Chapter

OPTIMIZATION FOR FEATURE SELECTION IN DNA MICROARRAYS

Vincent Gardeux[1]*, René Natowicz[2], Maria Fernanda Barbosa Wanderley[3] and Rachid Chelouah[1]

[1]L@RIS, EISTI, Avenue du Parc, Cergy, France
[2]ESIEE-Paris, University of Paris-Est, Noisy-le-Grand, France
[3]Federal University of Minas Gerais, Belo Horizonte, Brazil

ABSTRACT

We present two methods for feature selection in high throughput transcriptomic data, in which the subsets of selected variables (the genes) are optima of a multi-objective function. In the clinical trials, the number of embedded patient cases is never higher than in the hundreds, while the number of gene expressions measured for each patient is higher than tens of thousands. These trials aim to better understand the biology of the phenotypes at the genomic level, and to better predict the phenotypes in order to give each patient the best treatment.

Our first method states that the gene subsets are the optima of a bi-objective function. This function is a tradeoff between the size of the gene subset and the discrimination of the phenotypes, expressed as the inter-class distance. Because the gene selection stage is independent of the prediction model, it is a filter method of feature selection. The second method aims to select gene subsets that will optimize the performance of a specific prediction model. It is a wrapper approach of the feature selection problem. The optimal gene subsets are computed by a line search optimization heuristic which maximizes the performances of a linear discriminant analysis.

Using public datasets in oncology we compared our results to those of the main previous methods. Our optimization approach of the gene subset selection almost always returned subsets that were significantly smaller than those of the previous methods, the performance of our predictors almost always being higher, and being more robust. In the two methods we searched the space of gene subsets for optima of an explicit multi-objective function. Meta-heuristic methods are well suited to address these optimization problems, specifically in high dimensional spaces.

Keywords: Optimization, DNA Microarray Analysis, Feature Selection, Bioinformatics, Oncology

* E-mail address: research@gardeux-vincent.eu

1. INTRODUCTION

The various fields of bioinformatics generate ever increasing amounts of data of ever increasing dimensions (genomics, genetic sequencing, proteomics, and beyond). Optimization is involved in many of these dimensions, from the reverse engineering of gene networks to the modeling of the three-dimensional structures of proteins as well as the modeling of their folding, to name but a few.

In genomics, cDNA microarrays are part of a technology that allows the simultaneous measurement of the expressions of tens of thousands of genes. Among these features, biologists and physicians search for the most relevant ones, i.e. the features that help to understand the biological processes that underlie specific pathological phenotypes. Cancer research is a very active field and, considering the huge amount of very high dimensional data, it has a high demand for methods which could focus the attention on small subsets of features, as they could be the drivers of complex phenotypes or resistance to treatments. These features could be promising targets for new therapeutic strategies.

In oncology, clinical trials provide this gene expression data, to which clinical and biological information is attached, as is the outcome of a treatment, and most of the times follow up information. A lot of data mining techniques are available to extract human-understandable information and structures from large datasets, but genomic applications are specific. Each piece of data is a class-labeled high-dimensional vector of gene expressions, and the number of cases is far smaller than the number of dimensions, i.e. the number of variables. Hence, there is a need for methods that can select small sets of variables (compared to the number of cases), able to predict particular biological phenotypes.

The first reason for searching small subsets of variables is the robustness, because a large set of variables is likely to model the specificities of the dataset, losing its ability to generalize and leading to overfitting. The second reason is the usefulness for biological modeling. Genetic interaction networks are the biological mechanisms underlying phenotypes, and the complexity of these networks is far more than linear in the number of its components. The last reason comes from the use in clinical routine, since small subsets of genes are amenable to robust and cheap low-throughput technologies.

A predictive modeling is the whole process of feature selection stage, model design, and statistical validation of a predictor[1]. In this chapter we focus our attention on the feature selection process, specifically performed using optimization procedures. The search space is the set of the gene subsets. Because the microarrays measure the expressions of thousands of genes (let's say $n \approx 20000$), the number of gene subsets is more than "*astronomic*" ($2^n \approx 10^{6000}$). The size of the search space is so high that *classic* search methods in optimization and even metaheuristic methods are not directly applicable. Thus, we have to design specific methods, dedicated to the optimization in very high-dimensional spaces in order to alleviate the *curse of dimensionality* [3].

In the first section of this chapter we present a feature selection method by the optimization of a bi-objective function. This method returns exact optima of the function.

[1] The subset of genes selected after feature selection

Although very efficient, this method can only address the optimization of functions which are tradeoffs between independent criteria. In the second section we address a more general problem in which the bi-objective function is relaxed to a single objective function, and solved by a metaheuristic method specifically designed to fit high-dimensional problems. Then we present the performances of these two methods on nine different genomic datasets in oncology.

2. A FILTER-LIKE OPTIMIZATION

Feature selection methods usually fall into two main categories, *filters* and *wrappers*. Filter methods rank the features by a valuation criterion and retain only the features whose values are above a fixed threshold. These methods are completely independent of the classifiers, i.e. the classification models which are used afterwards to make a prediction. Conversely, wrapper methods search the set of features for optimal subsets given a particular classifier. A performance measure is attached to each subset depending on the performance it obtains using this particular classifier on the learning dataset. Thus, the classifier is a part of the valuation criterion of the subset of features, if not the valuation function itself.

In this section, we present the δ-test method, which is specific because, although the selection is independent of the classifier (as is also the case for the filter methods), it does not have any filtering threshold. We set up an optimization procedure driven by the selection of the variables in the subset. A solution is thus represented as a binary vector, each component representing the belonging (or not) of the variable to the subset. Because we searched subsets of features that discriminated two classes as much as possible and were as small as possible, we chose to select these features depending on two criteria: the first objective being the interclass distance induced by the subset of features, to be maximized, and the second being the size of the subset itself, to minimize [12]. We chose to combine these two conflicting objectives in one single aggregate objective function $F_\omega(S)$, defined as follows:

$$F_\omega(S) = \omega \times d\big(c(S), c'(S)\big) + (1-\omega) \times (1 - |S|) \quad (1)$$

where S is a subset of variables (a possible solution), $c(S)$ and $c'(S)$ are the class centroids restricted to the variables of the subset S, $d(c(S), c'(S))$ is the Euclidean distance between the class centroids, and $\omega \in [0,1]$ is a weight parameter.

As regards to the weight parameter ω, the two limit cases are $\omega = 0$, at which $F_\omega(S) = 1 - |S|$ whose maximum is reached at $S^*(\omega) = \emptyset$, and $\omega = 1$ at which $F_\omega(S) = d\big(c(S), c'(S)\big)$ whose maximum is reached at $S^*(\omega) = \mathcal{S}$ (the whole set of features.) Between these limit cases, the optimal solution $S^*(\omega)$ is a subset which, given the weight parameter ω, achieves the best balance between the two conflicting objectives.

The first objective of this function, the interclass distance induced by the subset S, is separable, which means that the contribution of any variable s of S to the interclass distance is independent of the contribution of the other variables in S. This contribution $\delta(s)$ is:

$$\delta(s) = \sqrt{\bar{x}_s^2 - \bar{x}'^2_s} \quad (2)$$

where $\bar{x}_s$ and $\bar{x}'_s$ are the means of s on the two classes. It follows that $F_\omega(S \cup \{s\})$ is higher than $F_\omega(S)$ for any variable s whose contribution to the interclass distance is higher than $(1-\omega)/\omega$. Such a variable increases the function F_ω. By recurrence starting at $S = \emptyset$, for any ω the optimum $S^*(\omega)$ is the set of variables that increase the function F_ω. Reciprocally, given the subset $S(k)$ of the k variables of highest contribution to the interclass distance, there exists a maximum interval $I(\omega) = [\omega_{k-1}, \omega_k[$ such that $\forall \omega \in I(\omega),\ S(k) = S^*(\omega)$. The consequence is that the family of functions $F_\omega(S),\ \omega \in [0,1]$ has exactly $n+1$ optima. These optima are the subsets $S(k),\ k \in \{0,1,\dots,n\}$, of the k variables of highest contribution to the interclass distance.

This feature selection method is available online with an **R** package at:

http://gardeux-vincent.eu/DeltaTest.php

2.1. Automatic Selection of the Optimal Size

A main issue of the filter method approaches is to find the *right* filtering threshold. In the case of the above filter-like optimization, the issue is to select among the $n+1$ optimal subsets of variables, the one of optimal size, i.e. an optimal set $S(k^*)$.

In binary classification, the performance criteria are three folds: the *accuracy* of the predictor (the probability of a case to be allocated to its proper class), its *sensitivity* (the probability of a positive case to be properly allocated), and *specificity* (that of a negative case to be properly allocated). The performance of each optimal subset $S(k)$ is assessed on the learning dataset by a given classifier model and the smallest optimal subset of highest performance is the optimal subset of optimal size, i.e. the set of variables $S(k^*)$. This computation of the optimal size k^* is very efficient because it is conducted in the set of optimal subsets, whose size is $n+1$.

3. A WRAPPER OPTIMIZATION

The limit of use of the previous method is the separability of the criterion. In the specific case of the function F_ω, the contribution of a variable s to the interclass distance is independent of the set S to which the variable is added. In many cases this situation may be oversimplifying the problem. For instance, the previous approach could not fit an objective function that has to take into account the correlations between the variables. In such a situation, when adding a variable to a set, the variation of the objective function not only depends on the added variable, but also on the other variables contained in the set. The same argument would hold if the objective function had to account for the performance of the predictor. The aim of the second method that we present is to select a set of variables which maximizes the performance of a given classifier model. In this specific case, we chose to define two objectives to optimize: the accuracy of the prediction and the size of the variable set. They are aggregated in a convex linear combination as above.

Hence, the aggregate objective function $F_\omega(S)$ becomes:

$$F_\omega(S) = \omega \times \text{accuracy}(S) + (1-\omega) \times (1 - |S|) \qquad (3)$$

where $\text{accuracy}(S)$ is the accuracy of the predictor whose inputs are the values of the variables belonging to S. This accuracy is measured on the learning set of case. Because the performance of the predictor is part of the optimization function, this method is a wrapper approach of the feature selection problem.

3.1. ABEUS: A High-Dimensional Optimization Procedure

The objective function defined in the equation (3) is similar to the ones used by usual wrapper methods. It requires the tuning of the weight parameter ω in order to hopefully find a set of variables of both high learning accuracy and small size.

On high throughput expression data, the number of variables, which can amount to more than 50,000, is so high that the usual metaheuristic methods cannot be applied directly. Thus, we addressed this new issue by an approach that we had previously developed for solving high dimensional problems in continuous spaces [8] [9]. This previous method, named Enhanced Unidimensional Search (EUS), was a decomposition scheme combined with a line search procedure similar to a relaxation method. The main idea was to optimize the high-dimensional problem by splitting the problem according to its individual dimensions and addressing independent optimizations on each dimension. This method showed quick convergence rates, did not require any parameter tuning, and had been successfully applied to large scale optimization of problems in continuous spaces.

In our case, the main specificity of the optimization modeling is the elementary step of the optimization process which is to increase the dimension of the problem by adding a variable to a set of variables, or to decrease it by removing a variable. Thus, we modified the EUS method to address this specific issue. The resulting method is the Binary EUS (BEUS) method. Its scheme is exactly that of the EUS. The difference is in the definition of the neighboring of an n-dimensional solution: Given a local optima S of the function, the neighboring solutions are the sets of variables obtained by adding a variable to, or removing a variable from, the set S.

In order to simplify the optimization process, we made a second modification to the EUS procedure. Our aim was to simplify the two-aggregated-objectives function to a single-objective one, by moving the size criterion from the function to the optimization scheme. The function F_ω (equation (3)) is thus modified into the single objective function f:

$$f(S) = \text{accuracy}(S) \tag{4}$$

The minimization of the size of the solution is no longer explicit, it becomes a part of the optimization scheme: Given a local optimum S, a neighboring solution S' is retained by the optimization procedure if the value $f(S')$ is higher than $f(S)$ or, in case of equality, if the size of S' is lower. In the resulting method, the minimization of the size of the solutions is achieved by the search scheme. Because the size is automatically minimized, we called this method Automatic BEUS (ABEUS). The *pseudo-code* of this method is in figure 1.

```
Procedure ABEUS
(Let S be the current solution, represented by a binary vector)
begin
    S is initialized at random
    do
        v = f(S)
        for i = 1 to n
            S_i = 1 − S_i (S_i is the ith component of the solution)
            v' = f(S)
            if (v' < v) then
                (the new solution has poorer performance)
                S_i = 1 − S_i (back to S)
            else if (v' = v)
                (same performance of both solutions)
                if S_i = 0 then
                    (the new solution is smaller, we keep it)
                else
                    (same performance but larger size)
                    S_i = 1 − S_i (back to S)
                end if
            end if
        end for
    while a better solution was found
end
```

Figure 1. ABEUS optimization procedure

It is important to notice that this transformation is only possible because we have considered the size criterion to be less important than the accuracy criterion. Thus, it becomes possible to put aside the minimization of the size of the solution, and to focus on that task later.

4. APPLICATION

In order to build a classification model from the optimal subsets of variables found with our methods, we had to choose a classifier method. The Linear Discriminant Analysis (LDA) classifier is often used for prediction in DNA microarrays studies. Indeed, Dudoit *et al*. [6] showed that LDA was one of the best classifier models for discriminating and classifying gene expression data. In addition, Miller *et al*. [14] showed that Diagonal LDA predictors had higher sensitivity values. Furthermore, Hess *et al*. [13] reported a thorough study of the prediction of the response to preoperative chemotherapy in breast cancer, in which the subsets of genes were computed in the ranking of the p-value to a t-test. In this study, many classifier models were evaluated by the authors (support vector machines, DLDA, K-nearest neighbors) under various parameter settings. The performances of a total of 780 distinct classifiers (sets of genes and classifier models) were assessed by cross-validation procedures. The conclusion of the study was that DLDA had better performances than the other classifier models on the

datasets under study. For these reasons we chose the DLDA classifier to address the classification of the genomic data with our methods of feature selection.

5. RESULTS

We compared the performance of our predictors on nine different datasets in oncology. For each dataset we have assessed the robustness of our predictive modeling by means of cross-validation procedures. In genomic studies, a set of genes is called a molecular signature, or simply a signature, and a variable is called a DNA-probe (or a gene by abuse of language).

5.1. Benchmark on Datasets in Oncology

We first considered a set of six publicly available datasets in oncology, on which studies had been conducted. Each of them was a two-class dataset, whose characteristics are listed in table 1.

Table 1. Reference articles of six datasets for benchmarking

data type	reference	# cases	# DNA-probes
colon	[1] [2]	62	2000
lymphom	[22] [3]	77	5469
leukemia	[11] [4]	72	7129
prostate	[24] [5]	102	10509
brain	[18] [6]	60	7129
ovaries	[4] [7]	54	22283

We applied our predictive modeling to these datasets and conducted a three-fold cross-validation procedure whose results are in tables 2 and 3.

In table 4 we have the comparison of our results with previously published results, selected due to their lack of bias (i.e. the signatures, their sizes and the parameters of the classification model had been computed without any reference to the test subsets on which the predictor performances had been measured). [2] [23] [7]. Concerning the dataset of ovarian tumors, we did not find any non-biased studies. The last two lines of the table, δ-DLDA and *ABEUS*, are the performances of our predictive modelings.

[2] http://genomics-pubs.princeton.edu/oncology/
[3] http://www.gems-system.org/
[4] http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
[5] http://www.gems-system.org/
[6] http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
[7] http://data.cgt.duke.edu/clinicalcancerresearch.php

Table 2. Three-fold cross-validation of the δ-DLDA: δ-valued signatures and optimal sizes

data	colon	lymphoma	leukemia	prostate	brain	ovaries
Learning DataSet						
#probes	6.790±4.203	4.070±1.976	3.140±0.928	3.500±1.025	9.490±5.017	11.39±5.737
Accuracy	0.905±0.027	0.904±0.029	0.983±0.017	0.934±0.020	0.855±0.038	0.969±0.022
Sensitivity	0.910±0.036	0.921±0.066	0.982±0.031	0.937±0.040	0.823±0.070	0.950±0.041
Specificity	0.896±0.069	0.899±0.032	0.983±0.023	0.931±0.030	0.872±0.049	0.984±0.024
Test DataSet						
#probes	6.790±4.203	4.070±1.976	3.140±0.928	3.500±1.025	9.490±5.017	11.39±5.737
Accuracy	0.807±0.080	0.857±0.064	0.950±0.046	0.901±0.053	0.660±0.094	0.695±0.092
Sensitivity	0.853±0.101	0.796±0.174	0.939±0.087	0.898±0.093	0.533±0.172	0.619±0.168
Specificity	0.722±0.168	0.877±0.065	0.956±0.059	0.904±0.078	0.728±0.128	0.756±0.133

Table 3. Three-fold cross-validation of ABEUS-DLDA modeling

data	colon	lymphoma	leukemia	prostate	brain	ovaries
Learning DataSet						
#probes	4.260±1.753	3.520±1.044	2.440±0.637	4.200±1.131	6.860±1.949	3.800±0.775
Accuracy	0.964±0.022	0.993±0.012	1.000±0.000	0.984±0.011	0.978±0.021	1.000±0.000
Sensitivity	0.952±0.031	1.000±0.000	1.000±0.000	0.989±0.015	0.967±0.046	1.000±0.000
Specificity	0.984±0.030	0.990±0.015	1.000±0.000	0.979±0.016	0.984±0.022	1.000±0.000
Test DataSet						
#probes	4.260±1.753	3.520±1.044	2.440±0.637	4.200±1.131	6.860±1.949	3.800±0.775
Accuracy	0.785±0.064	0.881±0.065	0.908±0.063	0.891±0.048	0.593±0.101	0.666±0.110
Sensitivity	0.813±0.090	0.889±0.135	0.875±0.139	0.880±0.091	0.503±0.165	0.628±0.149
Specificity	0.732±0.198	0.879±0.073	0.925±0.071	0.901±0.068	0.642±0.135	0.696±0.162

5.2. Prediction of Preoperative Chemotherapy in Breast Cancer

The second application was performed on two different datasets coming from clinical trials conducted at MD Anderson Cancer Center (Houston, Texas, USA) [13], aiming to predict the response to chemotherapy in breast cancer. In [13] the dataset of 133 patient cases was split into a training set of 82 patient cases and a test set of 51 patient cases, each one with the same ratio of responder cases and residual disease (respectively 1/3 and 2/3.)

In [25], a third independent dataset of 91 patient cases is reported[8]. This dataset is from a cohort of the same clinical trial. The measurements of the expression levels were conducted with the same microarray (Affymetrix U133A), platform, and protocol. The performance of the classifications [9] are in table 5. In this table δ-DLDA-30 and δ-DLDA-11 are the DLDA classifiers respectively designed on our signatures of size 30 and 11. The third column corresponds to the signature returned by the ABEUS method, whose size was 7 genes. The two last columns are the best previously published results.

[8] Dataset publicly available on the Gene Expression Omnibus (GEO) database with accession number GSE20271

Table 4. Mean accuracies (in %) and mean number of probes ($\bar{p}$) of the non-biased results published for the benchmark datasets

bibl. ref.	colon		leukemia		prostate		brain		lymphoma		ovaries	
	Ac.	$\bar{p}$	Ac.	$\bar{p}$	Ac.	$\bar{p}$	Ac.	$\bar{p}$	Ac.	$\bar{p}$	Ac.	$\bar{p}$
[20]	-	-	-	-	-	-	60.00	21	-	-	-	-
[26]	85.83	20	-	-	-	-	-	-	-	-	-	-
[5]	-	-	-	-	-	-	-	-	83.33	6	-	-
[19]	82.33	20	-	-	-	-	-	-	-	-	-	-
[17]	82.03	-	94.40	-	91.22	-	-	-	-	-	-	-
[21]	-	-	-	-	94.12	22	-	-	-	-	-	-
[16]	85.71	30	-	-	94.11	20	-	-	-	-	-	-
[10]												
F-test	84.05	15.1	-	-	91.18	126.4	-	-	-	-	-	-
∂W	76.70	35.1	-	-	94.60	756.6	-	-	-	-	-	-
∂RW	78.60	43.3	-	-	94.70	573.3	-	-	-	-	-	-
∂Spb	80.30	31.8	-	-	94.80	95.5	-	-	-	-	-	-
SVM-RFE	85.48	26.4	-	-	94.18	43.2	-	-	-	-	-	-
GLM Path	81.91	1.3	-	-	94.09	1.6	-	-	-	-	-	-
Rand. Forest	89.40	49.8	-	-	94.10	81	-	-	-	-	-	-
δ-DLDA	85.50	9.05	98.60	2.97	94.10	4.00	68.30	8.27	88.30	4.16	75.90	15.98
ABEUS	83.90	7.00	97.20	6.74	87.30	7.29	70.00	14.03	90.90	9.31	68.50	8.76

Table 5. Performance of five different predictors

Method #probes	δ-DLDA-30 30	δ-DLDA-11 11	ABEUS 7	DLDA-30 [13] 30	BI-Maj-30 [15] 30
Training dataset (82 cases)					
Accuracy	0.780	0.804	0.951	-	-
Sensitivity	0.761	0.809	0.904	-	-
Specificity	0.786	0.803	0.967	-	-
Independent test dataset (51 cases)					
Accuracy	0.863	0.882	0.803	0.765	0.863
Sensitivity	0.846	0.846	0.538	0.923	0.923
Specificity	0.868	0.894	0.894	0.711	0.842
Independent test dataset (91 cases)					
Accuracy	0.670	0.659	0.758	0.725	0.681
Sensitivity	0.631	0.631	0.578	0.632	0.579
Specificity	0.680	0.666	0.805	0.750	0.708

6. DISCUSSION

6.1. Benchmark of Six Oncology Datasets

The performances of our straightforward optimization method (δ-test method) are in Table 2. Although this method always returns the optimum of the bi-objective function, the performances of the predictors were not optimal. This comes from the fact that the subsets of genes were computed with no regard to the predictor model. Moreover, we can see on the brain and ovaries datasets that the performances of our predictor decreased sharply on the testing datasets. This kind of behavior is the sign of data overfitting. However, this overfitting seems to be less severe with the δ-test method than with the ABEUS method whose performances are listed in Table 3.

In this table, one can see clearly that the objective function we chose is directly related to the classifier model, because the performances on the learning dataset are those obtained by maximizing the objective function, which is the accuracy of the prediction. Our optimization procedure can effectively solve the optimization problem with a high accuracy and few genes. But the consequence is a higher overfitting, although the performance is close to the ones of the δ-test predictive modeling.

In Table 4, we compared the performances of our modelings to those of previously published modelings. We can see that the performance of our modelings was at the level of, or were higher than almost all of the others, and that the size of our signatures was almost always significantly smaller. A noticeable exception is GLMPath [10] whose performance, even though slightly lower than ours, was obtained with remarkably small numbers of genes.

6.2. Prediction of Preoperative Chemotherapy Outcome in Breast Cancer

Table 5 summarizes the results of our modelings on three datasets in breast cancer. Once again, we can see that the ABEUS method successfully maximized the accuracy on the learning dataset and reduced the number of genes to seven (which is remarkably low for this specific problem). The performance of the δ-test method is not outperformed on the 51-case dataset but the performance is slightly lower on the 91-case dataset. It is interesting to compare this result to the ones of the DLDA-30 predictor [13] because the feature selection method used in this study was a filter method in which the genes were ranked by the p-value of a t-test. Compared to this filter method on the 51-case dataset, the performance of the δ-test modeling is significantly higher.

CONCLUSION

The very high-dimensional data coming from high throughput technologies is challenging for the field of optimization. On such bioinformatic data, the main concerns are on the one hand the reduction of the dimensionality, and on the other hand the robustness of the predictive modeling. Compared to the statistical approaches of feature selection which are the mainstream in bioinformatics, the optimization approaches are simple, efficient to compute, and most of the time the performances and the dimension reduction rates are higher. But, as with the statistical approaches, the lack of robustness of the performances remains an issue.

One lesson of our studies in optimization for bioinformatic problems is that the overfitting of the data is a phenomenon that occurs even with very small sized optimal solutions. This counter-intuitive result may indicate that the initial dimensionality of the problem is so high that the space of solutions is under-sampled by the datasets to a degree that does not only lead to imprecisions of the predictions outside the learning sets, but to high distortions. The consequence of this under-sampling is that the robustness cannot be assessed afterwards, but should be included in the optimization process itself. This could be achieved by giving the robustness the position of an explicit criterion in the multi-objective function, or by including the robustness assessment as a part of the search procedure.

Hence, new optimization schemes are to be developed to address feature selection in this high-dimensional data which pushes our metaheuristic optimization methods to their limits.

REFERENCES

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences (PNAS),* 96(12):6745–6750, 1999.

[2] C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Science*s *(PNAS),* 99(10):6562–6566, 2002.

[3] R. E. Bellman. *Adaptive Control - A Guided Tour*. Princeton University Press, 1961.

[4] A. Berchuck, E. S. Iversen, J. M. Lancaster, J. Pittman, J. Luo, P. Lee, S. Murphy, H. K. Dressman, P. G. Febbo, M. West, et al. Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. *Clinical cancer research*, 11(10):3686–3696, 2005.

[5] J. M. Deutsch. Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics*, 19(1):45–52, 2003.

[6] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Sta-tistical Association*, 97(457):77–87, 2002.

[7] A. Dupuy and R. M. Simon. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute*, 99(2):147–157, 2007.

[8] V. Gardeux, R. Chelouah, P. Siarry, and F. Glover. Unidimensional search for solving continuous high-dimensional optimization problems. In *Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 1096–1101, Pisa, Italy, November 30 - December 2, 2009. IEEE Computer Society.

[9] V. Gardeux, R. Natowicz, R. Chelouah, R. Rouzier, A. Padua Braga, and P. Siarry. Un algorithme d'optimisation à haute dimension pour la fouille de données: méthode et application en onco-pharmacogénomique. In *Proceedings of Recherche Opérationnelle et Aide à la Décision*, Saint-Étienne, France, March 2-4, 2011.

[10] B. Ghattas and A. Ben Ishak. Sélection de variables pour la classification binaire en grande dimension: comparaisons et application aux données de biopuces. *Journal de la société française de statistique*, 149(3):43–66, 2008.

[11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[12] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[13] K. R. Hess, K. Anderson, W. F. Symmans, V. Valero, N. Ibrahim, J. A. Mejia, D. Booser, R. L. Theriault, A. U. Buzdar, P. J. Dempsey, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of clinical oncology*, 24(26):4236–4244, 2006.

[14] L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13550–13555, September 2005.

[15] R. Natowicz, R. Incitti, B. C. Euler Horta, P. Guinot, K. Yan, C. Coutant, F. Andre, L. Pusztai, and R. Rouzier. Prediction of the outcome of preoperative chemotherapy in breast cancer using DNA probes that provide information on both complete and incomplete responses. *BMC bioinformatics*, 9(149), 2008.

[16] C. Orsenigo. Gene selection and cancer microarray data classification via mixed-integer optimization. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, 4973:141–152, 2008.

[17] N. Pochet, F. De Smet, J. A. K. Suykens, and B. L. R. De Moor. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*, 20(17):3185–3195, 2004.

[18] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[19] A. Rakotomamonjy. Variable selection using SVM based criteria. *The Journal of Machine Learning Research*, 3:1357–1370, 2003.

[20] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub. A molecular signature of metastasis in primary solid tumors. *Nature genetics*, 33(1):49–54, 2002.

[21] S. Shah and A. Kusiak. Cancer gene search with data-mining and genetic algorithms. *Computers in Biology and Medicine*, 37(2):251–261, 2007.

[22] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, et al. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1):68–74, 2002.

[23] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95(1):14–18, 2003.

[24] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

[25] Adel Tabchy, Vicente Valero, Tatiana Vidaurre, Ana Lluch, Henry Gomez, Miguel Martin, Yuan Qi, Luis Javier Barajas-Figueroa, Eduardo Souchon, Charles Coutant, Franco D. Doimi, Nuhad K. Ibrahim, Yun Gong, Gabriel N. Hortobagyi, Kenneth R. Hess, W. Fraser Symmans, and Lajos Pusztai. Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response pre-dictor in a multicenter randomized trial in breast cancer. *Clinical Cancer Research*, 16(21):5351–5361, 2010.

[26] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. *The Journal of Machine Learning Research*, 3(7-8):1439–1461, 2003.