Feature Selection for Cancer Classification using Reduced Variable Neighborhood Search

Angelos Pentelas

University of Macedonia, School of Information Sciences,
Department of Applied Informatics, Egnatia Street 156, 54636, Thessaloniki

25 June, 2019

Agenda

- Problem Statement
- Theoretical Background
- Methodology
- Experimental Evaluation
- Related Work and Comparison
- Conclusion and Future Directions

The Feature Selection (FS) problem

Goal: Find the minimally-sized feature subset that is necessary and sufficient to the target concept, e.g., classification, clustering

Class: NP-hard, combinatorial problem with $O(2^d)$ complexity, where d is the number of features

Gains:

- model simplification
- shorter computing time
- curse of dimensionality avoidance
- enhanced generalization by overfit avoidance

Feature Selection techniques (1/2)

Traditional FS techniques:

- Filter
- Wrapper
- Embedded

	Comp. time	Exploration	Accuracy	Subset size	Overfit risk
Filter	✓	✗	✗	✗	✓
Wrapper	✗	✓	✓	✓	✗
Embedded	✗	✗	✓	✓	✗

Figure: The pros and cons of traditional FS techniques.

Feature Selection techniques (2/2)

Emerging FS techniques:

- Ensemble: The final subset is obtained by specifically handling a multitude of subsets yielded from single FS methods, e.g., union, weighted, etc.
- Hybrid: Comprises of multi-level search phases, performed by the same or different FS methods

Reduced Variable Neighborhood Search (RVNS)

- Benefits from systematic changes in different neighborhood structures
- Avoids local search $\rightarrow$ computationally lighter than basic VNS
- Suitable for large problem instances

Algorithm 1 : RVNS pseudocode for a minimization problem

```
initialize solution x;
while stopping criteria are not met do
    k = 1;
    while k < k_max do
        generate x' a random solution from neighborhood N_k(x);
        if evalutate(x') < evaluate(x) then
            x = x';
            k = 1;
        else
            k = k + 1;
        end
    end
end
return x;
```

Figure: RVNS pseudocode for a minimization problem.

Solution Representation and Neighborhood Definitions

Solution representation:

- Binary, one-dimensional, d-length array $\rightarrow$ [0,1,1,0,...,1,0]

Neighborhood structures:

- Replace a selected feature with an un-selected one
- Replace two selected features with an un-selected one
- Add an un-selected feature

... in this order.

Fitness Function

A solution's quality is determined by:

$$F(s) = \alpha * \underbrace{\frac{3}{a_1(s) + a_2(s) + a_3(s)}}_{\text{a(s)}} + (1 - \alpha) * g(s)$$

where:

- $a_i(s) \rightarrow$ accuracy from classifier$_i$, $i \in \{1, 2, 3\}$
- $g(s) \rightarrow$ number of selected features, i.e., ones in s
- $\alpha \rightarrow$ weight coefficient

Optimization criterion: Minimization

Motivation

- Support Vector Machine (SVM), k Nearest Neighbors (kNN) and Random Forest (RF)
- Different-learning $\rightarrow$ more reliable results

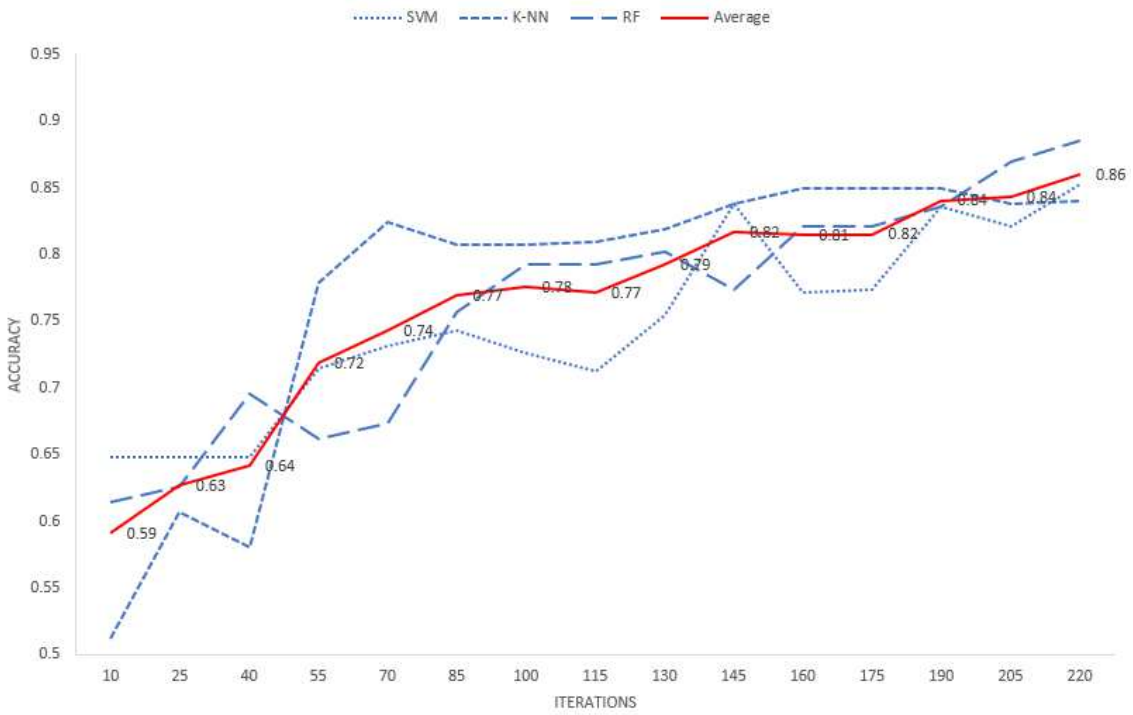

Figure: The classifiers are learning in a different manner, e.g., iterations 40 and 145.

Towards Hybridization - RFE

Recursive Feature Elimination (RFE):

- Embedded FS technique
- Heuristically ranks feature importance
- Ranking via a model's attributes, e.g., SVM's coefficients
- Eliminates a certain % of *unimportant* features at each iteration
- Terminates when the specified % of features is reached

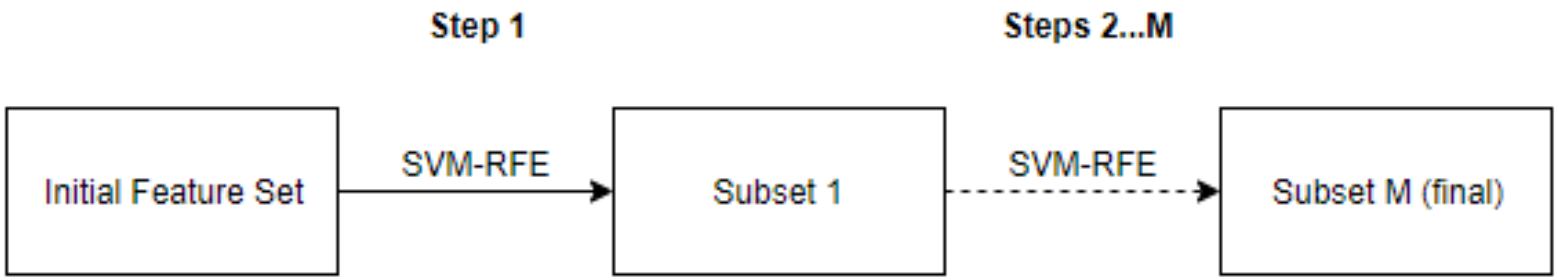

Figure: The SMV-RFE FS scheme.

Hybrid RFE-RVNS

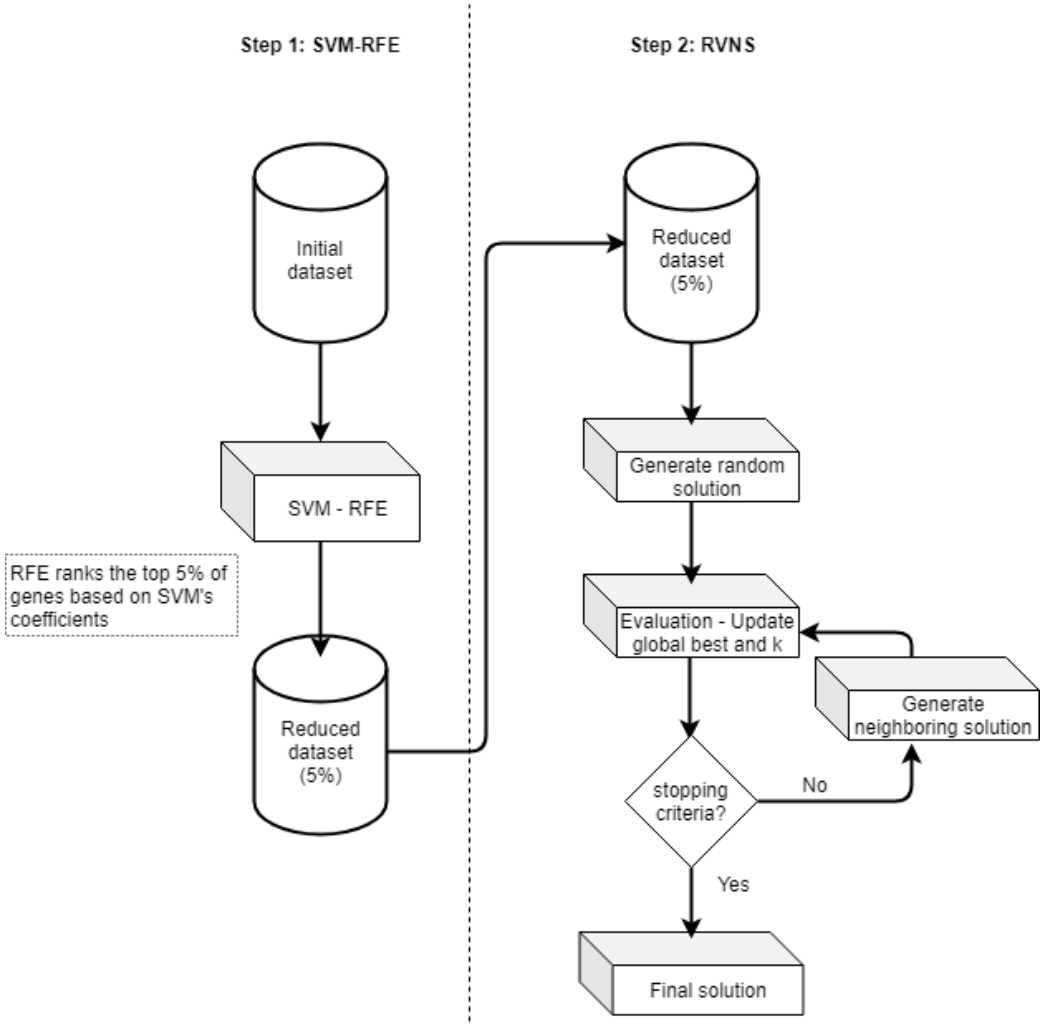

Figure: The hybrid RFE-RVNS FS scheme.

Cancer-Related Datasets

Five cancer-related datasets (microarray, mass spectroscopy):

- Small sample size
- High dimensional
- Features $\rightarrow$ Genes

Dataset	Sample size	Number of genes	Number of classes
Leukemia	72	7,129	2
Lung	181	12,533	2
Ovarian	253	15,154	2
Colon	62	2,000	2
Breast	590	17,814	2

Figure: Dataset characteristics.

Sources: `http://leo.ugr.es/elvira/DBCRepository/`, `https://data.mendeley.com/datasets/v3cc2p38hb/1`

Parameter settings

Experimentally selected parameters:

- Initial solution: random, 2 arbitrarily selected genes
- RVNS: 300 iterations, neighborhoods as defined
- Fitness function: α equals to 0.99
- SVM[1]: Linear kernel
- kNN[1]: k equals to 5
- RF[1]: 20 10-depth decision trees
- SVM-RFE: 5% of top ranked genes, 10% elimination step

[1] Evaluates model accuracy with 10-fold cross validation

Performance (1/2)

Worst performance observed on the Colon dataset:

	RFE-RVNS			RVNS		
Metric	Best	Mean	Worst	Best	Mean	Worst
Average accuracy	93.73	91.23	88.57	89.68	87.72	85.24
SVM accuracy	96.90	93.36	88.57	93.33	89.50	84.05
k-NN accuracy	96.90	92.69	89.05	90.00	87.79	84.05
RF accuracy	90.24	87.64	84.05	88.57	85.88	80.48
#Genes	4	5.5	8	3	5.5	8

Figure: Performance on the Colon dataset after 10 independent runs.

Performance (2/2)

	RFE-RVNS			RVNS		
Metric	Best	Mean	Worst	Best	Mean	Worst
Average accuracy	99.58	98.88	97.64	97.44	94.26	89.35
SVM accuracy	100	98.75	94.58	100	95.18	86.67
k-NN accuracy	100	99.58	97.08	97.5	93.84	87.56
RF accuracy	100	98.32	97.08	97.5	93.76	87.2
#Genes	3	3.8	5	2	4.7	8

Figure: Performance on the Leukemia dataset after 10 independent runs.

... even better performance on the rest of the datasets.

Comparison

Adjustments:

- only SVM classifier
- RFE-RVNS, RVNS $\rightarrow$ 500 and 1000 iterations, respectively

Reference	Leukemia	Lung	Ovarian	Breast	Colon
RFE-RVNS	99.86[4]	99.51[3]	99.80[3]	99.12[2]	96.69[5]
RVNS	98.84[5]	98.67[3]	98.55[4]	98.66[2]	93.74[6]
GA	95.86[4]	99.49[4]	98.83[4]	-	100[3]
GPSO	97.38[3]	99.00[4]	99.44[4]	-	100[2]
GBC	96.43[5]	-	-	-	91.51[5]
MFMW	-	98.34[6]	-	-	95.16[6]
MRMR+SVM	98.35[37]	-	-	-	91.68[78]
ENN	-	-	99.21[75]	-	81.48[-]
FS+RP	-	-	-	98.97[>100]	-

Figure: The performance of our algorithms, when applied only with the SVM classifier, compared to similar methods.

Conclusion

Methodology's perks:

- Fast Wrapper FS technique
- Successful neighborhood structures
- Different-learning classifiers $\rightarrow$ trustworthy results
- High average accuracy
- Small gene subset size
- Enhanced performance with the hybridization, i.e., RFE-RVNS

Methodology's drawbacks:

- Weaker exploration than population-based metaheuristics

Future Directions

- Hybridization of RVNS with other FS techniques
- Testing of presented methods on more datasets, e.g., multi-class cancer-related, text classification
- Cancer classification necessitates datasets with larger sample

End of presentation

Questions?