Machine Learning for Fish Oil Analysis *

Jesse Wood[1[0000−1111−2222−3333]], Bach Hoai Nguyen[1[1111−2222−3333−4444]], Bing Xue[1[2222−−3333−4444−5555]], Mengjie Zhang[1[2222−−3333−4444−5555]], and Daniel Killeen[2[2222−−3333−4444−5555]]

[1] Victoria University of Wellington, Te Herenge Waka, PO Box 600, Wellington 6140, New Zealand
jesse.wood@ecs.vuw.ac.nz
bach.nguyen@ecs.vuw.ac.nz
bing.xue@ecs.vuw.ac.nz
mengjie.zhang@ecs.vuw.ac.nz
[2] Plant and Food Research, Port Nelson, Nelson 7010, New Zealand
Daniel.Killeen@plantandfood.co.nz

Abstract. Gas chromatography (GC) can be used to identify chemical compounds present within tissue samples for quality assurance in food science. Existing analytical chemistry techniques for processing GC data are manual and time-consuming. Here, we explore classification algorithms for fish oil data that automate and significantly reduce the time required to process GC data. We find the Linear SVC model can predict the fish species with near-perfect accuracy. The fish oil data is high-dimensional and low sample size. We compare state-of-the-art feature selection methods to reduce the dimensionality of the data. High accuracy is possible with very few features for the MRMR and ReliefF feature selection methods. Visualisation is used to explore the interpretability of the models such that their efficacy can be verified for use in a factory setting. The exploration reveals there are many feature subsets all capable of producing high-accuracy predictions. No clear superset of important features emerges, which indicates there are many important features to choose from.

Keywords: Feature Selection · Gas Chromatography · Support Vector Machines · Food Science

1 Introduction

Fish oil analysis involves using analytical chemistry techniques, such as gas chromatrography, to anaylse the structure of the chemical compounds. We can use these structures to determine which chemicals are present in a given sample. This is important for quality assurance in a factory setting, especially in food science. We want to be confident that our food labels are accurate and reduce/eliminate cross contamination between different food products. To identify cross-contanimation we use fish classification. Given a fish oil sample, we

* Supported by organization Plant and Food Research.

can identify the fish species (i.e. Bluecod, Tarakihi), and part (Head, Fins). The existing techniques for performing fish classification are time consuming and laborious. Chemists compare a given sample to reference samples to determine which class it likely belongs to. Previous work on gas chromatrography CITE HERE, has shown machine learning can be used to automate classification.

In this paper we explore machine learning techniques to automate the process of identifying fish species and part on GC data. Firstly, classifications algorithms are evaluated for their ability to determine the fish species and part. Visualisation is used to explore the interpretability of successful models. It is important to verify their efficacy with domain knowledge before these algorithms can be deployed in a real-world setting. Secondly, feature selection is used to eliminate redundant features, whilst maintaining high-accuracy predictions.

Specifically, our work is divided into two main sections:

1. Classification Algorithms
2. Feature Selection

2 Background

- Chromatrography methods: how the raw fish oil data is collected.
- Classification algorithms: introduce classification algorithms used in the paper.
- Feature selection: main concepts.

3 Data processing

- Why the raw data is not applicable to existing classification algorithms?
- Extracting datasets that are ready for classification algorithms:
  - Sum up the intensity.
  - Aligning missing packets.
- Overview of extracted data.

4 Applying classification algorithms to the processed fish data

- Fish types:
  - Discussion on the accuracy of different classification algorithms.
  - Visualisation of SVM hyperplanes.
- Fish parts:
  - Discussion on the accuracy of the different classification algorithms.
  - Further discussion on the challenges of fish part in comparison.

5 Feature selection for fish data

- Why feature selection on this data?
- Breif the main ideas of the feature selection algorithms that were used.
- Compare the performance of selected features and usuing all features.
- (Optional): analyse the selected features.

6 Conclusions and future work

- Summarize the achieved results to show the effectiveness of machine learning algorithms that were used.
- Future work: further steps, improve fish part performance, more challenging datasets, different tasks.

References