AFRL-RH-WP-TR-2020-0113

HYBRID FEATURE SELECTION WITH GENETIC ALGORITHMS AND OTHER METHODS

Timothy Ho

Department of Biomedical Engineering
Rensselaer Polytechnic Institute

Daniel W. Cowan, Patrick M. McLendon
UES, Inc

Heather A. Pangburn
711 HPW/RHBB

OCTOBER 2020
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2020-0113 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

GROGG.MATTHEW.W.1107142980 Digitally signed by GROGG.MATTHEW.W.1107142980 Date: 2021.02.03 12:28:08 -05'00'

MATTHEW GROGG
In Vitro Models Team Lead
Biotechnology Branch

PANGBURN.HEATHER.ANN.1393596070 Digitally signed by PANGBURN.HEATHER.ANN.1393596070 Date: 2021.02.17 10:01:44 -05'00'

HEATHER PANGBURN, DR-III, PhD
Core Research Area Lead
Biotechnology Branch
Airman Biosciences Division

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
16 10 20	Interim	June 2020 – August 2020

4. TITLE AND SUBTITLE
Hybrid Feature Selection with Genetic Algorithms and other Methods

5a. CONTRACT NUMBER
FA8650-17-C-6834

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

TIMOTHY HO,[1] DANIEL W. COWAN,[2,3] PATRICK M. MCLENDON,[2,3] HEATHER A. PANGBURN[3]

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
Legacy RHM

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
1 Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180

2 UES Inc, 4401 Dayton-Xenia Rd, Dayton, OH 45232

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
3 Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Airman Biosciences Division
Biotechnology Branch
Wright-Patterson AFB, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHBB

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2020-0113

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A. Approved for public release.

13. SUPPLEMENTARY NOTES
88ABW-2020-3190, cleared on 16 October 2020

14. ABSTRACT
Feature Selection (FS) processes are becoming more important for image-base cell profiling as the field begins to rely more heavily on computational means of analysis. A single image analysis can be scanned for more than a thousand different features each with varying metrics. Current researchers have realized the holistic approach to measure every possible feature lead to the issue of determining the features that provide an adequate experimental analysis. Past FS methods were to preselect features from a smaller feature set, however, increases in computational speed and advances in machine learning methods have allowed quick expansive analysis to become more prevalent. After image-profiling, researchers obtain a plethora of data that is difficult for analysis. Many linear FS methods have been designed fail to account for the significant variability that exist with biological data. This work in this paper aims to solve the problem of linear feature selection methods by proposing a new feature selection technique using machine learning with a ridge metric.

15. SUBJECT TERMS
Feature Selection, Hybrid Genetic Algorithm, Genetic algorithms, Recursive feature elimination

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Matthew Grogg
Unclassified	Unclassified	Unclassified	SAR	9	19b. TELEPHONE NUMBER *(Include Area Code)* N/A

TABLE OF CONTENTS

Hybrid Feature Selection with Genetic Algorithms and other Methods

TIMOTHY HO,[1] DANIEL W. COWAN,[2,3] PATRICK M. MCLENDON,[2,3] HEATHER A. PANGBURN[3]

[1]*Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*
**Timothyho201@gmail.com*

2 UES Inc, 4401 Dayton-Xenia Rd, Dayton, OH 45232

3 711th HPW, Air Force Research Laboratory, WPAFB OH 43433

Abstract

Purpose Feature Selection (FS) processes are becoming more important for image-base cell profiling as the field begins to rely more heavily on computational means of analysis. A single image analysis can be scanned for more than a thousand different features each with varying metrics. Current researchers have realized the holistic approach to measure every possible feature lead to the issue of determining the features that provide an adequate experimental analysis. Past FS methods were to preselect features from a smaller feature set, however, increases in computational speed and advances in machine learning methods have allowed quick expansive analysis to become more prevalent. After image-profiling, researchers obtain a plethora of data that is difficult for analysis. Many linear FS methods have been designed fail to account for the significant variability that exist with biological data. This work in this paper aims to solve the problem of linear feature selection methods by proposing a new feature selection technique using machine learning with a ridge metric.

Methods This paper proposes the use of genetic algorithms to determine a feature set that provides a clear signal between a positive and negative control group to separate and identify features that provide clear distinctions between the two groups.

Results The Hybrid Genetic Algorithm Feature Selection model currently outperforms the previous method used for the analysis, providing a smaller distinct feature set that provides better separations between positive and negative controls. The results indicate that increasing feature sets possess increasing complexity that decreases the mean distance between the two control groups.

Conclusion A hybrid genetic algorithm feature selection outperform linear feature selection techniques for identifying a subset of features providing a clear distinct difference between the control groups.

References and links

1. J. C. Caicedo, S. Cooper, F. Heigwar, et al, “Data analysis strategies for image-based cell profiling” (2017)
2. G. Chandrashekar and F. Sahin, “A survey on feature selection methods” (2013)
3. Sklearn library
4. DEAP library

1. Introduction

Image-based cellular analysis has become more prevalent for quantifying the phenotypes of individuals through the morphological features observed [1]. Advancement in automated microscopy allow for the evaluation of many different variable in a more reasonable timeframe. The ability to quantify and measure every available feature gives researchers access to information that could be beneficial for genotype analysis. Although researchers are given a larger dataset, not every feature will be beneficial for analysis and complexity in determining those features become more difficult as more than a thousand features are measured within a given experiment. With the increased reliance on automated microscopy, biological data analysis inherits the same issues that plague machine learning processes.

Feature selection (FS) engineering is becoming more important for biological data as biology shift towards automated systems. A feature is a distinctive measurable attribute of an object [2]. The application of image-based analysis has expanded upon the domain of features from only a couple to tens of thousands of features. As Candrashekar et al [2] mentions, FS or feature elimination aims to reduce the number of features to solve three main issues with extraneous features: (i) reducing computational time, (ii) reducing the curse of dimensionality, and (iii) improving overall accuracy.

Recently different types of FS methods have been developed to reduce the feature domain. FS techniques have been classified into three main methods: (i) filter, (ii) embedded, and (iii) wrapper [1, 2]. Although many of these methods work well with smaller datasets, they tend to decrease in efficiency and accuracy as the size of feature domains exceed a certain number. In our application, the datasets exceed 10000 features which is far above the optimal limit with most linear FS techniques. Thus, a new method is needed to perform FS for our dataset.

2. Methods

The proposed hybrid FS technique is a combination of multiple FS techniques, the main components being GA and RFE. For my application, the feature set is filtered in four main steps, preprocessing, redundancy extraction, genetic algorithm filter, and recursive feature extraction ranking. The rest the section will introduce more technical information of terminologies and techniques incorporated into the proposed hybrid method.

2.1 Preprocessing

The preprocessing step, despite being classified as one step, is composed primarily of two separate algorithms, a constant value extractor and a two-sample t-test. The reason behind the two statistical analysis is to perform an initial inexpensive filter for removing inadequate features.

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}} \tag{1}$$

s = sample SD; X - individual value; $\bar{X}$- sample mean; n = sample size

The standard deviation is used to determine the variability of data. It generalizes the data of the population. For our data samples, it allows for the determination of features that are unable to provide a signal. This inability can be caused from a multitude of reasons. However, a feature that does not provide a signal is not helpful in a computational standpoint and will only increase computational time and complexity.

The next step in data preprocessing is examining each feature for their two-sample p-value. The definition of the statistical test is given by:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2}$$

For biological data, this test is important for determining the statistical difference between the means of two different population. It is a common application to determine if a new treatment is superior to a current treatment. For our data, the test can also serve as a determination of whether a variable or feature is providing a detectable signal through the mean.

2.2 Redundancy with Spearman Correlation

Spearman rank correlation is a function used to find highly correlated features. This is defined by:

$$\rho = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)\cdot\left(R(y_i) - \overline{R(y)}\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i) - \overline{R(y)}\right)^2\right)}} \quad (3)$$

With over ten-thousand features, there is inevitable that the chance that multiple features describe a similar result will be present within the datasets. Although having these features confirms the existence of a signal, it prevents machine learning algorithms from generalization because redundant features reinforces a certain behavior of machine learning algorithms.

2.3 Genetic Algorithm

Genetic algorithms (GA) are a heuristic search algorithm [2] and are used to find the global maximum. Mimicking the Darwin's survival-of-the-fittest, the FS is treated like a population of different individuals. The best individuals are kept, crossbred, and monitored for their fitness as describe by a given function. The model of a GA is given by:

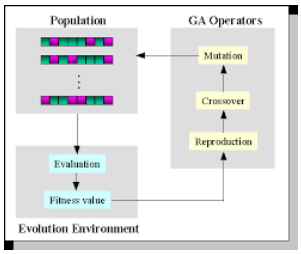

Fig. 1. Model of a general Genetic Algorithm function.

GA takes into consideration a wider subset of features, avoiding local minima and can determine a more generable subset of features. The disadvantages of this method are computational time and expenses. The genetic algorithm is used after the preprocessing step to reduce computation time and complexity for the GA.

2.4 Recursive Feature Extraction

Recursive feature elimination (RFE) is a type of FS method that removes the weakest features based on coefficients given by the metric supplied to the RFE function. This method is not only a FS method but also a tactic to rank the selected features. Features can then be filtered further for a more specific or more general subset.

2.5 Evaluation

The feature sets generated by different algorithms are evaluated by a binary classifier (BC), medoids and means distance, Silhouette Coefficient (SC), Calinski Harabasz distance (CH), and the Davies Bouldin (DB) Index. The purpose of the feature sets generation is to determine the clear boundaries and separation of the dataset. Clear separations are distinguished by clustering patterns within the feature sets. The methods mentioned are used to measure the wellness of clustering and the fitness of each point to its proposed cluster.

3. Model Assessment

DEAP was the python library used to construct the GA for the hybrid feature selection model [DEAP]. This was trained with the logistical regression algorithm to assess survival and fitness. Since the computational time for feature selection is expensive, features inputted were preprocessed for around 5000 features.

3.1 Preprocessing Inputs and Outputs

Three datasets were used to conduct FS with 11002 features with around 5000 to 7000 samples for combined positive and negative control wells. Time constraints lead to most results conducted from the first dataset. The first step of preprocessing was to run a standard deviation and mean filter to confirm constant value features that were present in both controls. Around 1000 features in all three datasets were found to not possess any signal and are eliminated immediately. After a two-sample T-test was conducted to determine features that had dissimilar means. Close to 3000 features possessed a p-value of above 0.05. The remaining features are then run through the spearman correlation las to eliminate redundant features reducing the number of features to 2800.

3.2 Genetic Algorithm for Feature Selection

The GA uses a logistic regression for analysis of the features. The population size is set to 100 and the number of generations is set to 100. After training, the best individual was tested for accuracy. The number for selected features can vary, however, the algorithm will usually select around half of the features inputted. Selected Features tend to be in the same category, cellNuc, environment, and cell, with a couple of exceptions. Feature sets sizes tend to center around half the number of inputs.

3.3 Recursive Feature Extraction Ranking

Recursive feature extraction (RFE) is a feature selection technique. However, it is computationally slow like all embedded methods. This step is necessary for determining the ranks of features and identify which features are more likely to represent the data. Of the thousand features chosen by the genetic algorithm, only the best twenty to hundred are selected. The metric of the RFE used was the logistic regression model provided by the sklearn library [sklearn].

3.4 Comparisons

The new GA/RFE process was compared to ElasticNet, Ridge, Lasso, Linear Regression, RFE with a logistic regression metric, and an RFE/GA hybrid method. Below are the results from evaluation metrics that provide values:

Table 1. Evaluation of Feature Selection Techniques on Dataset 1

Method	Set Size	BC (%)	SC	CH	DB
GA/RFE	20	98.04	0.4548	1325.3	0.9695
KS	853	97.76	0.1056	1301.59	1.7195
RFE Top 50	50	97.04	0.9735	2386.22	0.0182
RFE Top 100	100	96.96	0.9684	1783.75	0.0218
Lasso	99	95.22	0.9560	956.51	0.0305
RFE/GA	9	98.64	0.3110	2029.36	1.4440

Before discussing the analysis, each metric is evaluated differently. Binary Classification accuracy is calculated based on the percentage the neural network predicts correctly. The SC, CH, and DB are cluster evaluation methods. The SC ranged from -1 to 1 where high values indicate a great match to its neighboring cluster. CH criterion is a heuristic device and is used to determine clustering where higher values tend to indicate better cluster values. DB Index is the last metric used to analyze the feature sets. A lower value for DB indicates a better result.

The GA/RFE method performed well on the BC, the other metrics indicate subpar clustering as the SC was not the ideal value nor were the HC and DB values. Other feature sets such as RFE Top 50, RFE Top 100 and, obtained better SC and DB values. This indicates the behavior of the GA when generating a feature set. While aiming for higher accuracy, it tends to avoid features that are ideal for clustering. Despite this fact, the GA/RFE outperforms the KS metric in addition to finding a smaller and more efficient feature set.

The last metric was the means comparison as shown with the graph below:

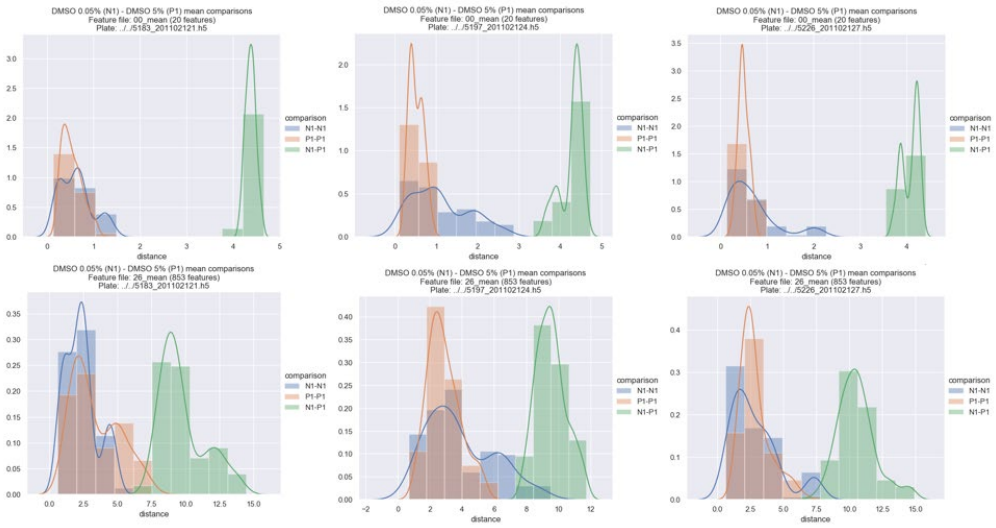

Fig. 2. Means graph where each column is a different dataset, the top is GA/RFE and the bottom is the KS.

The results indicate that the GA/RFE performed much better in terms of separation. Despite not having a larger range of values, the new feature set provides a clear distinction between the differences between the positive and negative group (N1-P1) and the positive (P1-P1) negative group (N1-N1). In addition to clear distinctions, the dataset also provide insight into another factor that influences FS, variability. Biological data can vary from sample to sample. As such, feature sets should be generated for each experimental dataset to ensure that the feature set is appropriate. Multiple feature sets could then be used to determine a universal feature set that can be applied for all experimental data in the future.

4. Conclusion

The GA/RFE method was an important development in the feature analysis for our team. Although it performed well, the results could be improved computationally. GA work best given a smaller sample sizes since complexity decreases exponentially with lower features. The most important step of the new method would be the preprocessing step. By considering for features that are inadequate for analysis, more than half the number of features is discarded immediately while being computationally inexpensive. The GA/RFE step could potentially be replace with more suitable methods to obtain better feature sets. Future techniques that could be used are Ridge coefficients, and other hybrid methods to refine the process.