# Fishy Business - Multi-Tree GP Feature Construction for Multi-class Fish Classification

Jesse Wood[1[0000−0003−3756−2122]], 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, bach.nguyen, bing.xue, mengjie.zhang}@ecs.vuw.ac.nz
[2] Plant and Food Research, Port Nelson, Nelson 7010, New Zealand
daniel.killeen@plantandfood.co.nz

**Abstract.** Fish is approximately 40% edible fillet. The remaining 60% can be processed into low-value fertilizer or high-value pharmaceutical-grade omega-3 concentrates. High-value manufacturing options depend on the composition of the biomass, which varies with fish species, fish tissue and seasonally throughout the year. Fatty acid composition, measured by Gas Chromatography, is an important measure of marine biomass quality. This technique is accurate and precise, but processing and interpreting the results is time-consuming and requires domain-specific expertise. The paper investigates Multi-tree Feature Construction for Multi-Class classification of Gas Chromatography data.

**Keywords:** AI applications · Classification · Feature selection · High-dimensional data · Multidisciplinary · Gas Chromatography · Genetic Program · Fatty Acid

## 1 Introduction

"Kindly let me help you or you will drown said the monkey putting the fish safely up a tree." – Alan Watts.

– Wrapper-based Multi-Tree GP for Feature Construction for Multi-class Classification on Gas Chromatography data.

## 2 Gas Chromatography

– Destructive chemistry technique used to analyze chemical compounds in fish tissue.
– Manual laborious, time-consuming and expensive task.
– Gas chromatograph is intensity vs. time.
– Instrumental drift / alignment issue

## 3 Preprocessing

- Find missing timestamps.
- Impute with zero filling.
- Better classification accuracy for KNN.

## 4 Genetic Programming

Algorithm 1 shows the pseudo-code of the GP algorithm used for multiple-feature construction using multi-tree representation to construct $m$ new features, with elitism ratio $e$.

**Algorithm 1** GP-based multiple feature construction

**Input** : train_set, $m$;
**Output** : Best set of $m$ constructed features;
Initilize a population of GP invidiuals. Each individual is an array of $m$ trees;
best_inds ← the best $e$ individuals;
**while** Maimum generation is not reached **do**
  **for** $i = 1$ to Population Size **do**
    $transf\_train$ ← Calculate constructed features of individual $i$ on train_set;
    $fitness$ ← Apply fitness function on $transf\_train$;
    Update best_inds the best $e$ individuals from elitism and offspring combined;
  **end for**
  Select parent individuals using tournament selection for breeding;
  Create new individuals from selected parents using crossover or mutation;
  Place new individuals into population for next generation;
**end while**
Return best individual in best_inds;

## 5 MCIFC

A multiple class-independnet feature consturction method (MCIFC) [3].

### 5.1 Representation

MCIFC is a Multi-tree GP that constructs a smaller number of high-level features, proportional to the number of classes, from the original features. This method is based on the intuition that problems with more classes are likely to be more complex, and thus require more features to capture said complexity. The number of constructed features $m$, determined by $m = r \times c$, where $r$ is the construction ratio (set to 2), and $c$ is the number of classes. MCIFC constructs 8 features for the 4-class fish species problem and 12 features for the 6-class fish species problem.

### 5.2 Crossover and Mutation

MCIFC limits both the crossover and mutation operators to only one of the constructed features described in Algorithm 2. This approach favours exploitation over exploration, making small random changes to constructed features with monotonically increasing fitness due to elitism.

**Algorithm 2** MCIFC Crossover and Mutation.

$prob \leftarrow$ randomly generated probability;
$doMutation \leftarrow (prob < mutationRate)$;
**if** $doMutation$ **then**
  $p \leftarrow$ Randomly select an individual using tournament selection;
  $f \leftarrow$ Randomly select a feature/tree from $m$ trees of individual $p$;
  $s \leftarrow$ Randomly select a subtree in $f$;
  Replace $s$ with newly generated subtree;
  Return one new individual;
**else**
  $p1, p2 \leftarrow$ Randomly select 2 individuals using tournament selection;
  $f1, f2 \leftarrow$ Randomly select a features/trees from $m$ trees of $p1$ and $p2$, respectively;
  Swap $s1$ and $s2$;
  Return two new individuals;
**end if**

### 5.3 Fitness

MCIFC takes the balanced classification accuracy of an SVM classifier as the fitness function. The SVM classifier is known to be effective for fish oil data [1]. Balanced accuracy avoids results bias towards the majority class, which is relevant for the fish species dataset, with the majority class 44% of samples belonging to fish species blue cod. The balanced accuracy is given by

$$\text{Balanced Accuracy} = \frac{1}{c}\sum_{i=1}^{c}\frac{TP_i}{TP_i + FN_i} \tag{1}$$

Where $TP_i$ is the number of true positives for class $i$, and $FN_i$ is the number of false negatives for class $i$, c is the number of classes.

## 6 Experimental Setup

Table 1 shows the datasets used in the experiments and their respective characteristics. Due to the high dimensionality of gas chromatography data, this paper employs a GP-based FC approach. The dataset is suited towards dimensionality

Table 1: Datasets.

| Dataset | Features | Instances | Classes | Class Distribution |
|---|---|---|---|---|
| Fish Parts | 4800 | 153 | 4 | 44% 17% 20% 19% |
| Body Parts | 4800 | 153 | 6 | 15% 22% 14% 22% 14% 13% |

reduction, as previous work [1] demonstrated FS can improve classification accuracy. The small number of instances is due to the expensive and time-consuming nature of performing Gas Chromatography on fish tissue.

The data is pre-processed to fix the instrumental drift by imputing missing timestamps with zero filling. Features are normalized in the range [0,1] based on the training set.

Table 2: Paramter settings.

| | |
|---|---|
| Function Set | $+, -, *$, neg, protectedDiv |
| Teriminal Set | $x_1, x_2, ..., x_n$, $r \in [-1, 1]$ |
| Maximum Tree Depth | 8 |
| Population size | 100 |
| Initial Population | Ramped Half and Half |
| Generations | 300 |
| Crossover | 0.9 |
| Mutation | 0.1 |
| Elitism | 0.1 |
| Selection | Tournament |
| Tournament Size | 3 |
| Construction ratio | 2 |

Table 2 describes the parameter settings of all GP-based methods used in the experiments. The function set has standard arithmetic operators $+, -, \times$, a protected division operator that prevents division by zero returning 0 instead, and the unary *neg* operator reverses the sign. The feature set, and randomly generated constant $r \in [-1, 1]$, are used in the terminal set. A population of 100 individuals is used for all experiments, with 300 generations. The construction ratio $r$ used to determine the number of features constructed is experimentally chosen as 2.

## 7 Results

Table 3 shows the results of the experiments. We give the train and test accuracy for the best individual found in the final generation for the fish species dataset. The fitness is measures as the balanced classification accuracy on a train-test split of the dataset (66% train, 33% test).

Table 3: Initial Results - Fish Species.

| Method | Train | Test |
|---|---|---|
| ST | 0.53 | – |
| MT | 0.55 | – |
| MCIFC | – | 0.8673 |
| SVM | 1.0 | 92.4 |

The table gives results for:

1. ST - Single-Tree GP with classification map [2]
2. MT - Multi-tree GP with one-vs-rest approach
3. MCIFC - Multiple Class-Independent Feature Construction [3]
4. SVM - Linear Support Vector Machine one-vs-rest approach [1]

## 8 Discussion

The balanced classification accuracy of 86.73% is a significant improvement on the previous Single-Tree and Multi-tree GP classification, whose best results both could not exceed 55% accuracy on the training set. Both these approaches fail to capture the complexity of the fish species dataset. For Single-tree GP, this is likely due to the difficulty of the GP finding a single expression that can fit the class boundaires for the classification map. For Mulit-tree GP using the one-vs-rest approach, we see improvement over Single-Tree GP, but still not as good as MCIFC.

The Multi-tree GP has to learn how to perform accurate classification, which is a difficult task. MCIFC is a wrapper-based method, which plays to the strengths of GP for feature construction, and SVM for classification tasks.

The Linear SVM classifier has a balanced classification accuracy of 92.4% on the test set, which is a significant improvement on the previous best result of 86.73% using MCIFC. However, after tweaking the parameters of MCIFC we expect to match (or exceed) the SVM classifier in the future.

(*Note:* The code needs to be adjusted to record train and test accuracy for each individual to get more comprehensive results, but these initial results were a proof of concept.)

## 9 Conclusion

## 10 Future Work

Add the Czekanowski distance metric to the fitness function, which puts pressure on the GP to construct features that differentiate effectively between classes in relation to their distance in the feature space [3]. Their paper suggests using an $\alpha = 0.8$ term to bias the fitness function towards accuracy.

Run with the same parameters settings from [3], a single generation took 1 hour and 28 minutes to evaluate, and with random tree generation, produced the best individual with a test accuracy of 71%. Compared to an initial fitness of 58% for the parameter settings given above. It would take 48 hours to evaluate 30 generations, so the results are not included in this paper, but reducing the construction ratio $r$, or population coefficient $\beta$, would significantly reduce this computation.

## References

1. Long, G.: AI 2021: Advances in Artificial Intelligence, vol. 13151. Springer Nature (2022), placeholder for AJCAI 2022
2. Smart, W.R.: Genetic programming for multiclass object classification. BSc (Honours) Research Project (2005)
3. Tran, B., Xue, B., Zhang, M.: Genetic programming for multiple-feature construction on high-dimensional classification. Pattern Recognition **93**, 404–417 (2019)