AFRL-RH-WP-TR-2020-0112

EMBEDDED TECHNIQUES FOR HIGH CONTENT ANALYSIS FEATURE SELECTION

Guanpeng Andy Xu

Massachusetts Institute of Technology

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-2020-0112 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 12:24:19 -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 10:05:15 -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)*
23-10-2020	Interim	June 2020-August 2020

4. TITLE AND SUBTITLE
Embedded Techniques for High Content Analysis Feature Selection

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

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Guanpeng Xu1, Daniel W. Cowan2,3, Patrick M. McLendon2,3, Heather A. Pangburn3

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
Legacy RHM

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
1Massachusetts Institute of Technology
2UES Inc
4401 Dayton-Xenia Rd
Dayton, OH 45232

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
3Air 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-2020-0112

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

13. SUPPLEMENTARY NOTES
88ABW-2020-3287, cleared on 23 October 2020

14. ABSTRACT
High content analysis (HCA) is a useful technique for extracting unprejudiced explanations for phenotypic responses. However, the massive number of features generated -- often exceeding the number of samples -- necessitates an intermediate feature selection step as part of an overall analytic pipeline. While typical feature selection techniques in the literature focus on more modest feature sizes $p < 100$, we found that our large feature sets diminished the feasibility of direct wrapper-based approaches, whereas filter-based approaches produced limited feature complementarity. We ultimately propose two embedded methods for feature selection: one based on an ensemble of feature rankers, and one based on iterative selection. Both methods score well with respect to accuracy and clustering metrics and are readily tunable for hyperparameter searches.

15. SUBJECT TERMS
High content analysis, image analysis, feature selection, phenotype, clustering Check DTIC Thesaurus

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	15	19b. TELEPHONE NUMBER *(Include Area Code)* N/A

TABLE OF CONTENTS

Embedded Techniques for High Content Analysis Feature Selection

Guanpeng Andy Xu, CEE Intern

Mentors: Dr. Heather Pangburn, Dr. Patrick McLendon, Mr. Daniel Cowan

ABSTRACT

High content analysis (HCA) is a useful technique for extracting unprejudiced explanations for phenotypic responses. However, the massive number of features generated -- often exceeding the number of samples -- necessitates an intermediate feature selection step as part of an overall analytic pipeline. While typical feature selection techniques in the literature focus on more modest feature sizes $p < 100$, we found that our large feature sets diminished the feasibility of direct wrapper-based approaches, whereas filter-based approaches produced limited feature complementarity. We ultimately propose two embedded methods for feature selection: one based on an ensemble of feature rankers, and one based on iterative selection. Both methods score well with respect to accuracy and clustering metrics and are readily tunable for hyperparameter searches.

BACKGROUND

Our research group ultimately aims to develop an analytic platform to understand individual susceptibility to various toxicants based on underlying genetic differences. Initially, we collect cells from a variety of genetic donors, and expose them to various compounds at different concentrations. After staining and imaging the samples, we then perform a segmentation step to identify individual cells; we measure around 11000 distinct features for each cell. Lastly, we identify a set of features that best characterize cell responses and associate differences in such responses with their genetic causes. My personal contribution to the project focused on the *feature selection* stage of the pipeline.

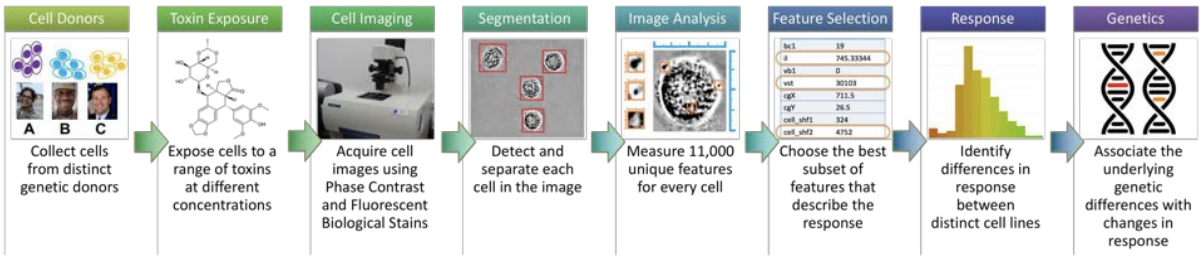

Figure 1: Overall analytic pipeline. My work focused on the feature selection phase.

INTRODUCTION

In the world of data analysis, more is not always better. While adding measurements to a dataset can often provide insight and improve model performance, attempting to model datasets with too many features can give rise to the *curse of dimensionality* (Nasreen 2014). The presence of extraneous features can severely hamper model training, and the increased sparsity of high-dimensional feature spaces is detrimental to many distance-based learners, such as k-means. Lastly, overfitting becomes a prominent concern whenever the dimensionality of the data exceeds the number of training samples. *Feature extraction* approaches, such as principal component analysis, reduce data dimensionality by aggregating features at the cost of reducing the intelligibility of the resulting model -- a feature that adds a mass to a length, after all, has little physical meaning. We instead focus on *feature selection* as a means of mitigating the curse of dimensionality while maintaining the semantic content of our chosen features.

Filter and wrapper methods represent the two predominant paradigms for feature selection. Filtering methods typically examine statistical properties of the dataset to evaluate individual features, whereas wrapper methods use machine learning methods to evaluate feature subsets. Since wrapper methods inherently take into account feature interactions and model performance, they usually achieve superior accuracy. However, their exponentially increasing subset search space means that wrapper methods are usually outpaced by their filter counterparts. This is not usually problematic in many applications of feature selection with smaller dimensionality (Ververidis et al 2008), but does restrict the usage of wrapper methods in general.

Embedded methods offer a compromise between the speed of filters and the accuracy of wrappers. Though embedded methods often use machine learning models in order to identify relevant features, they do not search for optimal subsets as do wrapper methods. Instead, embedded methods are able to select features *in the process* of learning the dataset. These approaches therefore embed the feature selection routine within the training process itself. An

example of an embedded method is LASSO selection, which uses an L1 regularized logistic regression model. A noted tendency of L1 regularization as opposed to L2 'ridge' regression is that certain feature weights are often reduced to zero, thanks to the 'sharp' penalty expression. This makes L1 regression a natural candidate for feature selection, as we may simply select features with nonzero weights. A variation of LASSO regularization, elastic net regularization, incorporates a mixture of L1 and L2 penalties. Elastic net selection generally produces results intermediate between LASSO and ridge penalties, and varying the ratio of the L1 to L2 terms serves as an alternative means to control the feature set size. However, all three L1- and L2-based regularization schemes can be used naturally for *feature ranking*, provided features have been adequately normalized beforehand.

METHODS

Our HCA data comprised several 384 well plates, each of which was analyzed via cell segmentation to produce p ~ 11000 features for each cell; each well contained between n ~ 200 and n ~ 1000 individual cells. Each feature was normalized across the cells to have zero mean and unit variance.

Initially, we considered an approach based on Sequential Floating Forward Selection (SFFS). SFFS is an extension of the simpler Sequential Forward Selection (SFS) and Sequential Backward Selection (SFS) that aims to solve the 'nesting problem', where once selected or eliminated, features cannot be removed or re-introduced later in the search. While SFFS has the potential to identify superior feature subsets compared to SFS and SBS, such an improvement in performance must be balanced by its increased computational burden. SFFS often evaluates markedly more feature subsets than its simpler counterparts: Reunanen observes that for certain feature set sizes on the sonar dataset of the UCI Machine Learning Repository, SFFS evaluates over 300 subsets, whereas SFS evaluates under 50 (Reunanen

2012). Indeed, many of the noted performance gains of SFFS over SFS may be attributed simply to the larger number of feature sets considered under the former wrapper. In any case, our initial wrapper-based scheme did not attempt to execute SFFS over the entire HCA dataset. After normalizing our dataset using a zero-mean, unit-variance scaler, we pre-selected p ~ 30 features using a Fisher-score filter. Lastly, we performed SFFS on this reduced feature set (Figure 2). While SFFS is efficient on this smaller feature set size, this approach remains limited in performance by the effectiveness of our filtering technique.

Figure 2: Proposed wrapper-based feature selection pipeline.

Our approaches to embedded feature selection were based on the LASSO and elastic-net regularization schemes. The increased efficiency of these algorithms enabled us to be far less aggressive in our initial feature filter, selecting p ~ 3000 features instead. To address the instability of LASSO selection (Garauha 2016), we first used an ensemble of elastic net logistic regression models to generate an aggregate ranking of features, as the normalized weights provide a natural index of feature importance. Various methods of constructing such an ensemble exist; we considered a range of models with differing L1 ratios to identify features that performed well both within small and large feature set sizes. Lastly, we selected features from our ranking until the cumulative feature set no longer increased the accuracy (Figure 3.)

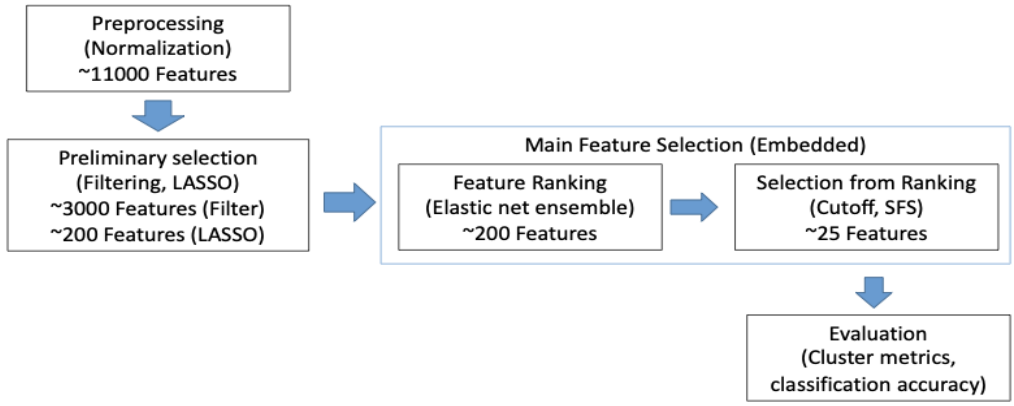

Figure 3: Proposed ensemble feature selection pipeline based on feature ranking.

The post-filtering pre-selection phase reduces the necessity of having to rank all p ~ 3000 features to around p ~ 200. If such a feature subset proves insufficient, the ranking phase may include the remaining p ~ 2800 features behind the other features; we did not find this necessary during testing.

An alternative approach to mitigating instability is iterative selection. After normalization and preselecting features, we repeatedly selected additional features via the LASSO algorithm. In order to avoid 'masking' issues where the presence of certain strong features could discourage the selection of weaker, relevant features, we withheld previously selected features at each additional stage of selection. The number of features selected at each stage, as well as the number of rounds of iterative selection, are very tunable thanks to the parameters of the underlying models. We typically selected p ~ 25 features per round and incorporated a 5 repetition maximum for our testing, although the number of features selected can vary between rounds (Figure 4). This form of repeated selection is potentially vulnerable to some measure of feature redundancy. As such, we implemented a pruning stage where less relevant individual features would be removed so long as doing so would not impact the accuracy of a model trained without them. This stage is not crucial to the selection procedure and omitting it did not significantly degrade the algorithm's performance.

Figure 4: Proposed embedded feature selection pipeline based on iterative selection.

Equally as important as clever algorithms for producing feature sets are effective metrics for evaluating them. We identified two categories of such metrics: accuracy, as trained by some sort of statistical or machine learning model, and cluster quality. Importantly, we do not emphasize an ideal size for feature sets, nor do we explicitly penalize correlated features; Guyon et al note that including significantly correlated features can often improve performance (Guyon et al 2003). For the former category, we simply evaluated the feature set by its ability to enable a machine learning model -- in our case, a support vector machine -- to accurately distinguish between several populations. Classification accuracy tends to reward complementary feature sets that provide more information about the data as a whole, and a low accuracy score may suggest that a feature set is not sufficiently comprehensive.

While many cluster quality metrics exist, we chose to focus on the Davies-Bouldin score and silhouette score for our task (see Petrovic 2006); both indices compare within-group similarity to out-of-group similarity. In contrast to accuracy, cluster quality metrics reward feature sets with strong, relevant features that produce large separations between data in distinct categories. Feature sets with many extraneous features will perform poorly under this metric, even if they succeed in distinguishing populations via classification. By evaluating feature complementarity and feature strength with these two indices, we hoped to gain a multifaceted

understanding of the relative advantages and disadvantages of various feature selection methods.

RESULTS

Initially, we compared the accuracy of existing and proposed feature selection methods on both the Wisconsin breast cancer dataset from the UCI Machine Learning Repository (Dua et al) and control populations between HCA plate datasets (Table 1); we limited the selection to p = 30 features for the HCA dataset, with the top 10 features excluded. We considered our SFFS and ensemble approaches in conjunction with SVM and LASSO selection, as the number of features would be too small to properly evaluate an iterative approach. While our SFFS-based approach was able to identify a highly accurate feature set for the breast cancer data, it did so at the cost of a large number of selected features and significantly increased runtime. Moreover, the performance gains of SFFS disappeared on the HCA dataset, suggesting that this approach would be unsuitable for full-scale feature selection on the entire set of 11000 features. In contrast, the three embedded approaches evaluated here performed efficiently and accurately on both the breast cancer and HCA datasets. These observations encouraged us to discontinue wrapper-based selection in favor of the embedded and iterative approaches.

Table 1: Accuracy comparisons of various feature selection methods.

Selection Method	Feature Set Size (Breast Cancer)	Accuracy (Breast Cancer)	Feature Set Size (HCA)	Accuracy (HCA)
SFFS	18	97.2%	12	97.0%
SVM-RFE	15	97.0%	15	97.7%
L1 (LASSO)	7	96.7%	14	97.2%
Ensemble	6	96.8%	10	97.7%
None	30	96.7%	30	97.9%

Our second performance evaluation focused on the HCA dataset and compared filtering, SVM, LASSO, ensemble, and iterative approaches with respect to accuracy and cluster quality metrics (Table 2). Again, we excluded the top 10 features for more meaningful comparison between methods. While the filter and SVM selectors enable direct control over the number of selected features, each method enables adjusting internal parameters which directly impact the size of the final feature set. Unsurprisingly, the filter method was the fastest and produced the most separated clusters, at the cost of having the worst classification accuracy. The other four methods performed similarly on the accuracy metrics, with the ensemble method performing best; our iterative selection process was most successful on the clustering metrics.

Table 2: Accuracy and cluster quality comparisons of various feature selection methods on our HCA dataset.

Selection Method	Feature Set Size	Davies - Bouldin	Silhouette	Accuracy
Filter	30 (manual)	0.958	0.385	96.4%
SVM	50 (manual)	0.620	0.291	99.1%
L1 (LASSO)	117	0.618	0.297	99.2%
L1 Iterative	135	0.728	0.359	99.0%
Ensemble	27	0.572	0.250	99.3%

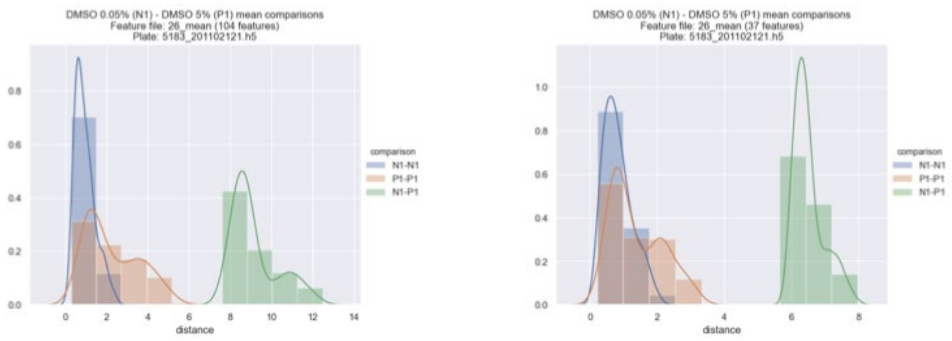

Figure 5. Mean distance comparisons of iterative selection (left) and ensemble selection (right) on the control populations of the HCA dataset. Note that the distances between positive control and negative control well means (green) significantly exceed those between the positive controls (orange) and negative controls (purple).

Also of note are the mean distance comparisons of the feature sets generated by the iterative and ensemble selection methods. Both methods produce significantly higher inter-cluster distances than intra-cluster distances between the control populations (Figure 5). More interesting, however, are the results of mean distance comparisons between compound-treated wells. Although all compound-treated wells showed responses differing from the negative control wells, such responses were not identical. 5-Fluorouracil and nocodazole have highly

distinct cellular mechanisms and exhibit distinct feature profiles, whereas etoposide and oxaliplatin both impact DNA synthesis and do not show a discrepancy between inter- and intra-compound comparisons (Figure 6).

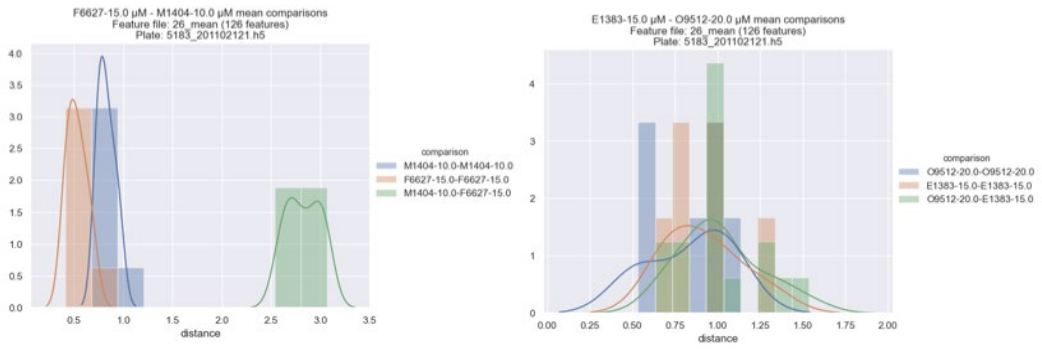

Figure 6: Mean distance comparisons between 5-Fluorouracil and nocodazole (left) and between etoposide and oxaliplatin (right). The former pair show distinct feature profiles, while the latter pair do not.

These results suggest that our feature selection processes may provide insight into the mechanisms of certain compounds; however, more work would be needed to explore such a connection.

CONCLUSIONS

Our research highlighted two embedded approaches for feature selection for high-content analysis. The former approach, ensemble selection, identified smaller, highly complementary feature sets with good accuracy. The latter approach, iterative selection, generated larger sets of strong features with comparable accuracy. Both methods are efficient for our high dimensional datasets and are tunable for hyperparameter searches.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my mentors Dr. Heather Pangburn, Dr. Patrick McLendon, and Mr. Daniel Cowan for their invaluable feedback and guidance throughout my

internship. I also wish to recognize the other members of my research group for their productive discussions and suggestions. Lastly, I would like to thank the Air Force Research Lab, the Department of Defense, and the Center for Educational Excellence for making my experience possible.

REFERENCES

Altidor, W., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2011). Ensemble feature ranking methods for data intensive computing applications. In Handbook of data intensive computing (pp. 349-376). Springer, New York, NY.

Dougherty, E. R., Hua, J., & Sima, C. (2009). Performance of feature selection methods. Current genomics, 10(6), 365–374. https://doi.org/10.2174/138920209789177629

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Gauraha, N. (2016, February). Stability feature selection using cluster representative lasso. In International Conference on Pattern Recognition Applications and Methods (Vol. 2, pp. 381-386). SCITEPRESS.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Nasreen, Shamila. (2014). A Survey Of Feature Selection And Feature Extraction Techniques In Machine Learning, SAI.

Petrovic, S. (2006, October). A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In Proceedings of the 11th Nordic Workshop of Secure IT Systems (pp. 53-64). sn.

Reunanen, J. (2012). Overfitting in feature selection: Pitfalls and solutions.

Ververidis, D., & Kotropoulos, C. (2008). Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. Signal processing, 88(12), 2956-2970.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., & Sugiyama, M. (2014). High-dimensional feature selection by feature-wise kernelized lasso. Neural computation, 26(1), 185-207.

Zanaty, E. A. (2012). Support vector machines (SVMs) versus multilayer perception (MLP) in data classification. Egyptian Informatics Journal, 13(3), 177-183.