A Hybrid Approach to Feature Ranking for Microarray Data Classification

Dusan Popovic[1], Alejandro Sifrim[1], Charalampos Moschopoulos[1], Yves Moreau[1], Bart De Moor[1]

[1] ESAT-SCD / iMinds-KU Leuven Future Health Department, KU Leuven, Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium

{Dusan.Popovic,Alejandro.Sifrim, Charalampos.Moschopoulos, Yves.Moreau, Bart.DeMoor}@esat.kuleuven.be

Abstract. We present a novel approach to multivariate feature ranking in context of microarray data classification that employs a simple genetic algorithm in conjunction with Random forest feature importance measures. We demonstrate performance of the algorithm by comparing it against three popular feature ranking and selection methods on a colon cancer recurrence prediction problem. In addition, we investigate biological relevance of the selected features, finding functional associations of corresponding genes with cancer.

Keywords. genetic algorithms, Random forest, feature ranking, feature selection, gene prioritization, microarrays, classification, cancer

1 Background

High-throughput technologies, such are mass-spectrometry, microarrays and next-generation sequencing, recently empowered biomedical researchers with capabilities to study biological phenomena at the molecular level. Consequently, these technological advances facilitated development of targeted therapies and non-invasive diagnostic tests for certain diseases [1]. However, ever-increasing utilization of these techniques also gave rise to plethora of new problems that seriously challenge traditional views on data analysis. The data sets resulting from high-throughput experiments are often characterized by a large number of highly correlated features, severe signal/noise ratios and a small number of biologically very heterogeneous samples. The described problems are especially prominent when analyzing diseases that display complex patterns of molecular changes, such as cancer. These issues eventually promoted extensive utilization of machine learning algorithms in the field. In this work we present a method that aids identification of relevant predictor variables in this context by combining optimization capabilities of genetic algorithms with robust Random forest feature importance estimation.

Genetic algorithms [2],[3] (GA) are class of search and optimization meta-heuristics inspired by the process of natural selection. They represent a potential solution to the problem at hand as an individual (also called chromosome) that is defined over several, usually binary, variables, called genes. A set of individuals constitutes a population, from which the fittest individuals are selected to be combined and sometimes otherwise altered in order to produce a next generation of solutions. The fitness value reflects the desired quantitative aspect(s) of an individual, and is obtained trough application of a user-defined fitness function. This is essentially an iterative stochastic process that terminates when the optimization objective is achieved or when certain stopping criterion is met. Applications of the genetic algorithms in bioinformatics include amongst others: multiple sequence alignment [4], RNA structure prediction [5] and biomarker discovery [6].

The Random forest [7] (RF) is popular classification method that has been applied to numerous scientific fields so far. It essentially operates by constructing an ensemble of fully-grown decision trees built on different bootstraps extracted from the data, with additional randomness injected in algorithm by selecting splitting variables from random subsets of all possible candidate splits. Random forests are especially suitable for dealing with problems characterized by high dimensionality and severe correlation between predictors. They also produces internally estimated feature importance scores (for discussion on different methods together with their possible biases see [8]) that take into account interactions between features, which is a utility of great interest in many applications. The later capability is often exploited for multivariate feature selection, as well as in explanatory analyses ("opening a black-box"). These two advantages are the main reason behind growing popularity of the method in bioinformatics, where it has been used for analysis of microarray data [9] and DNA sequencing [10] , among other applications.

2 Materials and Methods

2.1 A hybrid approach to feature ranking problem

The proposed algorithm is depicted in Figure 1. It is essentially wrapper approach to feature selection/ranking [11]. We represent subset of biological genes to be used for subsequent classification (chromosome in GA) as a vector of binary values. The initial generation is created randomly, with probability of 1% for each feature to be selected in single chromosome. Prior to a GA generation run we take class-balanced bootstrap [12] from the original data, after which subsets are created for every chromosome given their particular selections of features. These subsets are then used for training Random forest classifiers, whose out-of-bag accuracy values are then fed back to the genetic algorithm as a fitness of corresponding chromosomes. We apply bootstrapping aiming to mitigate the GA potential for over-fitting, which is an espe-

cially severe problem when the later is used in wrapper-based feature selection context [13]. Also, this procedure should promote "longevity" of solutions that are robust against small perturbations in data, and thus generalize well. Settings of other parameters of GA and RF are provided in Table 1.

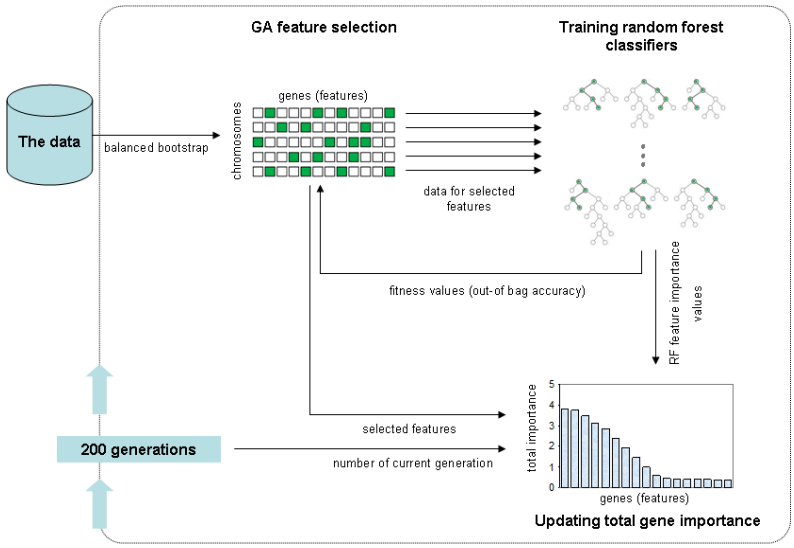

Fig. 1. Workflow of the feature ranking using hybrid approach

When GA converges to the general region of maximal fitness; longevity and classification performance of particular genes are continuously rewarded until the end of algorithm execution. In particular, the importance of a biological gene is increased by corresponding RF-FI value during every generation and for every chromosome where the gene is present. This principle essentially mimics usage of gene conservation scores for prediction [14]. Formally, if s_{ijk} is a binary function that indicates if feature j ($j=1..f$, here f stands for number of biological genes) is selected in chromosome i ($i = 1..c$, here $c=100$) during generation k ($k = 0..g$, here $g=200$), r_{ijk} is the random forest feature importance measure (note that $r_{ijk}=0$ *if* $s_{ijk}=0$) and **1** is an unit vector of size $1xf$; vector of importance across all of the genes and for a single generation becomes :

$$GI_k = \mathbf{1}\left(\begin{bmatrix} s_{11k} & s_{12k} & \cdot & \cdot & s_{1fk} \\ s_{21k} & s_{22k} & \cdot & \cdot & s_{2fk} \\ \cdot & \cdot & & & \cdot \\ s_{c1k} & s_{c2k} & \cdot & \cdot & s_{cfk} \end{bmatrix} \circ \begin{bmatrix} r_{11k} & r_{12k} & \cdot & \cdot & r_{1fk} \\ r_{21k} & r_{22k} & \cdot & \cdot & r_{2fk} \\ \cdot & \cdot & & & \cdot \\ r_{c1k} & r_{c2k} & \cdot & \cdot & r_{cfk} \end{bmatrix}\right) \quad (1)$$

That is, values of s_{ijk} multiplied by r_{ijk} are summed across chromosomes for each gene, resulting in a gene importance vector for given generation. Accordingly, the final gene importance vector is:

$$GI = \sum_{k=100}^{g} GI_k \tag{2}$$

,where k starts from 100, which is the approximate moment when the algorithm starts exploring the general area of a solution (half of the total number of generations).

Finally, to assess utility of our method for aiding microarray classification we compare it against three additional feature selection methods that are often used in microarray analyses:

- *Wilcoxon rank-sum test :* Features are ranked by values of the test obtained by comparing a vector of single variable values corresponding to positive cases and that corresponding to negatives. Final subset of features can be than selected according to predefined p-value or cut-off .
- *t-test :* Similarly to former.
- *Random forest feature importance :* The features are ranked by difference between out-of-bag mean square error obtained by single trees when values of a feature are shuffled to that achieved on unaffected data. To produce global estimate these values are averaged over the entire ensemble and divided by standard deviation. As before, the cut-off for selecting feature subset can be chosen arbitrary or by statistical modeling.

	Parameter	Value
Genetic algorithm	Type of selection	Stochastic universal sampling
	Sigma scaling	On
	Sigma scaling coefficient	1
	Size of the population	100
	Number of generation	200
	Type of crossover	Uniform
	Probability of crossover	0.7
	Type of mutation	Simple (flipping a value of single bit)
	Probability of mutation per bit	0.5/number of genes
RF	Feature importance estimation method	Permutation accuracy importance
	Number of trees in ensemble	10
	Number of variables randomly selected for a split	Square root of total number of variables
	Minimal number of observations in a leaf	1

Table 1. The parameter setting of the genetic algorithm and random forest classifiers that are used within fitness function

Once we have variables ordered by all four methods, we train classifier on the whole training set, selecting the first five, ten and fifty top-ranked features from each list. After this, resulting classifiers (12 in total) are tested against independent test set. The classifier of choice is a Random forest, with the number of trees set to one hundred.

2.2 Data sets

We demonstrate our method using three publicly available microarray data sets of colon cancer samples that have been generated on Affymetrix HG U133 2.0 Plus platform. For all three, we consider cancer recurrence as an outcome of interest. Two data sets [15] have been merged to be used for training, while the third set [16] was used for testing. Further details on data can be found in Table 2.

Data set GEO accession no.	GSE17536	GSE17537	GSE5206
Author	Smith (MMC) [15]	Smith (VMC) [15]	Aronow [16]
Preprocessing method	MAS5	MAS5	RMA
No. positive outcomes	36	20	16
No. negative outcomes	109	35	58
Role in the study	a part of training set	a part of training set	the test set

Table 2. Details on the used data sets

3 Results and Discussion

The ROC curves obtained on the test set using four different feature selection methods and three different numbers of selected features are depicted in Figure 2. The corresponding AUC (area under the curve) values can be found in Table 3. It is immediately apparent from these that, in this particular setup, our method outperforms the other three regardless to the number of features selected. Furthermore, most of AUC gain obtained with the hybrid approach seems to be concentrated in regions of low false positive rates. This is often of great importance in the real-world applications, due to the high costs associated with confirmatory functional experiments.

Additionally, these results are comparable or better than that reported in literature for the same classification problem. For example, the genetic signature reported in Wang et al. [17] achieves AUC of 0.74 on the data set that has been used in the study, while further refinement of the same method [18] reaches an AUC of 0.66 on an external validation set. Also, a study by Lin et al. [19] reports AUCs of 0.73 and 0.80 obtained on the two data sets using genetic signatures augmented with the clinical data.

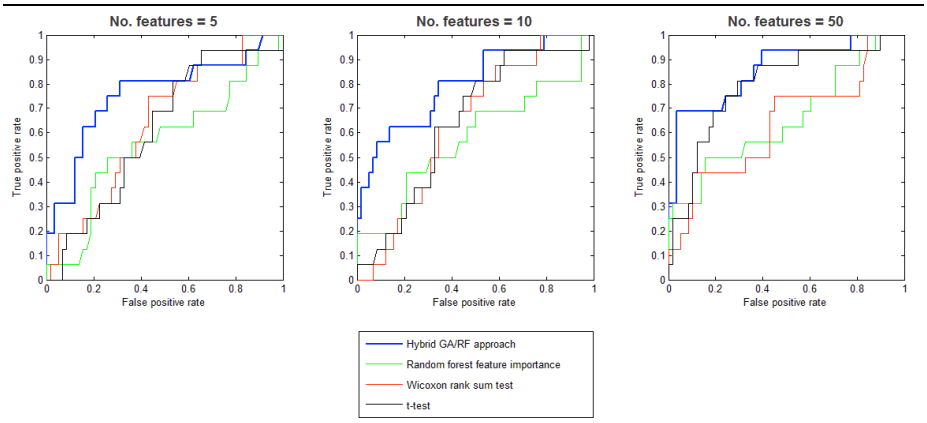

Fig. 2. ROC curves obtained on the test set using four different feature ranking methods and three different numbers of selected features

Method	Number of features		
	5	10	50
Hybrid GA/RF approach	**0.7575**	**0.7985**	**0.8572**
Random forest feature importance	0.5571	0.5765	0.6546
Wilcoxon rank sum test	0.6342	0.6325	0.6352
T-test	0.6121	0.6433	0.7936

Table 3. Values of area under the ROC curve for classifiers using different number of top-ranked variables as suggested by four feature ranking methods. The best values per number of selected features are indicated in bold.

To investigate a possible functional relation of resulting highly ranked genes to colorectal cancer, we also performed a functional analysis using the Ingenuity Pathway Analysis suite. Out of the twenty top ranked genes, eight (ALDH1A3, DNAJA3, FAM65C, HOXA7, MCM8, TM4SF1, PXDN, SEC31A) have been reported to be cancer-related. ALDH1A3, DNAJA3 and HOXA7 play a role in apoptopis, an important hallmark in oncogenesis. However, TM4SF1 is the only gene to be reported specifically for colorectal cancer recurrence. Interestingly a gene of unknown function, TSPAN11, was found which belongs to the same protein family of tetraspanins as TM4SF1. We blasted the nucleotide sequence of TSPAN11 to find sequence paralogs. One of the top ranking hits was CD151 which has been shown to show differential expression in colorectal cancer [20]. However we can't exclude the possibility that the oligoprobe on the microarray platform shows aspecific hybridization in relation to CD151 and TSPAN11 due to their high sequence similarity.

4 Conclusions and the Future Work

We propose and demonstrate a novel method for feature ranking that combines genetic algorithm-facilitated search and Random forest feature importance measures. We tested it against three feature ranking algorithms in context of microarray-based colon cancer classification, achieving superior results in terms of area under the ROC curves. Furthermore, we observe functional association of several genes from top of our prioritized list with cancer, indicating that the method might be usable in wider context of biomarker discovery research.

However, as these genes have been judged predictive thought "guilt-by-association" rather than by proving their causality given the disease, further analyses are needed to establish the utility of the method beyond feature ranking. In the future, we also plan to test this hybrid approach in different high-throughput setups and for various biological classification problems. In addition, we will further investigate the idea of utilizing artificial "conservation scores" in optimization by genetic algorithm in general, perhaps for guiding the search process via a fitness function.

Acknowledgements: The authors would like to acknowledge support from:

- Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, START 1, OT 09/052 Biomarker, several PhD/postdoc & fellow grants
- Industrial Research fund (IOF): IOF/HB/10/039 Logic Insulin, IOF: HB/12/022 Endometriosis
- Flemish Government:
    - FWO: PhD/postdoc grants, projects: G.0871.12N (Neural circuits), research community MLDM
    - IWT: PhD Grants; TBM-Logic Insulin, TBM Haplotyping, TBM Rectal Cancer
    - Hercules Stichting: Hercules III PacBio RS
    - iMinds: SBO 2013; Art&D Instance
    - IMEC: phd grant
- Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate)
- COST: Action BM1104: Mass Spectrometry Imaging, Action BM1006: NGS Data analysis network

The scientific responsibility is assumed by its authors.

5 References

1. Glas A.M., Floore A., Delahaye L.J., Witteveen A.T., Pover R.C., Bakx N., Lahti-Domenici J.S., Bruinsma T.J., Warmoes M.O., Bernards R., Wessels L.F., Van't Veer L.J. (2006) Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics 7:278
2. Fraser A. (1957) Simulation of genetic systems by automatic digital computers. I. Introduction. Aust. J. Biol. Sci. 10, 484–491
3. Holland J.H. (1975) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press
4. Gondro C, Kinghorn BP (2007). A simple genetic algorithm for multiple sequence alignment. Genetics and Molecular Research 6 (4): 964–982. PMID 18058716.
5. Van Batenburg FH, Gultyaev AP, Pleij CW (1995). An APL-programmed genetic algorithm for the prediction of RNA secondary structure. Journal of Theoretical Biology 174 (3): 269–280. doi:10.1006/jtbi.1995.0098. PMID 7545258.
6. Popovic D, Sifrim A, Pavlopoulos GA, Moreau Y, De Moor B (2012). A simple ge netic algorithm for biomarker mining. In Proceedings of the 7th IAPR international conference on Pattern Recognition in Bioinformatics (PRIB'12),Springer-Verlag, Berlin, Heidelberg, 222-232.
7. Breiman, Leo (2001). Random Forests. Machine Learning 45 (1): 5–32.
8. Strobl C, Boulesteix AL, Zeileis A, Hothorn T(2007). Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8:25.
9. Huang X, Pan W, Grindle S, Han X, Chen Y, Park SJ, Miller LW, Hall J(2005). A comparative study of discriminating human heart failure etiology using gene expression profiles. BMC Bioinformatics, 6:205.
10. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, et al. (2005) Identifying SNPs predictive of phenotype using random forests. Genetic epidemiology 28: 171–182
11. Saeys,Y et al. (2007) A review of feature selection techniques in bioinformatics. Bioinformatics, 23, 2507–2517.
12. Efron B, Tibshirani R (1993). An Introduction to the Bootstrap. Chapman & Hall/CRC, Boca Raton FL
13. Loughrey J, Cunningham P (2004). Overfitting in wrapper-based feature subset se lection: the harder you try the worse it gets. Proceedings of International Conference on Innovative Techniques and Applications of Artificial Intelligence, vol. 33, p. 43.
14. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288: 136–140.
15. Smith JJ, Deane NG, Wu F, Merchant NB et al. (2010). Experimentally derived me tastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 2010 Mar;138(3):958-68.
16. Kaiser S, Park YK, Franklin JL, Halberg RB et al. (2007) Transcriptional recapitula tion and subversion of embryonic colon development by mouse colon tumor models and human colon cancer. Genome Biol;8(7):R131.

17. Wang Y., Jatkoe T., Zhang Y., Mutch M.G., Talantov D., Jiang J., McLeod H.L., Atkins D. (2004): Gene expression profiles and molecular markers to predict recur rence of Dukes' B co-lon cancer. J. Clin. Oncol. 22, 1564–1571
18. Jiang Y., Casey G., Lavery I.C., Zhang Y., Talantov D., Martin-McGreevy M., Skacel M., Manilich E., Mazumder A., Atkins D., Delaney C.P., Wang Y. (2008): Development of a clinically feasible molecular assay to predict recurrence of stage II colon cancer. J. Mol. Diagn. 10, 346–354
19. Lin Y.H., Friederichs J., Black M.A., Mages J., Rosenberg R., Guilford P.J., Phillips V., Thompson-Fawcett M., Kasabov N., Toro T., Merrie A.E., van Rij A., Yoon H.S., McCall J.L., Siewert J.R., Holzmann B., Reeve A.E.: Multiple gene expression classi fiers from different array platforms predict poor prognosis of colorectal cancer. Clin. Cancer. Res., 13, 498–507 (2007)
20. Lin PC, Lin SC, Lee CT, Lin YJ, Lee JC (2011). Dynamic change of tetraspanin CD151 membrane protein expression in colorectal cancer patients. Cancer Invest , 29(8): 542-7.