2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313269

OptSelect: An algorithm for ensemble feature selection and stability assessment

Eva K Lee, PhD
Center for Operations Research in Medicine and Healthcare and School of Biological Sciences, Georgia Institute of Technology
Atlanta, USA

Karan Uppal, PhD
Center for Operations Research in Medicine and Healthcare and School of Medicine, Emory University Atlanta, USA

***Abstract* — Recent studies have shown that ensemble feature selection approaches can improve the robustness and stability of final classification models. Existing methods for aggregating feature lists from different methods require use of arbitrary thresholds for selecting the top ranked features and are often based on metrics independent of the classification accuracy while selecting the optimal set. In this paper, we develop the OptSelect tool for ensemble feature selection and stability assessment of individual features for improved biomarker discovery. The software tool is packaged in R for broad dis-semination.**

OptSelect is a multi agent-based stochastic optimization tool designed for ensemble feature selection. Stage one involves function perturbation, where ranked list of features is generated using multiple feature selection methods. Stage two in-volves data perturbation, where feature selection is performed within randomly selected learning sets of the training da-ta. The agents are assigned to different behavior states and move according to a binary PSO algorithm. A multi-objective fitness function is used to evaluate the classification accuracy of the agents. We evaluate OptSelect system performance using the random probe method testing on five publicly available microarray datasets. The performance is compared with single feature selection techniques and existing aggregation methods. The results show that OptSelect improves classification accuracy when compared to both individual and existing rank aggregation methods. The PSO algorithm is able to uncover important discriminatory features for predicting COVID-19 disease severity, demonstrating its important role within the optSelect tool. The algorithm is incorporated into an R package and disseminated via GitHub: https://github.com/kuppal2/optSelect.

Keywords* — *ensemble feature selection, agent-based, particle swarm optimization, classification, multi-objective fitness function

Corresponding author: Eva K Lee, evalee-gatech@pm.me

I. Introduction

Robust biomarker discovery is a key component of translational biomedical research and precision medicine [10, 20, 5, 42]. Most omics technologies measure thousands to hundreds of thousands of variables and fall under the category of n<<p problems that are prone to model over-fitting [10, 31, 4, 21]. The large feature space requires application of variable selection techniques to identify most salient variables and generate robust classifiers. This is crucial for targeted validation experiments, designing follow-up studies, diagnostic and treatment in clinical practice, and vaccine design and immunogenicity prediction [19, 26, 15, 25, 28].

Numerous feature selection algorithms have been developed over the last few decades [33, 22, 5]. The feature selection methods can be classified as: filter, wrapper, and embedded [33]. The filter methods use statistical criteria independent of the classifier to select relevant features and the selected features, e.g. p<0.05, are then used to build and train the model. Methods such as t-test, ANOVA, F-test, Chi-sq test, and mutual information are filter methods. The wrapper methods use a search strategy to evaluate different combinations of subsets of features and select the best model based on the evaluation using a classifier such as Support Vector Machine (SVM) [37]. Different search algorithms such as best sub-set, genetic algorithms, particle swarm optimization (PSO), etc. can be used to obtain an optimal set of features; how-ever challenges persist that the resulting feature set is too large and can lead to over-fitting [33, 5]. The embedded methods include methods such as recursive feature elimination based on SVM, random forests (RF), Lasso, Elasticnet where the variable selection is built-in [11, 3, 36, 40, 14, 24].

Recent articles have highlighted the importance of aggregating ranking results from multiple methods to achieve a set of stable features that are likely to be reproducible in future studies [33, 2, 1, 13]. Our vaccine immunogenicity prediction showcases consistency in feature selections is fundamental to clinical translational practice [18, 15]. The translation of basic science findings to new interventions has been limited due to irreproducible results [2, 13, 12]. Two main reasons of “unstable” results are a) data perturbation: inconsistency in selected feature subsets due to sampling variations; b) function perturbation: different rankings of relevance from different methods [2, 13]. Many feature selection approaches use arbitrary rank or significance thresholds to select the number of features without thorough evaluation, which could lead to suboptimal results. Moreover, different algorithms vary differently in performance depending on the distribution of the data and within-class variability [33, 2]. Here we introduce a novel and adaptable nested ensemble feature selection framework, OptSelect, that performs multi-objective optimization using agent-based modeling and binary particle swarm optimization (PSO) techniques to allow aggregation of results from different methods and performs nested feature selection using random subsets of training data to evaluate feature stability [16, 6]. In our own vaccine immunogenicity prediction work, consistency of selected features allows for robust, reproducible, and usable outcomes that can accelerate scientific advances and be translated in clinical practice rapidly [18, 29, 30].

II. Methods and Design

A. OptSelect: algorithm and implementation

A two-stage procedure is used for finding an optimal set of features. In Stage one, a nested feature selection procedure is used to generate ranked list of features using one or more feature selection algorithms selected by the user based on the training set. The training set (S) is randomly split into learning (L_n) and validation (V_n) sets N times using a Monte-Carlo cross-validation procedure such that the proportion of samples from each group is roughly the same in each learning set and validation set [35]. A ranked list of top k features is generated within each learning set, L_i, using one or more feature selection methods: t.test, f.test, recursive feature elimination, random forest, Wilcox Test, lasso, elasticnet [5, 33]. Each variable m is assigned a variable stability score (VSC), where

$$Variable\ stability\ score\ (m) = \frac{Number\ of\ learning\ sets\ in\ which\ a\ variable\ m\ is\ selected}{Total\ number\ of\ learning\ sets} \quad (1)$$

The variables are sorted by VSC. A backward or forward selection procedure is performed using the variables that are selected in at least one learning set using equation (2),

$$\begin{aligned} \text{fitness}_{(\text{Ln, fs, Vn})} = {} & w_1*(CV_{Lm} - CV_{Lm(perm)}) \quad (2) \\ & + w_2*(\text{balanced accuracy } V_n) \\ & + w_3*(\text{balanced accuracy } L_n)\ w_4*(\text{back\_accuracy}) \\ & - w_5(|f_s|) \end{aligned}$$

where w_1 is the weight for difference between average k-fold CV accuracy and average permuted k-fold CV accuracy based on the learning set L_m (default: 0.7); w_2 is the weight for balanced classification accuracy in validation set (V_n) (default: 0.2); w_3 is the weight for balanced classification accuracy in learning set (L_n) (default: 0.05); w_4 is the weight for back accuracy, which is the balanced classification accuracy using the validation set for training and the learning set for evaluation (default: 0.05); and w_5 is the penalty for model complexity determined based on size of the feature set f_s (default: 0.01).

In Stage two, the ranked lists of features from different feature selection methods are aggregated using a newly developed binary behavior-based particle swarm optimization (B3PSO) algorithm, which uses a multi-objective stochastic optimization procedure as described below.

B. Binary Behavior Based Particle Swarm Optimization (B3PSO) algorithm

Agent-based models involve three key elements [23]: 1) agents and their attributes/behaviors: each agent is assigned to a behavior or rule category, e.g. follows neighbors, moves randomly; 2) relationships and interactions among agents: the interactions define how the agents influence and cooperate with each other; 3) interaction with the environment: the environment provides feedback to the agents about their movements (e.g. 10-fold cross-validation accuracy). PSO, first introduced by James Kennedy and Russell Eberhart in 1995 [16], is a stochastic optimization technique based on the movement and intelligence of swarms. It comprises of a number of agents/particles that constitute a swarm moving around in the search space looking for the best solution determine based on a fitness evaluation function. The movement of each particle, p_i, is determined based on a velocity vector, v_i, and a position vector, x_i. In binary PSO, the position vector, x_i, has d dimensions, where d corresponds to the number of variables (genes, chemicals, etc.). And, $x_{id}=\{0,1\}$. Thus, a feature is selected if $x_{id}=1$ and the positive vector represents which features are selected. The velocity and position vectors are updated according to equations (p1) and (p2).

At iteration t+1 , the velocity of each particle, p_i, is updated according to the equation,

$$v_i^{t+1}=v_i^t+c_1*r_1*(pbest_i-s_i^t)+c_2*r_2*(gbest_i-x_i^t) \quad (p1)$$

where i is the current particle; c_1 and c_2 are constant learning factors to control the social influence and global influence respectively; r_1 and r_2 are random numbers between [0,1] interval; pbest is the best position of the particle has experienced based on the fitness function; gbest is the best position experienced by any particle in the swarm based on the fitness function; and x_i^t is the position in iteration t.

The velocity of the particle is restricted to be in the interval [-6,6]. The position is updated according to Equations (p2) and (p3) based on a sigmoid transformation, S, of its velocity (Figure 1).

For d in 1 to number of variables (total number of features):

$$S_{id} = 1/(1+\exp(-v_{id}^{t+1})) \quad (p2)$$

$$x_{id}^t = 1 \text{ if } S > r_{3,}\ 0 \text{ otherwise} \quad (p3)$$

where x_{id}^t is the position of the i^{th} agent at time t in dimension d (gene, chemical, etc.); v_{id}^{t+1} is the updated velocity of the i^{th} agent at iteration t+1 in dimension d; S_{id} is the sigmoid function with values between [0,1] interval for dimension d; r_3 is a random number between [0,1] interval.

In our implementation, users can provide a weight vector to bias the selection process based on expert knowledge or from literature. The random number in Equation (p3) is replaced by the weight of the feature [0 to 1] to bias the selection process.

Studies have shown that incorporating prior knowledge can improve the classification accuracy [13]. On the other hand, our own vaccine immunogenicity prediction showcases discovery of potential new knowledge when we search the features broadly in an unbiased manner [18,29,30].

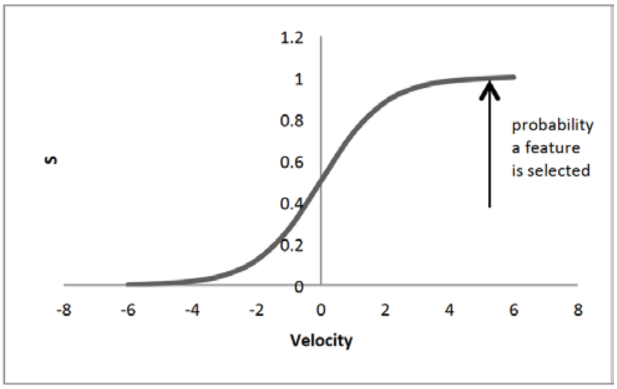


Figure 1. Relationship between velocity, sigmoid function, and probability of a feature being selected. The likelihood of a feature being selected increases as the velocity approaches 6 and S approaches 1.

It is common among most of the implemented binary PSO algorithms that all the particles are designed to behave uniformly (Chuang 2008). In our B3PSO algorithm, each particle behaves independently, with its position and velocity at each iteration determined based on its behavior at the immediate previous iteration. Specifically, each particle is assigned to one of the four behavioral states {C=Confusion, S=Self-influenced, N=Influenced by nearest neighbors, G=Influenced by swarm} based on the crowd model [38]. The velocities are updated as follows:

$$v_i^{t+1}=v_i^t+c_1*r_1*(pbest_i-s_i^t)+c_2*r_2*(choice_i-x_i^t) \qquad (p4)$$

where $choice_i$ is dictated by the particle's behavioral states:

- *C moves randomly*: $choice_i = rpos_i$, a random position vector
- *S only self-influenced:* $choice_i = 0$
- *N follows neighbors:* $choice_i = nbest_i$, the centroid (75th percentile) of the k nearest neighbors (default k=3)
- *G follows global leader:* $choice_i = gbest_i$, the global best position in the swarm

The behaviors are updated on regular intervals to prevent the population from getting stuck in local optima (user-defined parameter; default: 5 iterations). The fitness of each particle is evaluated using a nested cross-validation procedure and support vector machine classifier as shown in Figure 2. The search process is terminated when the distance between the centroid of the entire population and the global best agent is less than or equal to 2. In addition, users can provide a weight vector to bias the selection process based on expert knowledge or from literature. Studies have shown that incorporating prior knowledge can improve the classification accuracy [13]. The random number in Equation (p3) is replaced by the weight of the feature to bias the selection process.

We have developed numerous exact combinatorial feature selection algorithms and PSO heuristics algorithms in clinical translational and hospital applications with good results [18, 8, 10]. In general, PSO is fast and users can fine-tune different parameters to tailor to their applications.

C. Evaluation experiments

Two experiments were performed to evaluate the performance.

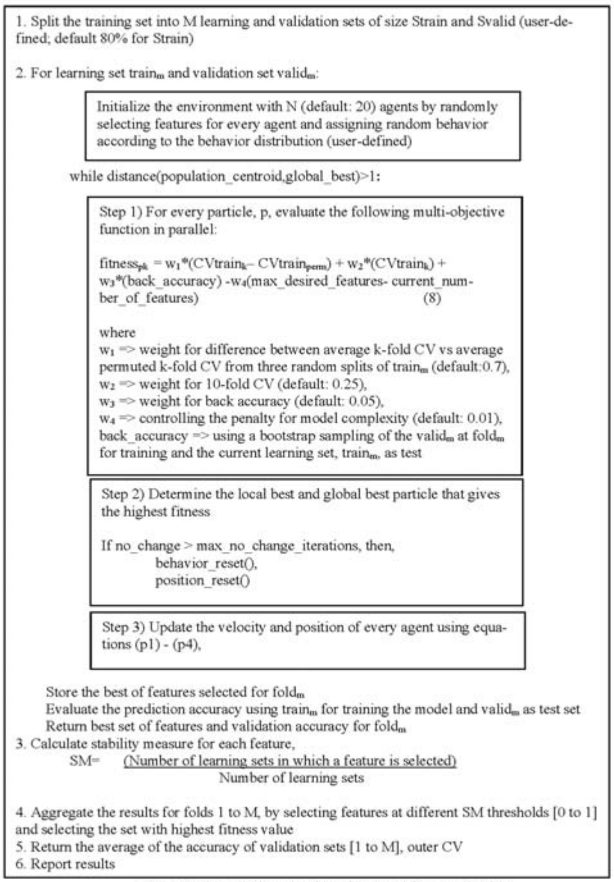
Experiment 1: We include five publicly available microarray datasets in this analysis(Table 1): Leukemia (7129 genes, train samples=38, test samples=34, 2 groups; [7]), SRBCT (2308 genes, train samples=63, test samples=25, 4 groups; [17]), Prostate cancer (6033 genes, train samples=61, test samples=41, 2 groups; [34]), MAQCII-ER and MAQCII-PCR (22,284 genes, train samples=130, test samples=100, 2 groups; [27]).

Experiment 2: A random probe evaluation was performed using the Iris dataset (Fisher 1950) to evaluate the ability of theStage two of the optSelect algorithm to identify real features. The original dataset has 4 real features and 150 instances belonging to 3 groups of the iris plant. Each real feature was scale normalized ($\mu = 0, \sigma = 1$) and 36 probe features with similar distribution were included in the dataset. The instances were divided into 60% train (90; 30 per class) and 40% test (60; 20 per class). Unlike the original iris data, the mixed data has proven to be difficult to classfy.

Table 1. Data information for Experiment 1

Dataset	Number of features	Training set	Test set	Number of groups
Leukemia	7,129	38	34	2
SRBCT	2,308	63	25	4
Prostate cancer	6,033	61	41	2
MAQCII-ER	22,284	130	100	2
MAQCII-PCR	22,284	130	100	2

Five commonly used methods including Limma, rfe-SVM, Lasso, Elasticnet, F-test implemented in the CMA R package were used for feature selection in Stage one. Two rank aggregation methods implemented in the rankAggreg R package, rankAggreg-Monte Carlo and rankAggreg-GA, that optimize the list of ranked features based on the Spearman's footrule were used for comparison in Stage two. The evaluation process was repeated using top 5 and top 15 features from Stage one. The final performance of different methods is evaluated using an independent blinded test set that was not seen by the optSelect algorithm during the model building phase. The balanced accuracy, average group-wise prediction accuracy, was used for evaluation purpose [7].



```
1. Split the training set into M learning and validation sets of size Strain and Svalid (user-defined; default 80% for Strain)
2. For learning set $train_m$ and validation set $valid_m$:
        Initialize the environment with N (default: 20) agents by randomly
        selecting features for every agent and assigning random behavior
        according to the behavior distribution (user-defined)
    while distance(population_centroid,global_best)>1:
        Step 1) For every particle, p, evaluate the following multi-objective
        function in parallel:
        $fitness_{pk} = w_1*(CVtrain_k - CVtrain_{perm}) + w_2*(CVtrain_k) + w_3*(back\_accuracy) - w_4(max\_desired\_features - current\_number\_of\_features)$    (8)
        where
        $w_1$ => weight for difference between average k-fold CV vs average
        permuted k-fold CV from three random splits of $train_m$ (default:0.7),
        $w_2$ => weight for 10-fold CV (default: 0.25),
        $w_3$ => weight for back accuracy (default: 0.05),
        $w_4$ => controlling the penalty for model complexity (default: 0.01),
        back_accuracy => using a bootstrap sampling of the $valid_m$ at $fold_m$
        for training and the current learning set, $train_m$, as test
        Step 2) Determine the local best and global best particle that gives
        the highest fitness
        If no_change > max_no_change_iterations, then,
                behavior_reset(),
                position_reset()
        Step 3) Update the velocity and position of every agent using equations (p1) - (p4),
    Store the best of features selected for $fold_m$
    Evaluate the prediction accuracy using $train_m$ for training the model and $valid_m$ as test set
    Return best set of features and validation accuracy for $fold_m$
3. Calculate stability measure for each feature,
        $SM = \frac{(\text{Number of learning sets in which a feature is selected})}{\text{Number of learning sets}}$
4. Aggregate the results for folds 1 to M, by selecting features at different SM thresholds [0 to 1] and selecting the set with highest fitness value
5. Return the average of the accuracy of validation sets [1 to M], outer CV
6. Report results
```

Figure 2. The optSelect algorithm for nested feature selection

COVID-19 Study: In addition, we also analyze the current COVID-19 severity trend among different cities using our PSO implementation to gauge its practical performance. In Lee 2020-COVID [41], we analyze COVID-19 severity across different nations and regions. Understand that various factors including different clinical practice may render different outcome, we attempt to focus on general population information including age distribution, demographics, socio-economic determinants, living conditions, population health status, hospital resources, poverty level, and blood types across 108 countries and 72 regions. This provides a rich feature set of 75 attributes. The goal is to identify discriminatory set of features that can predict severity using mortality per capita as the measured metric.

III. Results

A. Experiment 1

Tables 2 and 3 summarize the balanced accuracy results for the five datasets in Experiment 1 using the top 5 and top 15 ranked features selected by the five feature selection methods in Stage one followed by the two aggregation methods in Stage two. As discussed earlier, the arbitrary thresholds for selecting top ranked features does not guarantee reproducibility of results on the blinded test set. For instance, by increasing the number of selected features from 5 to 15. the performance of rfe-SVM and Elasticnet degrades by 28% for the SRBCT dataset. On the contrary, almost all methods showed improvements in accuracy for the Leukemia dataset by using greater number of features. Overall, the aggregation stage improved the classification accuracy with optSelect performing comparably or better in almost all cases.

Table 2. The balanced accuracies for the independent blind test sets, predicted using the top 5 ranked features selected in Stage 1.

Dataset / Method	Leukemia	SRBCT	MaqcII_ER	MaqcII_PCR	Prostate
Stage 1					
Limma	0.89	0.94	0.87	0.7	0.77
rfe-SVM	0.94	0.92	0.88	0.66	0.77
Lasso	0.94	0.9	0.87	0.62	0.75
Elasticnet	0.89	0.94	0.87	0.72	0.77
F.test	0.93	0.94	0.87	0.66	0.77
Stage 2 (Aggregation)					
rankAggreg -monte carlo	0.89	0.92	0.87	0.74	0.77
rankAggreg -GA	0.89	0.94	0.87	0.64	0.77
optSelect	**0.96**	**1**	**0.87**	**0.72**	**0.77**

Table 3. The balanced accuracies for the independent blind test sets, predicted using the top 15 ranked features selected in Stage 1.

Dataset / Method	Leukemia	SRBCT	MaqcII_ER	MaqcII_PCR	Prostate
Stage 1					
Limma	0.96	0.99	0.88	0.71	0.77
rfe-SVM	0.96	0.661	0.86	0.68	0.77
Lasso	0.96	1	0.87	0.64	0.77
Elasticnet	0.96	0.66	0.88	0.68	0.77
F.test	0.96	0.986	0.88	0.66	0.77
Stage 2 (Aggregation)					
rankAggreg -monte carlo	0.94	1	0.88	0.75	0.77
rankAggreg -GA	0.96	0.96	0.87	0.73	0.77
optSelect	**0.96**	**1**	**0.9**	**0.73**	**0.83**

B. Experiment 2

For Experiment 2, 3 out of 4 real features were selected in 8 out of 10 folds. 2 features were selected in all 10 folds. OptSelect discarded all probe (randomly generated). Both train

10-fold cross-validation accuracy and balanced accuracy for the test set were 95%.

C. COVID-19 Study

We briefly include one finding here. K-mean clustering was first performed which classified the 72 regions into 3 groups based on mortality per capita. Specifically, 64.8% are placed in Group Low, 18.5% in Group Medium, and 16.6% iin Group High. PSO was applied on the training set and the resulting classification rule was used to blind predict the remaining regions.

Table 4 shows the classification results. PSO selected seven features including percentage of population age 60-69, percentage of population age 80+, percentage of population of B− blood type, smoking rate, obesity rate, population density (sq mi), and poverty level. For example, New York City is in Group 3 while Miami is in Group 2. These results align with the fact that minority population is disproportionally affected by COVD-19 and suffers higher mortality rate as a result of their disadvantaged socio-economic determinants (poverty, dense dwelling, and poor health), while nursing home falls into the age group features. This illustrates the diverse features that PSO can uncover. It also showcases that along with our in-house discrete support vector machine (DAMIP) [8,18,19,20,25,28], PSO-DAMIP cam handle imbalance data better than these other methods.

Table 4. Classification results for 3-group COVID-19 severity prediction using PSO versus the five methods

	Training (50 regions) – 10-fold cross-validation results			**Blind Prediction Set (22 regions)**		
	Low	**Medium**	**High**	**Low**	**Medium**	**High**
limma	0.95	0.45	0.35	0.81	0.41	0.30
rfe-SVM	0.88	0.67	0.30	0.85	0.54	0.21
Lasso	0.91	0.49	0.51	0.80	0.35	0.45
Elasticnet	0.89	0.61	0.47	0.87	0.55	0.39
F.test	0.80	0.54	0.56	0.83	0.48	0.49
PSO-DAMIP[18]	**0.85**	**0.90**	**0.95**	**0.83**	**0.91**	**0.92**

IV. Discussion

Machine learning is increasingly being used to assist in biomedical knowledge discovery and decision making. However, feature instability and irreproducibility have persistently hindered the translation of results from basic science to clinical practice [2, 1, 13, 12]. Several methods for rank aggregation have been proposed [33]. However, most of these methods use criteria independent of the classification accuracy for aggregating the results and require selection of an arbitrary threshold for top features to determine the overlap. This could result in degradation of classification performance on an independent or new unseen data set.

In this paper we design a novel optimization based nested ensemble feature selection framework, optSelect, to address this challenge by using a two-stage approach for aggregating the results. Stage one involves selection of top ranked features using different methods. The top list of features obtained is then merged and used as input for a multi agent-based optimization procedure to find the most stable set of features with good classification accuracy. The "optimal" set is determined using a multi-objective fitness function. The algorithm incorporates the concepts of function perturbation and data perturbation to select the most optimal and stable set of features. Additionally, the algorithm is de-signed to prevent agents from getting stuck in local optima. Each agent interacts with other agents (in a heterogenous manner) based on its behavior state {confusion, follows neighbors, follows global leader, self-influenced}. The search for optimal solution terminates when the current global best and the centroid of the population converge, which is determined using the Euclidean distance.

Evaluation results for the five gene expression datasets (Experiment 1) show that optSelect allows aggregation of results from different methods without compromising on the classification accuracy. The results also highlight the dramatic changes in classification accuracies as a result of arbitrary thresholds for selecting top features. Furthermore, the performance of different independent feature selection techniques varies across different datasets. On the contrary, optSelect algorithm performed consistently well across all datasets and improved the classification results on independent test sets in most cases. The output includes outer CV estimates, optimal set of features, and stability measures for each feature. Figure 3 shows the stability measures of the 6 features from an optimal set selected by optSelect for the MAQCII-ER dataset, where the samples were classified as ER+ve vs ER-ve. The most stable feature that was reproducibly selected in all folds in the nested feature selection was estrogen receptor 1; whereas the stability measures for the features returned by the two aggregated approaches hover under 0.55.

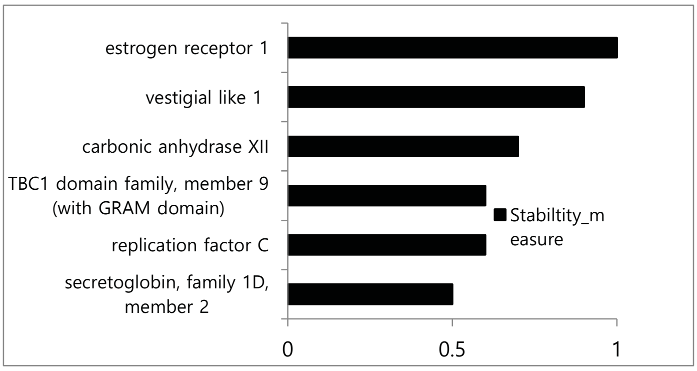


Figure 3. Stability measures for features in the optimal set selected by optSelect.

The performance evaluation of the multi agent-based optimization procedure on the random probe experiment (Experiment 2) shows that the behavior based search algorithm combined with the multi-objective classification based fitness function can detect relevant features even when majority of the features are randomly generated (noise).

Applying our implementation of PSO on recent COVID-19 severity prediction analysis illustrates that PSO can uncover critical features, along with an appropriate classifier, can handle imbalance data very well and offer robust classification outcome [41, 18, 8, 25, 28]

We emphasize that in practice, features selected constitute set of policies, clinical guidelines, rules to follow, or specific biological implications [43,8,18,25,28]. As such, it is important to identify the smallest possible feature set that has good predictive power for practical purpose. Hence, we report only top features ranging from 5-15 only.

Limitations and future work: We acknowledge that users should select different feature selection methods for Stage 1, and other aggregated approaches for Stage 2. OptSelect is designed in a flexible way and users can select the methods they prefer and perform comparisons. The current study did not assess the effect of using data normalization methods and classifiers on classification accuracy. Escalente et al. have showed application of PSO for full model selection (pre-processing methods, feature selection, and classification algorithms). We have also implemented extensively variations of PSO and exact combinatorial algorithms for feature selections within machine learning and artificial intelligence convolutional neural network framework [8, 18, 19]. Future work will focus on extending the current framework to full model selection. Computational time for each stage and components will also be analyzed,

Acknowledgement

This work is partially supported by grants from the National Science Foundation, IIP-0832390 and IIP-1361532. Find-ings and conclusions in this paper are those of the authors and do not necessarily reflect the views of the National Sci-ence Foundation. We acknowledge funding from Emory University in support of Karan Uppal while he was a full-time PhD student at Georgia Institute of Technology, supervised under Lee and worked on this project.

AReferences

1. Abeel T, Helleputte T, Peer YVD, Dupont P, Saeys Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 26(3), 392-398. http://dx.doi.org/10.1093/bioinformatics/btp630
2. Boulesteix AL, Slawski M. Stability and aggregation of ranked gene lists. Brief Bioinform. 2009 Sep;10(5):556-68. doi: 10.1093/bib/bbp034. Review.
3. Breiman, L. (2001). Random Forests. Mach Learn, 45(1), 5-32. http://dx.doi.org/10.1023/A:1010933404324
4. Cawley GC and Talbot NLC, Over-fitting in model selection and subsequent selection bias in performance evaluation, Journal of Machine Learning Research, 2010. Research, vol. 11, pp. 2079-2107, July 2010.
5. Christin C, Hoefsloot HC, Smilde AK, Hoekman B, Suits F, Bischoff R, Horvatovich P. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. Mol Cell Proteomics. 2013 Jan;12(1):263-76. doi: 10.1074/mcp.M112.022566. Epub 2012 Oct 31.
6. Chuang LY, Chang HW, Tu CJ, Yang CH. Improved binary PSO for feature selection using gene expression data. Comput Biol Chem. 2008 Feb;32(1):29-37. Epub 2007 Sep 25.
7. Dreyfus G and Guyon I. Chapter 2: Assessment Methods. Feature Extraction. Studies in Fuzziness and Soft Computing. Vol 207. 2006
8. Eva K Lee, Haozheng Tian, Jinha Lee, Xin Wie, John Neeld Jr, K Doug Smith, Alan R Kaplan. Investigating a Needle-Based Epidural Procedure in Obstetric Anesthesia. AMIA Annu Symp Proc. 2018; 2018: 720–729.
9. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by gene expression profiling   Science 15 October 1999: Vol. 286. No. 5439, pp. 531 - 537
10. Guyon et al. An introduction to variable and feature selection Journal of Machine Learning Research 2003 (3): 1157-1182
11. Guyon I, Weston J, Barnhill S, and Vapnik V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn, 46(1-3), 389-422. http://dx.doi.org/10.1023/A:1012487302797.
12. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015 Mar;12(3):179-85. doi: 10.1038/nmeth.3288.
13. He Z, Yu W. Stable feature selection for biomarker discovery. Comput Biol Chem. 2010 Aug;34(4):215-25. doi: 10.1016/j.compbiolchem.2010.07.002. Epub 2010 Aug 10. Review.
14. Jović* A, K. Brkić* and N. Bogunović. A review of feature selection methods with applications
15. Kazmin D, HI Nakaya, EK Lee, et al. Systems analysis of protective immune responses to RTS,S malaria vaccination in humans. Proceedings of the National Academy of Science.. 2017; 114(9) 2425–2430.
16. Kennedy J. and Eberhart R. (1995). "Particle Swarm Optimization". Proceedings of IEEE International Conference on Neural Networks IV. pp. 1942–1948. doi:10.1109/ICNN.1995.488968
17. Khan et al. Classification and diagnostic prediction of cancers using gene expression profiling and artficial neural networks Nature Medicine June 2001: Volume 7 no. 6, pp 673-679
18. Lee EK, H Nakaya, F Yuan, T Querec, F Pietz, BA Benecke, G Burel, B Pulendra. Machine learning framework for predicting vaccine immunogenicity. Interfaces – The Daniel H. Wagner Prize for Excellence in Operations Research Practice (Winner). 2016; 46(5): 368 - 390.
19. Lee EK, HZ Tian, HI Nakaya. Antigenicity prediction and vaccine recommendation of human influenza virus A using convolutional neural networks. Human Vaccines & Immunotherapeutics. 2020:1-19.
20. Lee EK. Large-scale optimization-based classification models in medicine and biology. Ann Biomed Eng 35: 1095–1109, 2007
21. Lever, J., Krzywinski, M. & Altman, N. Model selection and overfitting. Nat Methods 13, 703–704 (2016). https://doi.org/10.1038/nmeth.3968
22. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. Brief Bioinform. 2008 Sep;9(5):392-403. doi: 10.1093/bib/bbn027. Epub 2008 Jun 18.
23. Macal CM and North MJ. Tutorial on agent-based modeling and simulation. Journal of Simulation. 2010. (4). 151-162.
24. Masoudi-Sobhanzadeh, Y., Motieghader, H. & Masoudi-Nejad, A. FeatureSelect: a software for feature selection based on machine learning approaches. BMC Bioinformatics 20, 170 (2019). https://doi.org/10.1186/s12859-019-2754-0
25. Nakaya HI, T Hagan, M Kwissad, EK Lee, et al. Systems analysis of immunity to influenza vaccination across multiple years reveals universal predictors of immunity in the young &elderly. Immunology 2015,43(6):1186-98.
26. P Gonzalez-Dias, EK Lee, S Sorgi, DS de Lima, AH Urbanski, ELV Silveira, HI Nakaya. Methods for predicting vaccine immunogenicity and Reactogenicity. Human Vaccines & Immunotherapeutics. 2020; 16 (2): 269–276.
27. Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, Nikolsky Y, Tsyganova M, Ishkin A, Nikolskaya T, Hess KR, Valero V, Booser D, Delorenzi M, Hortobagyi GN, Shi L, Symmans WF, Pusztai L. Effect of training-sample size and classification difficulty on the accuracy of genomic

predictors. Breast Cancer Res. 2010;12(1):R5. doi: 10.1186/bcr2468. Epub 2010 Jan 11.

28. Querec TD, R Akondy, EK Lee, et al. Systems biology approach predicts the immunogenicity of the yellow fever vaccine in humans. Nature Immunology 2009; 10:116-125. PMCID: PMC4049462

29. Ravindran R, Khan N, Nakaya HI, Li S, Loebbermann J, Maddur MS, Park Y, et al. (2014) Vaccine activation of the nutrient sensor GCN2 in dendritic cells enhances antigen presentation. Science 343(6168):313–317.

30. Ravindran R, Loebbermann J, Nakaya HI, Khan N, Ma H, Gama L, Machiah DK, et al. (2016) The amino acid sensor GCN2 controls gut inflammation by inhibiting inflammasome activation. Nature 531(7595):523–527.

31. Reunanen J. Overfitting in Making Comparisons Between Variable Selection Methods. Journal of Machine Learning Research 2003 (3): 1371-1382

32. Saeys Y, Abeel T, Peer YVD (2008). Robust Feature Selection Using Ensemble Feature Selection Techniques. In: Machine Learning and Knowledge Discovery in Databases (pp. 313-325). Springer Berlin Heidelberg. Retrieved August 23, 2013, from http://link.springer.com/chapter/10.1007/978-3-540-87481-2_21

33. Saeys Y, Inza I, Larrañaga P, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507.

34. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002;1(2):203–209.

35. Slawski et al. Classification for Microarrays – a comprehensive Bioconductor package for supervised classification with high dimensional data BMC Bioinformatics 2008 Vol. 9 pp439

36. Tibshirani R. Regression shrinkage and selection via the Lasso. J. Roy Stat Soc B 1996;58:267–88.

37. Vapnik, V. (1998). The Support Vector Method of Function Estimation. In Nonlinear Modeling (pp. 55-85). Springer US.. Retrieved January 2, 2014, from http://link.springer.com/chapter/10.1007/978-1-4615-5703- 6_3

38. Wu S, Sun Q. Computer simulation of leadership, consensus decision making and collective behaviour in humans. PLoS One. 2014 Jan 17;9(1):e80680. doi: 10.1371/journal.pone.0080680. eCollection 2014.

39. Yvan Saeys, Iñaki Inza, Pedro Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics, Volume 23, Issue 19, 1 October 2007, Pages 2507–2517, https://doi.org/10.1093/bioinformatics/btm344

40. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society Series B, 67(2), 301-320.

41. Lee, EK, ZH Li, H Prausnitz-Weinbaum. Predicting COVID-19 Severity Across Different Populations. 2020, Submitted.

42. Lever, J., Krzywinski, M. & Altman, N. Model selection and overfitting. Nat Methods 13, 703–704 (2016). https://doi.org/10.1038/nmeth.3968MJ

43. Wolf, EK Lee, SC Nicolson, GD Pearson, MK Witte, J Huckaby, M Gaies, LS Shekerdemian, WT Mahle. Rationale and methodology of a collaborative learning project in congenital cardiac care. American Heart Journal. 2016; 174:129-137

```
1. Split the training set into M learning and validation sets of size Strain and Svalid (user-defined; default 80% for Strain)

2. For learning set $train_m$ and validation set $valid_m$:

        Initialize the environment with N (default: 20) agents by randomly
        selecting features for every agent and assigning random behavior
        according to the behavior distribution (user-defined)

    while distance(population_centroid,global_best)>1:

        Step 1) For every particle, p, evaluate the following multi-objective
        function in parallel:

        $fitness_{pk} = w_1*(CVtrain_k – CVtrain_{perm}) + w_2*(CVtrain_k) + w_3*(back\_accuracy) - w_4(max\_desired\_features - current\_number\_of\_features)$     (8)

        where
        $w_1$ => weight for difference between average k-fold CV vs average
        permuted k-fold CV from three random splits of $train_m$ (default:0.7),
        $w_2$ => weight for 10-fold CV (default: 0.25),
        $w_3$ => weight for back accuracy (default: 0.05),
        $w_4$ => controlling the penalty for model complexity (default: 0.01),
        back_accuracy => using a bootstrap sampling of the $valid_m$ at $fold_m$
        for training and the current learning set, $train_m$, as test

        Step 2) Determine the local best and global best particle that gives
        the highest fitness

        If no_change > max_no_change_iterations, then,
                behavior_reset(),
                position_reset()

        Step 3) Update the velocity and position of every agent using equations (p1) - (p4),

    Store the best of features selected for $fold_m$
    Evaluate the prediction accuracy using $train_m$ for training the model and $valid_m$ as test set
    Return best set of features and validation accuracy for $fold_m$
3. Calculate stability measure for each feature,
        SM=   (Number of learning sets in which a feature is selected)
                          Number of learning sets

4. Aggregate the results for folds 1 to M, by selecting features at different SM thresholds [0 to 1] and selecting the set with highest fitness value
5. Return the average of the accuracy of validation sets [1 to M], outer CV
6. Report results
```

Figure 2. The optSelect algorithm for nested feature selection