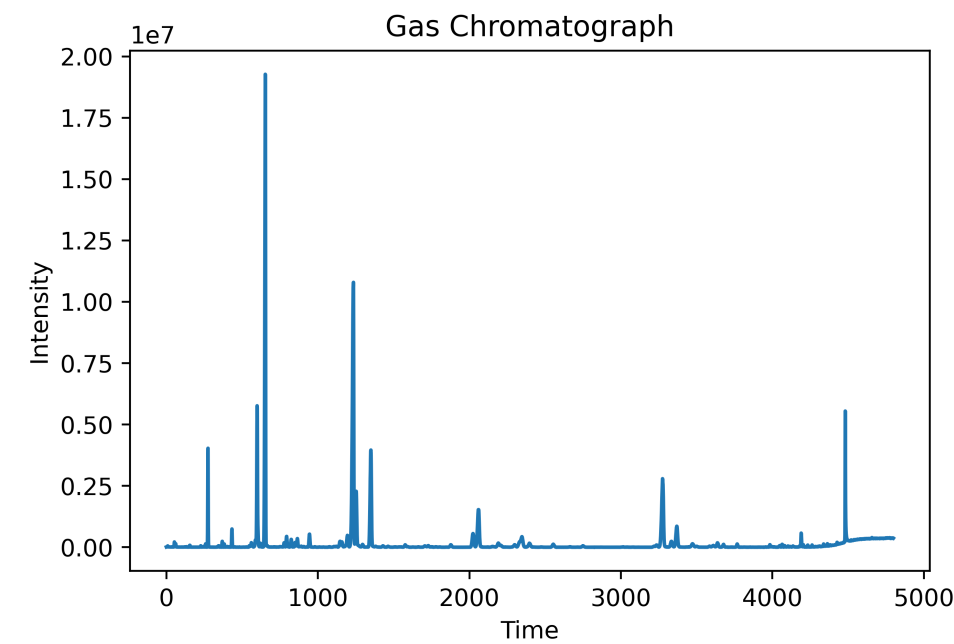


Linear SVM can accurately predict fish species, PSO makes that process 4 times faster, producing an accurate, interpretable and efficient model for Gas Chromatography.



Take a picture to download the full paper

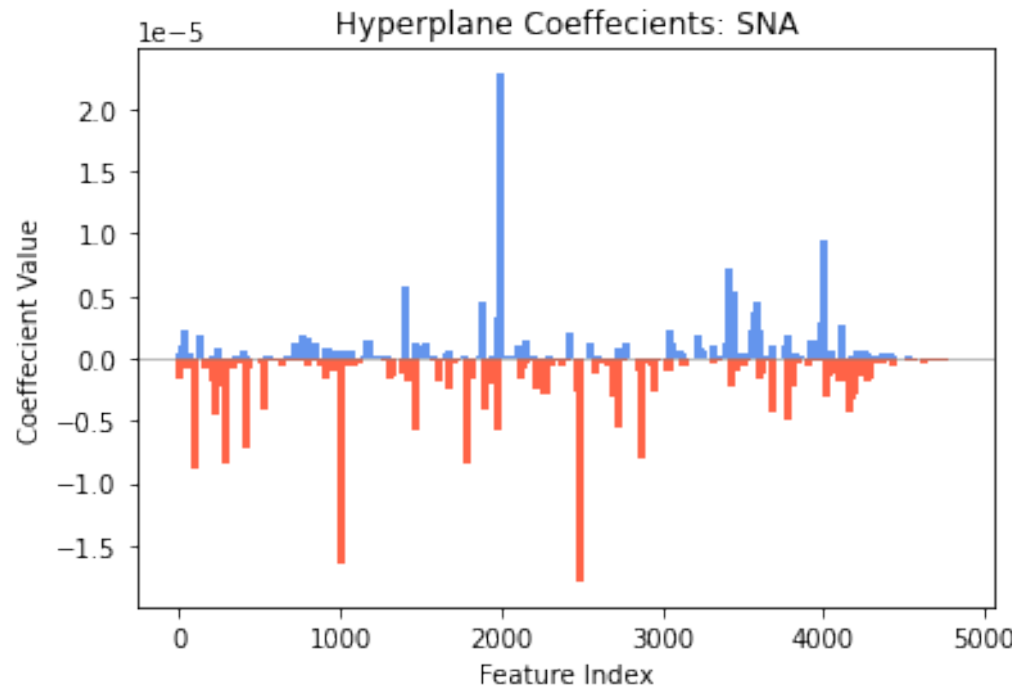
Gas Chromatogram - where x-axis is time, y-axis is intensity - for the Snapper Fish Species:



A table with classification results:

Dataset	Method	Train	Test
Species	KNN	83.57	74.88
	RF	100.0	85.65
	DT	100.0	76.98
	NB	79.54	75.27
	SVM	100.0	98.33
Parts	KNN	68.95	43.61
	RF	100.00	72.60
	DT	100.00	60.14
	NB	65.54	48.61
	SVM	100.00	79.86

Linear SVM hyperplane coefficients for the Snapper Fish Species:



A table best accuracy for feature selection methods:

Dataset	Method	# Features	Train	Test
Species	ReliefF	359	100.0	98.33
	mRMR	1500	100.0	99.17
	χ^2	3250	100.0	98.33
	PSO	1192	100.0	99.17
	Full	4800	100.0	98.33
Parts	ReliefF	1650	100.0	84.44
	mRMR	1500	100.0	86.94
	χ^2	1550	100.0	82.50
	PSO	1223	100.0	84.31
	Full	4800	100.0	79.86