



Automated Fish Classification



Using Unprocessed Fatty Acid Chromatographic Data

By Jesse Wood
Supervisors: Bach Nguyen, Bing Xue

Industry Partner: Daniel Killeen

Outline

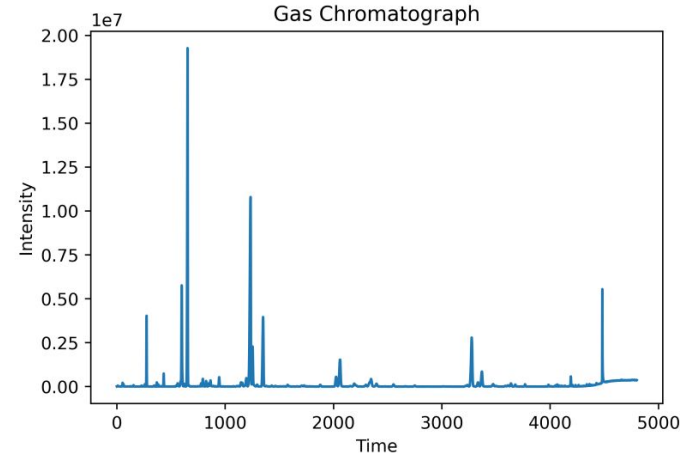
1. Gas Chromatography (GC)
2. Preprocessing
3. Classification
4. Feature Selection
5. Conclusions and Future Work



The New Zealand blue cod (*Paraperca colias*) is a temperate marine fish^[3] of the family Pinguipedidae.^[4] It is also known variously as Boston blue cod, New Zealand cod, sand perch, or its Māori names rāwaru, pākinikiri and patutuki.^[5]

Gas Chromatography (GC)

- Analysing GC (Eder 1995) is expensive and time consuming task.
- Steps:
 1. Apply heat to liquid.
 2. Evaporate into gas.
 3. Travel through long tube.
 4. Detector measures intensity.
- **Gist:** Molecules have distinct and different boiling points, these correspond to known timestamps.



The **Australasian snapper** (*Chrysophrys auratus*) or **silver seabream** is a species of porgie found in coastal waters of Australia, Philippines, Indonesia, mainland China, Taiwan, Japan and New Zealand. Its distribution areas in the Northern and Southern Hemispheres are disjunct.^[2] Although it is almost universally known in Australia and New Zealand as **snapper**, it does not belong to the snapper family, Lutjanidae. It is highly prized as an edible fish, with a sweet sea taste and a firm texture.



Prionotinae is a subfamily of demersal, marine ray-finned fishes, part of the family *Triglidae*. The fishes in this subfamily are called **sea robins** and are found in the Western Atlantic and Eastern Pacific Oceans, the other two *Triglidae* subfamilies are called gurnards.

Preprocessing

- Instrumental drift, an artifact of the gas chromatography technique, leads to missing packets in the data.
- Important to align GC data (fill in missing packets) for better results.
- Steps:
 1. Find missing timestamps,
 2. impute with zero filling
- Existing works, (Tomasi 2004, Zhang 2008), provide more advance techniques to handle instrumental drift.

Table 1: Inconsistent Timestamps

	Timestamp		
	Sample 1	Sample 2	Sample 3
Packet 1	51	50	50
Packet 2	52	51	51
Packet 3	53.05	53.1	53



Nemadactylus macropterus, the tarakihi, Jackass morwong or deep sea perch, is a species of marine ray-finned fish, traditionally regarded as belonging to the family Cheilodactylidae, the members of which are commonly known as morwongs. It is found in the south western Pacific Ocean, in Australia and New Zealand. Although there are records from the southern Indian Ocean and southwestern Atlantic, these may be due to misidentifications of similar species.

Classification: Motivations

- Reduce - byproduct in fish processing.
- Reuse - identify high-value fish oil.
- Recycle - refine fish oil, rich in omega-3, into supplements.
- Replace - Automate expensive/time-consuming of analyzing GC fish oil data.



<https://static.countdown.co.nz/assets/product-images/zoom/9400097038961.jpg>





<https://www.nutraingredients.com/Article/2016/03/10/Small-fish-oil-doses-enough-to-lower-blood-pressure-RCT>

Classification: Experiment

- Two datasets: fish species, body parts.
- 5 classifiers:
 1. KNN (Fix 1989)
 2. RF (Ho 1995)
 3. DT (Loh 2011)
 4. NB (Hand 2001)
 5. SVM (Cortes 1995)
- Run each classifier 30 independent runs, 10-fold cross-validation, balanced accuracy.
- SVM with Linear Kernel performs best for both datasets. Near perfect accuracy for fish species.
- Body parts more difficult to classify than fish species.



Table 2: Classification Accuracies

Dataset	Method	AvgTrain \pm Std	AveTest \pm Std
 Fish Species	KNN	83.57 \pm 1.80	74.88 \pm 12.54
	RF	100.0 \pm 0.00	85.65 \pm 10.76
	DT	100.0 \pm 0.00	76.98 \pm 13.12
	NB	79.54 \pm 1.60	75.27 \pm 4.35
	SVM	100.0 \pm 0.00	98.33 \pm 5.00
 Body Parts	KNN	68.95 \pm 3.49	43.61 \pm 13.48
	RF	100.00 \pm 0.00	72.60 \pm 16.15
	DT	100.00 \pm 0.00	60.14 \pm 14.57
	NB	65.54 \pm 2.69	48.61 \pm 12.19
	SVM	100.00 \pm 0.00	79.86 \pm 8.52

Classification: Results

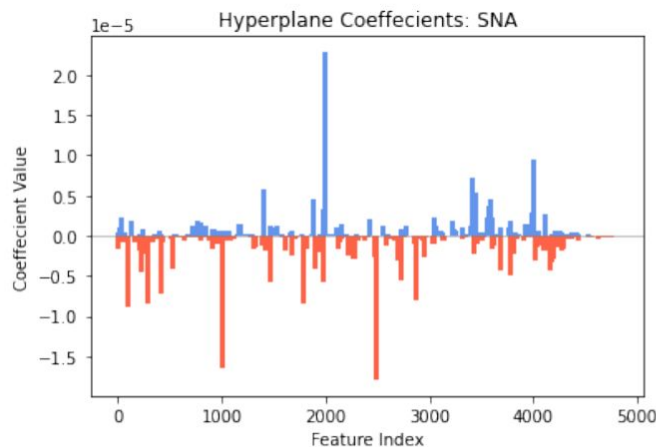
- RF, DT overfit the training data → they memorize, not learn.
- KNN performs worse → distance-based classification algorithm suffers in high-dimensions.
- NB performs poorly → Assumption of conditional independence may not hold.
- *Why is body parts is harder?*
Perhaps, more similarities between fish of same species, than body parts from different species.

Table 2: Classification Accuracies

Dataset	Method	AvgTrain \pm Std	AveTest \pm Std
 Fish Species	KNN	83.57 \pm 1.80	74.88 \pm 12.54
	RF	100.0 \pm 0.00	85.65 \pm 10.76
	DT	100.0 \pm 0.00	76.98 \pm 13.12
	NB	79.54 \pm 1.60	75.27 \pm 4.35
	SVM	100.0 \pm 0.00	98.33 \pm 5.00
 Body Parts	KNN	68.95 \pm 3.49	43.61 \pm 13.48
	RF	100.00 \pm 0.00	72.60 \pm 16.15
	DT	100.00 \pm 0.00	60.14 \pm 14.57
	NB	65.54 \pm 2.69	48.61 \pm 12.19
	SVM	100.00 \pm 0.00	79.86 \pm 8.52

Classification: Interpret Linear SVM

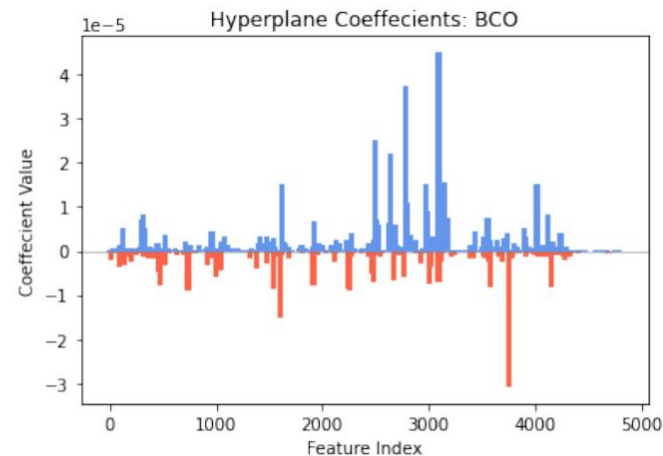
- SVM uses a one-vs-all approach for multi-class classification, it breaks into a series of binary classification tasks, with one hyperplane for each class.
- GC data is non-negative, only negative weights push toward the negative class, therefore positive weights are expected molecules, and the negative values are not.
- Both figures show most features have small weights, this suggests not all features are needed.



(a) Snapper



The **Australasian snapper** (*Chrysophrys auratus*) or **silver seabream** is a species of porgie found in coastal waters of Australia, Philippines, Indonesia, mainland China, Taiwan, Japan and New Zealand. Its distribution areas in the Northern and Southern Hemispheres are disjunct.^[2] Although it is almost universally known in Australia and New Zealand as **snapper**, it does not belong to the snapper family, Lutjanidae. It is highly prized as an edible fish, with a sweet sea taste and a firm texture.



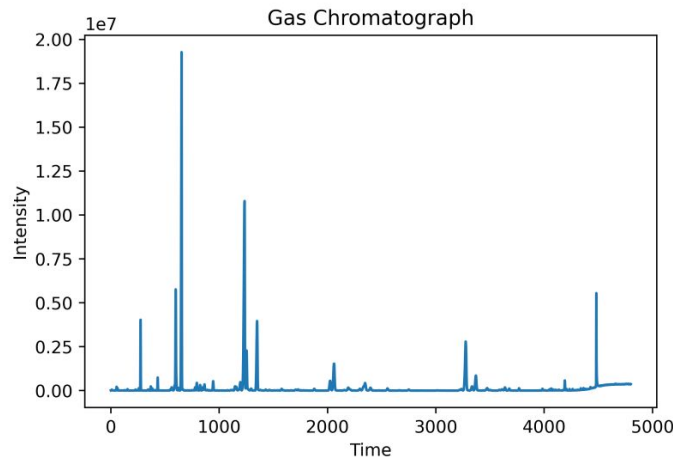
(b) Blue cod



The **New Zealand blue cod** (*Paraperis colias*) is a temperate marine fish^[3] of the family Pinguipedidae.^[4] It is also known variously as **Boston blue cod**, **New Zealand cod**, **sand perch**, or its Māori names **rāwaru**, **pākirikiri** and **patutuki**.^[5]

Feature Selection: Motivations

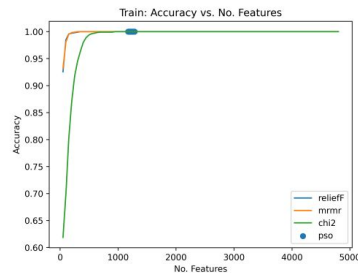
- High-dimensionality (4800 features).
- Redundant and correlated features hinder classification accuracy.
- Performance - less features improves compute and (potentially) accuracy.
- Interpretability - troubleshooting/diagnosis in a factory settings.



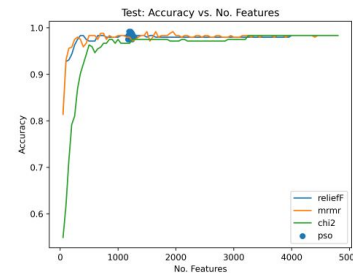
The **Australasian snapper** (*Chrysophrys auratus*) or **silver seabream** is a species of porgie found in coastal waters of Australia, Philippines, Indonesia, mainland China, Taiwan, Japan and New Zealand. Its distribution areas in the Northern and Southern Hemispheres are disjunct.^[2] Although it is almost universally known in Australia and New Zealand as **snapper**, it does not belong to the snapper family, Lutjanidae. It is highly prized as an edible fish, with a sweet sea taste and a firm texture.

Feature Selection: Experiment

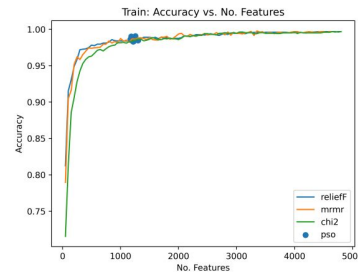
- Two datasets: fish species, body parts
- 4 feature selection methods:
 1. mRMR (Ding 2005)
 2. Relief-F
 3. Chi2 (Liu 1995)
 4. Wrapper PSO (Kennedy 1995, Nguyen 2020)
- 10-fold cross-validation, balanced accuracy.
- Run 1-3 for a range of k in $[0, 4800]$ in increments of 50.
- PSO automatically selects k , evaluate with 30 independent runs. PSO wrapped in SVM classifier.



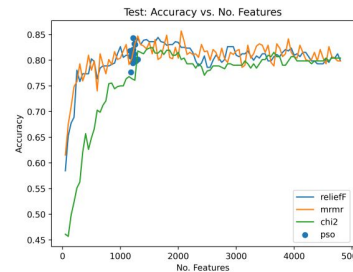
(a) Species: Training set



(b) Species: Test set



(a) Part: Training set



(b) Part: Test set

Feature Selection: Results

- Body parts still harder than classifying fish species → PSO and ReliefF requires more features this dataset.
- PSO achieves best classification performance on fish species, overfits on body parts.
- Most features are not needed
 - a. Fish species → 75% features removed.
 - b. Body parts → 60% features removed.
- Automated classification can be up to 4 times faster.



Table 3: Best accuracy on Fish Species.

Method	Number of features	Training accuracy	Testing accuracy
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



Table 4: Best Accuracy on Fish Body Parts

Method	Number of features	Training Accuracy	Testing Accuracy
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

Conclusions and Future Work

- Linear SVM is interpretable and accurate for fish oil classification.
- It is more difficult to classify body parts than fish species.
- FS methods improved efficiency and accuracy on both datasets.
- Future work would improve performance on body parts dataset:
 - More sophisticated imputation for preprocessing
 - feature construction
 - transfer learning

References

1. Eder, K. (1995). Gas chromatographic analysis of fatty acid methyl esters. *Journal of Chromatography B: Biomedical Sciences and Applications*, 671(1-2), 113-131.
2. Tomasi, G., Van Den Berg, F., & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(5), 231-241.
3. Zhang, D., Huang, X., Regnier, F. E., & Zhang, M. (2008). Two-dimensional correlation optimized warping algorithm for aligning GC× GC- MS data. *Analytical Chemistry*, 80(8), 2664-2671.
4. Bi, K., Zhang, D., Qiu, T., & Huang, Y. (2019). GC-MS Fingerprints Profiling Using Machine Learning Models for Food Flavor Prediction. *Processes*, 8(1), 23.
5. Matyushin, D. D., & Buryak, A. K. (2020). Gas Chromatographic Retention Index Prediction Using Multimodal Machine Learning. *Ieee Access*, 8, 223140-223155.
6. Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3), 238-247.
7. Hand, D. J., & Yu, K. (2001). Idiot's Bayes—not so stupid after all?. *International statistical review*, 69(3), 385-398.
8. Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.
9. Loh, W. Y. (2011). *Classification and regression trees*. Wiley interdisciplinary reviews: data mining and knowledge discovery,
10. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
11. Liu, H., & Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of 7th IEEE international conference on tools with artificial intelligence* (pp. 388-391). IEEE.
12. Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02), 185-205.
13. Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning*, 53(1), 23-69.
14. Eberhart, R., & Kennedy, J. (1995, November). Particle swarm optimization. In *Proceedings of the IEEE international conference on neural networks* (Vol. 4, pp. 1942-1948).
15. Nguyen, H. B., Xue, B., Andreae, P., & Zhang, M. (2017, June). Particle swarm optimisation with genetic operators for feature selection. In *2017 IEEE Congress on Evolutionary Computation (CEC)* (pp. 286-293). IEEE.

