Interdisciplinary Sciences: Computational Life Sciences (2018) 10:12–23
https://doi.org/10.1007/s12539-017-0274-z

ORIGINAL RESEARCH ARTICLE

An Ensemble Framework Coping with Instability in the Gene Selection Process

José A. Castellanos-Garzón[1,2] · Juan Ramos[1] · Daniel López-Sánchez[1] · Juan F. de Paz[1] · Juan M. Corchado[1,3]

Received: 22 August 2017 / Revised: 6 November 2017 / Accepted: 8 November 2017 / Published online: 8 January 2018

Abstract

This paper proposes an ensemble framework for gene selection, which is aimed at addressing instability problems presented in the gene filtering task. The complex process of gene selection from gene expression data faces different instability problems from the informative gene subsets found by different filter methods. This makes the identification of significant genes by the experts difficult. The instability of results can come from filter methods, gene classifier methods, different datasets of the same disease and multiple valid groups of biomarkers. Even though there is a wide number of proposals, the complexity imposed by this problem remains a challenge today. This work proposes a framework involving five stages of gene filtering to discover biomarkers for diagnosis and classification tasks. This framework performs a process of stable feature selection, facing the problems above and, thus, providing a more suitable and reliable solution for clinical and research purposes. Our proposal involves a process of multistage gene filtering, in which several ensemble strategies for gene selection were added in such a way that different classifiers simultaneously assess gene subsets to face instability. Firstly, we apply an ensemble of recent gene selection methods to obtain diversity in the genes found (stability according to filter methods). Next, we apply an ensemble of known classifiers to filter genes relevant to all classifiers at a time (stability according to classification methods). The achieved results were evaluated in two different datasets of the same disease (pancreatic ductal adenocarcinoma), in search of stability according to the disease, for which promising results were achieved.

Keywords Gene selection · Filter method · Ensemble method · Wrapper method · Machine learning · Data mining · Gene expression data

1 Introduction

The study of gene expression data from the new chip technologies is of great interest for bioinformatics (and functional genomics), because they allow us to simultaneously analyze expression levels from hundreds of thousands of genes in a living organism sample. This feature makes gene expression analysis a fundamental tool of research for human health. It provides identification of new genes that are key factors in the genesis and development of diseases. However, the exploration of these large data sets is an important, yet difficult problem. The development of new hybrid methods fusing statistical, data mining and machine learning techniques to discover knowledge can help to face the challenges imposed by this technology [1, 2]

An area playing a major role in the analysis of gene expression data is gene selection. Gene selection involves the study of genes significant for a target disease. Such genes should be able to differentiate samples from different

✉ José A. Castellanos-Garzón
jantonio@usal.es
http://bisite.usal.es
https://www.cisuc.uc.pt/groups/show/ecos

Juan Ramos
juanrg@usal.es

Daniel López-Sánchez
lope@usal.es

Juan F. de Paz
fcofds@usal.es

Juan M. Corchado
corchado@usal.es
http://www.oit.ac.jp

[1] IBSAL/BISITE Research Group, University of Salamanca, Edificio I+D+i, 37007 Salamanca, Spain

[2] CISUC, ECOS Research Group, University of Coimbra, Pólo II-Pinhal de Marrocos, 3030-290 Coimbra, Portugal

[3] Osaka Institute of Technology, Osaka 535-8585, Japan

populations [3]. The discovery of these genes is the basis for the development of the diagnosis and prognosis methodologies of diseases. Hence, pharmaceutical companies are concerned about the identification of genes that can be modified across drugs [4–6]. Due to their importance, those genes are known as informative genes or differentially expressed genes, since they can differentiate distinct classes or subtypes of a target disease.

A common problem related to the large list of filter methods proposed in the literature is its instability [7, 8]. The instability problem considers that the significance of the discovered genes is closely related to the applied filter and classifier method, as well as specific characteristics of the dataset used. This indicates that when evaluating statistically significant genes in the laboratory, they are not really relevant to the target disease. Solutions to this challenge should be focused on assessing genes with respect to different classifiers or using measures regardless of the classification model [3, 9, 10]. Hence, our research consists of applying hybrid techniques to the gene selection process to build more stable solutions [5, 11, 12]. In this case, we have developed a five-staged ensemble framework, where each stage is responsible for carrying out a filtering process of genes significant for the next stage. The instability problem is faced by looking for a gene subset from the result of the previous stages, being able to simultaneously maximize the accuracy of a classifier set. The early stages of the framework are responsible for processing and removing noise from the dataset by using two ranking-based filter methods (simple methods). After that, different gene selection methods (compound methods) are applied to the result obtained from the stages above. The genes achieved by each applied method are combined in a single set to be subsequently filtered through two wrapper methods. At this point, we introduce two wrapper methods based on a list of different classifiers outlined in the framework. Then, a new gene set is created from the individual results reached by each combination of wrapper method and used classifier. Finally, we provide an algorithm to filter a gene subset from the set above by selecting a combination of *k* genes able to simultaneously maximize the accuracy of all classifiers. In addition, the results achieved for one of the used datasets were extended to other datasets of the same disease, for which it turned out that the genes discovered in the first dataset were also significant in the latter dataset.

The research developed in this area (gene selection) has generated a wide list of gene selection methods (filter methods), designed to discover informative gene subsets associated with a target annotation. For their better understanding, these methods have been divided into four main categories: filters, wrappers, embedded and ensemble [13–16]. Filter methods determine the relevance of features by ranking them based on statistical criteria, whereas wrappers use a classifier to determine feature sets with high discrimination power. Like wrappers, embedded methods are based on learning methods, but allowing to interact with them, which decreases the runtime taken by wrappers. Meanwhile, ensembles are the most recent among feature selection methods and merge different strategies to face instability problems presented by other methods due to data perturbations. Some of the most recent ensemble methods proposed in literature are: [17], which states an ensemble method called EGSG to select multiple gene subsets for classification purposes. The method selects salient gene subsets from gene expression data based on information theory and approximate Markov blanket. In [18], an ensemble of filters methods and classifiers has been proposed, where five methods of gene filtering are applied using different metrics. The result of each method is used to train a specific classifier and the outputs of the used classifiers are combined through a voting process. Recently in [19, 20], new ensemble approaches have been proposed. In [19], a feature selection method based on a bi-objective genetic algorithm has been proposed. Concepts of the theories of rough set and multivariate mutual information (information theory) are used as objectives in the fitness function of the genetic algorithm. The ensemble is built from different feature selectors based on the genetic algorithm to yield a much generalized feature subset. In [20], two gene selection approaches are described. The approach of homogeneous distributed ensemble deals with generating *n* models using the same feature selection method but different training data, whereas the heterogeneous centralized ensemble deals with *n* models generated using different feature selection methods, but the same training data. Unlike previous works, our framework provides an approach of successive reduction based on different linked filtering stages, which are responsible for reaching different stability levels about filter methods, classification methods and the data domain. Additionally, the framework has been designed in a flexible way, in the sense that new gene selection and classification methods can be added to improve the global filtering process.

1.1 Pancreatic Ductal Adenocarcinoma as Experimental Data

Pancreatic ductal adenocarcinoma (PDAC) has been considered as one of the most aggressive types of cancer [21], having a 5-year survival rate of 8% [22]. Pancreatic cancers are usually asymptomatic in early stages, obstructing early detection and, thus, contributing to low survival. Moreover, the available chemotherapeutic drugs exhibit little effectiveness in PDAC. This phenomenon has been related to the dynamic relation between the stroma and tumor cells [23]. PDAC originates due to a successive accumulation of mutations affecting different oncogenes and tumor suppressors. Many of these genes have a major role in key signaling pathways. Genes such as RAS, AKT, CDKN2A, TP53 and DPC4, among others, are affected by punctual mutations or allelic loss in pancreatic cancer [24, 25]. In addition to all the above, PDAC has been identified as one of the cancers with high drug resistance. In this regard, we have that the desmoplastic stroma constitutes a protective barrier against drugs, which hinders the effectiveness of any medical treatment applied.

Pancreatic ductal adenocarcinoma is a particularly unstable cancer from the molecular point of view. Unfortunately, gene expression datasets often present problems such as differences in methodology applied to their preparation, low sample size, unbalanced classes, and missing data, among others. Thus, genes obtained through gene selection methods present low generalization capability for classification under different datasets. In consequence, there is a wide range of approaches for obtaining a reduced set of biomarkers potential for a given disease. However, by considering the biological properties of PDAC and its instability, ensemble approaches qualify for finding more consistent signatures by involving enough genes to classify different datasets. This allows us to select a reduced subset from the signature in accordance with the datasets used. The current proposal faces the mentioned problems by finding a stable group of biomarkers to simultaneously optimize classification from different datasets. For this purpose, two different datasets have been used to select a biomarker set by means of several filtering stages.

To reach the goals of this research, the rest of this paper has been structured as follows. Section 2 explains the main features of the proposed framework and the filtering process in stages. Section 3 outlines the experiments developed to evaluate the performance of the framework for two datasets of the same disease (PDAC). Section 4 deals with the conclusions of this research, whereas the used references have been listed at the end of this paper.

2 An Ensemble Proposal for Successive Filtering of Relevant Genes (EF-GMS)

This section deals with the explanation of each component of our gene selection proposal, which focuses on five linked stages, each developing a different task of gene filtering aimed at the next stage, until reaching the expected result. Figure 1 shows a general scheme of the steps developed by our approach, EF-GMS. By way of summary, we can say that our proposal consists of a starting stage (Stage-I), where different tasks of data preprocessing are applied to the input raw dataset. The following stage (Stage-II) oversees removing noise from the resulting data. Stage-III is responsible for running a set of gene selection methods on its input data and joining the results in a single set. Stage-IV finds gene subsets from its input set by maximizing the accuracy of each classifier separately. For this end, two wrapper methods are run by using each classifier given in the framework. The gene subsets achieved for each classifier are joined in a single set. Finally, Stage-V oversees stability in the gene set of the stage above. To do this, it runs an algorithm which can choose a combination of k genes, simultaneously maximizing the accuracy of all involved classifiers. Note that the idea pursued by this framework is to successively reduce the initial dataset, based on different criteria to achieve generic genes, until reaching a small subset of relevant genes. We will describe in detail each stage shown in Fig. 1.

2.1 Stage-I: Data Preprocessing

As previously explained, this stage is responsible for cleaning the input raw data, where different data treatments are applied, so that this stage prepares the data for the next stages to use them. The tasks involved in this stage among others are removing control and constant probes, missing value treatment and data normalization, if needed. Once all needed processes have been carried out, a new subdataset is built for the next stage.

2.2 Stage-II: Noise Removing Methods

This stage is responsible for removing noise in the data and involving two gene filter methods, which are linked to make a double filter process through different techniques. By applying the Mann–Whitney test to the input dataset as the first filter method, we will have a gene significance test, relating genes to the studied disease. The Mann–Whitney test is commonly used in the literature as a filter method to filter out differentially expressed genes according to tissue sample classes [26]. Moreover, this test does not assume a specific data distribution (nonparametric test) as is the case of other tests. This test relates samples belonging to the same population to the null hypothesis, whereas samples belonging to different population are related to the alternative hypothesis [27]. The Mann–Whitney test establishes a ranking of significant values (p values) for the genes of a dataset. Hence, genes with p value < 0.05 are taken out as genes rejecting the null hypothesis and so they are the most statistically significant. Therefore, such genes are passed to the following filter method of this stage, S2N.

S2N (Signal to noise [3, 28]) is the second filter method applied to reduce noise from the input dataset to this stage. S2N computes the existing correlation of each gene according to the positive and negative class of the dataset. Hence, this statistic assigns positive values to genes correlated with the positive class, whereas negative values are assigned to genes correlated with the negative class of the dataset. To select the most significant genes related

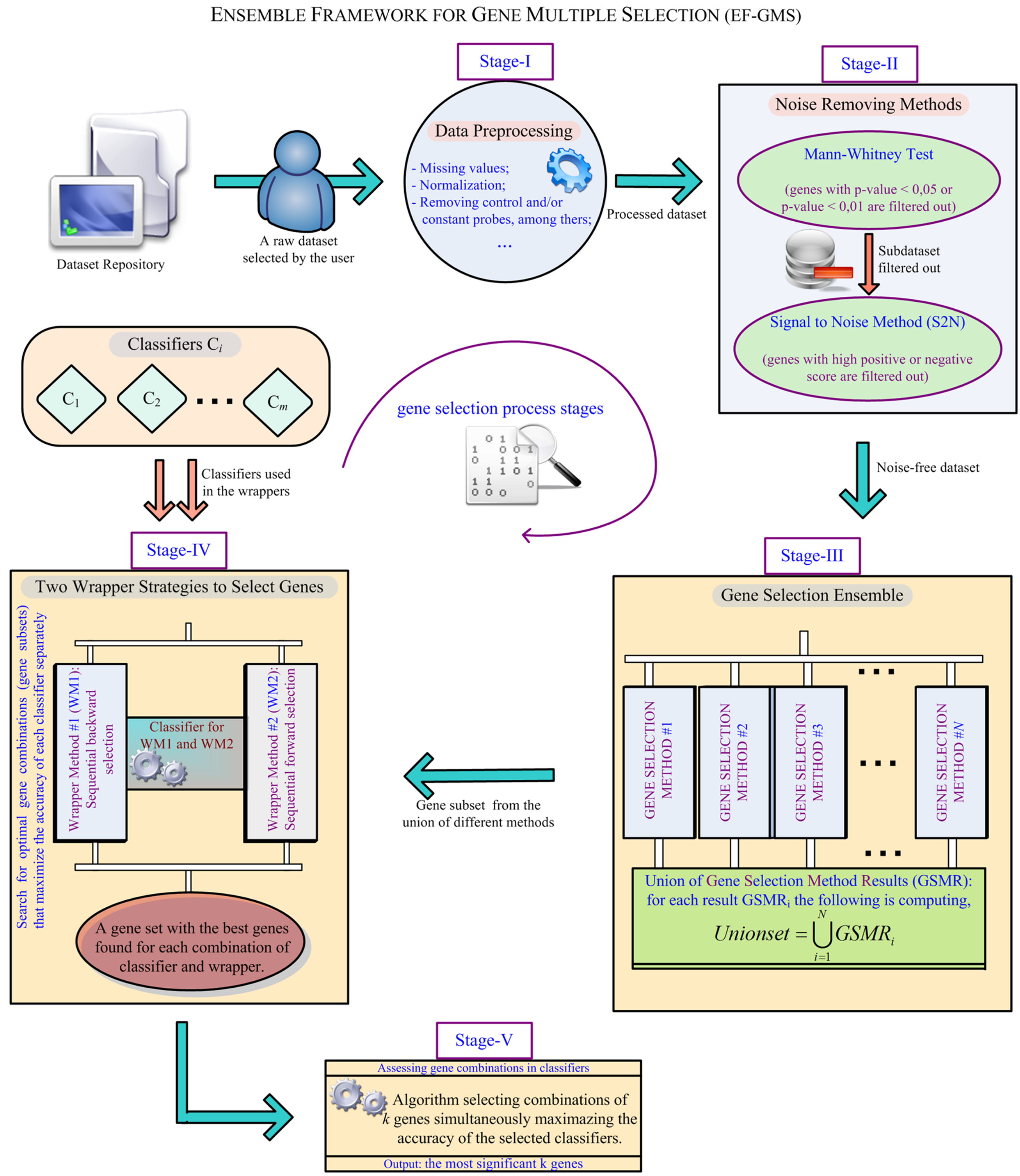
ENSEMBLE FRAMEWORK FOR GENE MULTIPLE SELECTION (EF-GMS)
Stage-I
Stage-II
Dataset Repository
A raw dataset selected by the user
Data Preprocessing
- Missing values;
- Normalization;
- Removing control and/or constant probes, among thers;
...
Processed dataset
Noise Removing Methods
Mann-Whitney Test
(genes with p-value < 0,05 or p-value < 0,01 are filtered out)
Subdataset filtered out
Signal to Noise Method (S2N)
(genes with high positive or negative score are filtered out)
Classifiers C_i
C_1
C_2
...
C_m
Classifiers used in the wrappers
gene selection process stages
Noise-free dataset
Stage-IV
Stage-III
Two Wrapper Strategies to Select Genes
Search for optimal gene combinations (gene subsets) that maximize the accuracy of each classifier separately
Wrapper Method #1 (WM1): Sequential backward selection
Classifier for WM1 and WM2
Wrapper Method #2 (WM2): Sequential forward selection
A gene set with the best genes found for each combination of classifier and wrapper.
Gene subset from the union of different methods
Gene Selection Ensemble
GENE SELECTION METHOD #1
GENE SELECTION METHOD #2
GENE SELECTION METHOD #3
...
GENE SELECTION METHOD #N
Union of Gene Selection Method Results (GSMR): for each result $GSMR_i$ the following is computing,
$Unionset = \bigcup_{i=1}^{N} GSMR_i$
Stage-V
Assessing gene combinations in classifiers
Algorithm selecting combinations of k genes simultaneously maximazing the accuracy of the selected classifiers.
Output: the most significant k genes

Fig. 1 Flowchart representing the five linked stages of the gene selection process of the EF-GMS framework: data preprocessing, noise removing, gene selection ensemble, gene filtering with two wrapper strategies and final selection of a stable gene subset

to both classes, a threshold should be fixed in such a way that genes with big positive and negative values can be filtered out. The result of this method is a new dataset (without noise) whose genes can be considered significant for PDAC and, therefore, the following stages of the framework can continue to apply their filtering processes based on other criteria to ensure stability in the final result.

2.3 Stage-III: Gene Selection Ensemble

This stage is in charge of carrying out a filtering process based on a set of gene selection methods (compound filter methods) predefined in the framework. Additionally, the framework can add new filter methods to improve the results toward those with more diversity. Therefore, this stage is assumed as an ensemble method where each filter method is individually applied to the input dataset. Then, the individual result of each method is added to a single set (called Unionset) through the union operation defined for sets. Hence, this stage returns a gene subset holding different selection strategies and criteria to ensure diversity. In consequence with the above, the goal of building Unionset is to find a gene combination from this set, which can be representative of the dataset and maximize the accuracy of different classifiers used in the study of the disease represented. This will be the task to perform by the next stages of the framework. Thus, the next step to develop in Stage-IV is to select gene subsets from Unionset across wrapper strategies, which maximize the accuracy of each classifier given by the framework.

2.4 Stage-IV: Two Wrapper Methods

This stage is in charge of finding a small gene subset from Unionset for each classifier used. Genes in each subset must maximize the accuracy of each classifier separately. To find such gene subsets, two greedy strategies acting as wrapper methods have been implemented. We have implemented a gene removing strategy (known as backward gene selection), which we call WM1, and gene addition strategy (known as forward gene selection), which we call WM2. Both strategies are run for all existing classifiers to maximize their accuracies and obtain the corresponding gene subsets.

Note that this stage has a list of classifiers to be used by the wrapper methods. So the idea pursued with this stage is to find a gene subset from Unionset for each wrapper (WM1 and WM2) and each classifier in the list. Each subset must maximize the accuracy of its corresponding classifier. This way, we can create a single set by joining all subsets found from WM1 and WM2, which will have the best genes for each classifier. The operation mode developed by both strategies (WM1 and WM2) is presented as follows:

- Backward selection, WM1: This method starts from both, a gene subset (in this case, Unionset) and a classifier as input. Then, the method iteratively removes a gene from Unionset and evaluates the accuracy of the remaining genes in Unionset by using the input classifier. If the accuracy for the new subset is greater than the previous subset, then the new subset replaces the previous Unionset. Otherwise, the gene removed from Unionset is returned (because it is a significant gene) and another gene is selected to be removed from Unionset. The processes above are repeated until all genes in Unionset have been selected to be removed. At the end of the algorithm, only a subset of genes meaningful to the classifier used will maximize the precession of the classifier. Note that in our case, WM1 will be run for each classifier existing in the classifier list of the framework. Thus, a different gene subset from Unionset will be obtained for each classifier.
- Forward selection, WM2: As in WM1, WM2 starts from a gene subset (in this case, Unionset) and a classifier as its input. Unlike WM1, this strategy selects a gene from Unionset to be added to a set *NS* in each iteration. *NS* is an empty set in the first iteration of the algorithm. The gene added to *NS* in each iteration must be such that together with the rest of genes in *NS*, it maximizes the accuracy of the input classifier. That is, the *NS* accuracy along with the new gene added from Unionset must be greater than the *NS* accuracy without that gene. Otherwise, the selected gene is not added to *NS* and another gene should be selected from Unionset. The algorithm ends when no new gene can be added to *NS* in such a way that its accuracy improves. At this point, *NS* will have a gene subset from Unionset whose genes will be the most significant according the input classifier. Also note that for each classifier outlined in the framework, an *NS* subset is achieved, whose gene selection from all subsets will be solved in the next stage.

2.5 Stage-V: An Algorithm Looking for a Stable Gene Subset

Starting from the genes in the subset returned by the stage above are associated with the classifiers used by the wrapper methods, the goal of this stage is to find a stable gene subset able to simultaneously maximize the accuracy of all given classifiers. To do this, we have developed an algorithm (Algorithm 1) that given a k, it carries out a search for all k gene combinations from the input gene subset to find one combination maximizing the accuracy of all classifiers at the same time.

Note that for the search for a stable subset be effective, the gene set found in the stage above should be small whereas k should be much smaller than the size of the gene set. Our framework has been defined to reduce the input dataset as much as possible to insure the conditions above are met. Thus, this stage returns k stable genes as the result.

2.5.1 GeneCombine Algorithm

Input: T, a gene subset of the current dataset. k, the number of genes to select from T and L, a list of gene classifier methods.
Output: $\langle S, ma \rangle$, where S is a subset of k genes from T and ma is the mean accuracy from the classifier accuracies given in L.
Required: Accuracy function, computes the mean accuracy for a classifier trained from a gene subset by applying a stratified 10-fold cross-validation.

```
$S := \emptyset, ma := 0$;
% Select each different gene subset of size $k$ from $T$.
for each subset $P_k$ of $k$ genes from $T$ do
    $sum := 0$;
    % Evaluate the accuracy of $P_k$ for each classifier in $L$.
    for each classifier $C_i$ in $L$ do
        % Add the accuracy of each classifier applied to $P_k$ to $sum$.
        $sum := sum + Accuracy(C_i(P_k))$;
    endfor
    % Compute the mean accuracy of the classifiers evaluated on $P_k$.
    $mean := \frac{sum}{length(L)}$;
    % Update the accuracy and the gene subset.
    if $mean > ma$ then
        $ma := mean$;
        $S := P_k$;
        % If the mean accuracy is the maximum accuracy (100%) then the
        % algorithm ends.
        if $ma = 100$ then break endif
    endif
endfor
end.
```

3 Experimental Results

This current study has been developed for two Pancreas datasets (pancreatic ductal adenocarcinoma, PDAC), which we call PDAC#1 and PDAC#2. The proposed framework is basically applied to the first dataset (PDAC#1). After that, the PDAC#1 results are validated in PDAC#2 to discover genes significant for both datasets and, in general, genes relevant for the disease in question.

Therefore, the goal of this experiment is to assess our proposal in the process of discovering informative genes able to face the instability problem from different gene selection methods, classifiers and datasets. Thus, the significance of the genes found by our proposal from PDAC#1 in classification tasks is extended to the other dataset of the same disease (in this case, PDAC#2) to assess generality in the results. The latter deals with evaluating in PDAC#2, the same genes discovered by the EF-GMS framework (Fig. 1) in PDAC#1. We are also concerned about how the accuracy of a gene subset from a dataset is affected when it is evaluated in other dataset of the same disease but with very different features, which is a challenge in the literature.

3.1 Datasets of Pancreatic Ductal Adenocarcinoma

The two PDAC datasets used in this research have been extracted from the public repository of the National Center for Biotechnology Information (NCBI), https://www.ncbi.nlm.nih.gov/. The PDAC#1 dataset is composed of 54,675 gene probes evaluated under 25 tumor tissue samples plus 7 normal tissue samples, whereas the PDAC#2 dataset is composed of 54,675 gene probes evaluated under 39-paired tumor and non-tumor tissue samples. Summarizing, PDAC#1 has a gene expression matrix of size 54,675 × 32 whereas PDAC#2 has one of size 54,675 × 78. Both datasets are available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32676 and http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15471, respectively.

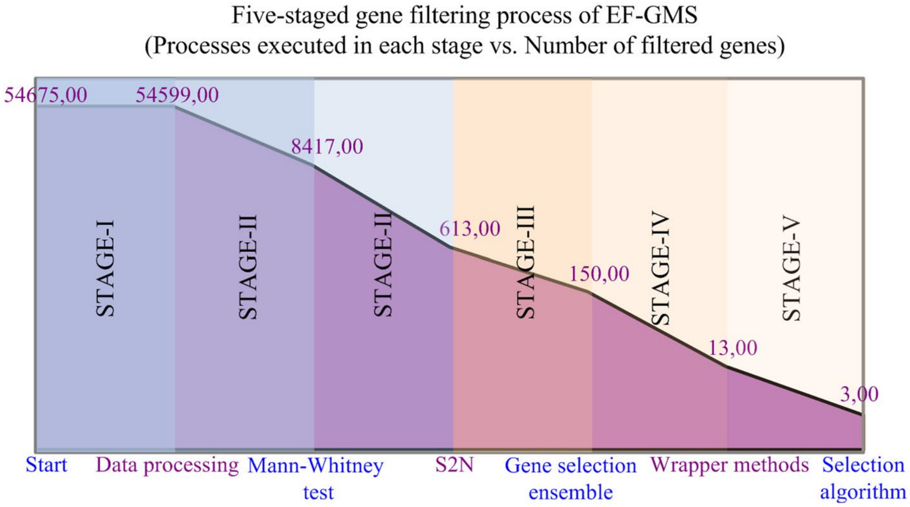

Fig. 2 EF-GMS chart summarizing the gene filtering process involved in each stage and the remaining number of genes when the tasks in each stage are applied to PDAC#1

3.2 Results from PDAC#1

This subsection shows the results reached in each stage of our framework (Fig. 1) for the PDAC#1 dataset. Additionally, Figure 2 presents a global view of the number of genes remaining after each filtering process. This figure shows the five-staged reduction progress of the gene number of PDAC#1 until reaching the final result in Stage-V, 3 informative genes. Reinforcing the information given in Fig. 2, Table 1 lists all stages of the framework along with the number of genes taken out and the reduction percentage (with respect to the stage above) of the remaining dataset after applying the filtering process in each stage. The filtering results of EF-GMS for genes in PDAC#1 is as follows:

- Stage-I: Once the data processing has been applied to the original dataset (54,675 probes), 54,599 probes are extracted to a new dataset, which will be the input to the next stage.
- Stage-II: This stage applies two linked filter methods to the input dataset (54,599 gene probes) to reduce noise. The first filter method applied is the Mann–Whitney test, which computes the p value related to each gene to establish a ranking and select those genes having *p* value below 0.05. In this case, the Mann–Whitney test achieved a new dataset composed of 8417 gene probes, which is passed to the second filter method, S2N. The S2N method computes and assigns positive or negative values to genes in accordance with their degrees of belonging to the positive or negative sample class. Thereafter, the genes obtaining the highest positive and negative values are selected as the most significant genes for each class. The threshold stated for both sample classes to choose the significant genes has been fixed in the middle point of the values computed from the genes in each class. Hence, according to the input dataset. S2N achieved a dataset with 613 gene probes as the result of this stage.
- Stage-III: Once the noise in the dataset has been removed, this stage is responsible for running different gene selection methods on the input dataset (613 gene probes, in this case) to achieve gene subsets holding different criteria. The result of each run method is added to a Unionset set through the union operation. Six methods have been used in this framework, which are: kofnGA in [29, 30], Boruta in [31, 32], propOverlap in [33, 34], SDA in [35, 36], Spikeslab in [37, 38] and SubLasso in [39, 40]. For the case of kofnGA which is a genetic algorithm, its main parameters have been initialized as follows: population size = 100, number of generations = 50,000, the fitness function used has been correlation between the genes, the remaining parameters have been initialized as stated by the method. The individual results of these methods for PDAC#1 have been listed in Table 2. This table shows the name of the method applied, number of genes found and the mean accuracy through a stratified tenfold cross-validation of such genes for the three classifiers outlined in the framework by the next stage. The used classifiers are [41, 42]: SVM: linear support vector machine, which finds the best hyperplane separating both classes; naive-Bayes: this model computes the probability of each class given the values of all attributes and assuming the attribute conditional independence; kNN: *k*-nearest neighbor classification. This is a lazy model which classifies the input pattern by using its *k*-nearest neighbors from the training set. Then, after carrying out the union of results, a new dataset with 150 gene probes was obtained and identified as Unionset. The mean accuracy for this last dataset has also been given in the table.

Table 1 Comparative table of the number of genes taken out and the reduction percentage of the remaining dataset (PDAC#1) in each filtering stage of the framework

Filtering and reduction stages	Number of genes taken out	Reduction percentage (%)
Stage-I	76	0.15
Stage-II (Mann–W. test)	46,182	84.58
Stage-II (S2N)	7804	92.73
Stage-III	463	75.53
Stage-IV	137	91.33
Stage-V	10	76.92

The reduction percentage is computed with respect to the size of the dataset achieved in the stage above

Table 2 Comparative table of the gene selection methods applied to PDAC#1

Method	Number of genes	SVM (%)	naiveBayes (%)	*k*	kNN (%)
KofnGA	20	100	100	1	96.87
Boruta	41	96.87	93.75	2	96.87
propOverlap	2	90.62	90.62	3	93.75
SDA	20	100	93.75	2	96.87
Spikeslab	100	93.75	93.75	4	96.87
SubLasso	6	96.87	93.75	1	93.75
Unionset	150	100	100	3	96.87

The methods used, their number of genes found and their accuracy for tree classifiers are listed. The accuracy by union of the results of these methods is also listed

– Stage-IV: This stage deals with two wrapper methods (WM1 and WM2) and a list of classifiers (three classifiers in this experiment: SVM, naiveBayes and kNN) to be used by each wrapper. The goal is to reduce the number of genes given by the input dataset (150 gene probes) while increasing the accuracy of the classifiers through the wrapper methods. Therefore, six gene subsets have been yielded, one for each combination wrapper–classifier. Both the number of genes and the accuracy reached for each combination above have been listed in Table 3 (for methods WM1 and WM2). There is an accuracy value italicized for each row of the table, which means that the wrapper method of that row has been run by using the classifier of the column corresponding to the italicized value. The remaining accuracy values on the same row have been obtained by using the genes found by the classifier whose value has been italicized. Note that the best scores have been reached for combinations WM1–naiveBayes and WM2–naiveBayes. As a final step of this stage, a single subset with 13 genes has been created by union of the genes listed in Table 3, which is passed to the final stage.

– Stage-V: This stage is responsible for finding a informative gene subset able to simultaneously maximize the accuracy for the three classifiers used in the framework. The idea is to search stable subsets with respect to different classifiers to find genes regardless of the methodology applied. In this case, 13 genes have been taken as input to this stage and we have taken $k = 3$, i.e., to find combinations of 3 genes from the set of 13 genes. Note that a strategy to define the value of k is to run the search algorithm for $k = 2$ and increase this value until finding a gene combination with a mean accuracy from the three classifiers less than the mean accuracy of the gene combination previously found. We have selected $k = 3$ by exploration of a parallel coordinate chart for the 13 genes processed in this stage. Figure 3 shows the parallel coordinate chart representing tissue samples vs. gene expression levels of PDAC#1, where each curve represents the profile of the corresponding gene. After that, the algorithm was run for $k = 3$, finding a gene combination reaching the maximal accuracy for the three classifiers, as shown in the last row of Table 3 (Stage-V algorithm). The three found genes are: *KDM6B, LOC100507632* and *SOX4*. This result proves the stability of such genes across different classifiers. Now, it remains to show that those genes are also stable when they are selected and evaluated from another dataset of the same disease, i.e., dataset PDAC#2 (Table 4).

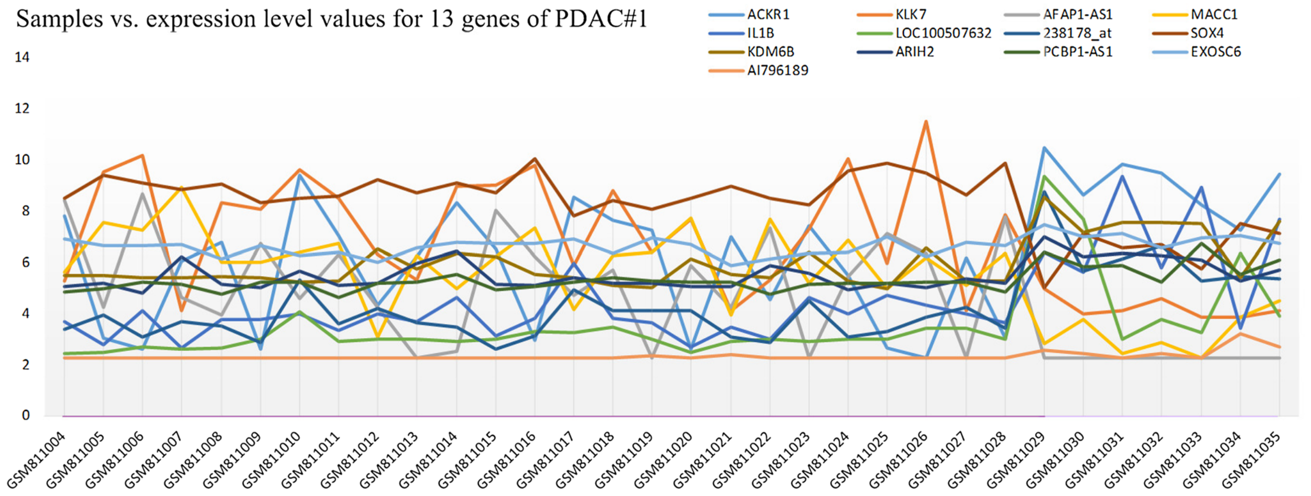

Fig. 3 Parallel coordinate chart associating tissue sample with gene expression level for 13 genes significant for PDAC#1

Table 3 Comparative table for the number of genes and accuracy reached by each wrapper for each classifier

Method	Number of genes	SVM (%)	naiveBayes (%)	*k*	kNN (%)
WM1	2	100	96.87	1	96.87
	5	96.87	100	1	100
	2	96.87	93.75	1	100
WM2	1	100	78.12	3	100
	3	93.75	100	3	100
	1	100	78.12	3	100
Stage-V algorithm	3	100	100	1	100

3.3 Evaluating in PDAC#2, the Genes Found in PDAC#1

An important goal of this research has been to evaluate the stability of the discovered genes with respect to different datasets of the same disease. In this sense, we have assessed the variation of accuracy for the genes found in the PDAC#1 dataset with respect to the same genes taken from the PDAC#2 dataset, representing the same cancer. Hence, this will allow us to appreciate the generality of the results obtained. The result of this test has been listed in Table 4. As shown, the best results have been obtained by our proposal, EF-GMS. Even though the accuracy for all methods has decreased because both datasets have very different features, the accuracy values for our proposal are above 87%, reaching a maximum value of 92.30% for the SVM classifier. This proves that the result achieved by EF-GMS in PDAC#1 is also significant in PDAC#2, which shows stability in the found genes across from different datasets of the same disease. Reinforcing the final results given in both datasets, Figure 4 displays two coordinate parallel charts representing the profile of the three genes found for both datasets.

3.4 Biological Insight

As a biological interpretation of some of the genes found and to analyze their involvement in the carried out experiment, we have revised their roles, in this case for SOX4 and KDM6B. That is, gene SOX4 (sex-determining region Y-related high-mobility-group box transcription factor 4) encodes a transcription factor that plays an important role during embryogenesis regulating development in different tissues [43]. This gene has been found to be deregulated in several types of cancer, in addition to being involved in increased cell proliferation, cell survival and apoptosis inhibition, epithelial-to-mesenchymal transition and metastasis. However, SOX4 has been reported as a tumor suppressor gene depending on the cellular context.

Meanwhile, the low survival imposed by PDAC is due to its invasiveness and the role played by desmoplastic stroma.

Table 4 Comparative table of gene selection methods for PDAC#2

Method	Number of genes	SVM (%)	naiveBayes (%)	*k*	kNN (%)
KofnGA	20	87.18	87.18	3	82.05
Boruta	41	88.46	84.62	4	91
propOverlap	2	69.23	71.79	5	78.20
SDA	20	84.61	82.05	1	83.33
Spikeslab	100	89.74	84.61	3	88.46
SubLasso	6	83.33	78.21	4	83.33
Unionset	150	89.74	85.90	2	89.74
EF-GMS: stage-V algorithm	3	92.30	87.18	1	91.03

The genes discovered by the methods given in Table 2 for PDAC#1 have been selected from PDAC#2 and evaluated on the three selected classifiers. The genes selected as the final result of EF-GMS in Table 3 have also been evaluated in PDAC#2

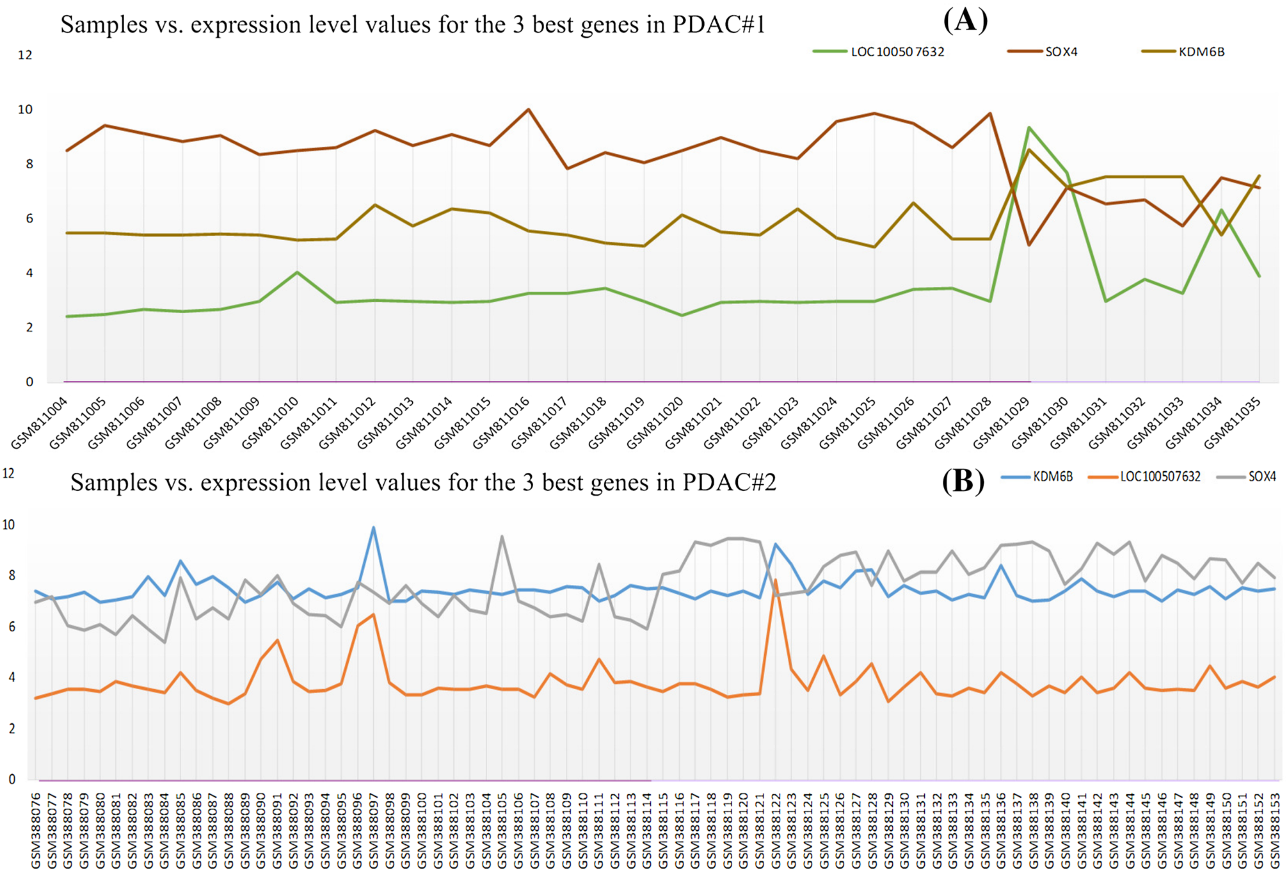

Fig. 4 Parallel coordinate charts associating tissue sample with gene expression level for the 3 genes relevant for both PDAC#1 and PDAC#2

The implication of EMT (a process that) is closely related to these phenomena. Considering the above, it is not surprising that SOX4 has been used as a PDAC biomarker in combination with other genes and found to be closely related to clinical outcome in patients [44].

Gene KDM6B encodes a histone demethylase that demethylates *Lys-2* of histone H3, activating gene expression. Histone methylation is an epigenetic modification playing a critical role in expression regulation of concrete genes and, thus, in cancer development. This gene is downregulated under numerous cancers and considered as a tumor suppressor in pancreatic cancers. However, this gene has been suggested to induce EMT in other cancer types, such as the case of renal cancer [45, 46].

4 Conclusions

This paper has presented EF-GMS, an ensemble framework for gene selection by considering the gene instability problem. Thus, the overall goal of our proposal in this research has been to provide a methodology able to find stable genes across from both different classifiers and datasets of the same disease, which is a challenge today. To do this, we have provided a framework of hybrid techniques by successively filtering in stages significant genes until reaching a small gene subset stable under different conditions.

Meanwhile, the goal of the proposed study has been to evaluate and compare the results of our proposal with respect to other methods in classification tasks from one of the datasets and generalize such results to the other dataset. The results achieved on the two datasets of pancreatic ductal adenocarcinoma (PDAC) have been very promising compared to other gene selection methods. Finally, the three genes discovered by our approach, *KDM6B, LOC100507632* and *SOX4* can be researched to gain insight into their behaviors.

References

1. Bourne PE, Wissig H (2003) Structural bioinformatics. Wiley-Liss Inc, Hoboken
2. Jiang D, Tang C, Zhang A (2004) Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng 16(11):1370–1386
3. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, deSchaetzen V, Duque R, Bersini H, Nowé A (2012) A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinform 9(4):1106–1118
4. Inza I, Larrañaga P, Blanco R, Cerrolaza A (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 31:91–103
5. Jager J, Sengupta R, Ruzzo W (2003) Improved gene selection for classification of microarrays. In: Pacific symposium on biocomputing (UW CSE Computational Biology Group)
6. Kumari B, Swarnkar T (2011) Filter versus wrapper feature subset selection in large dimensionality microarray: a review. Int J Comput Sci Inf Technol (IJCSIT) 2(3):1048–1053
7. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y (2009) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26(3):392–398
8. He Z, Yu W (2010) Stable feature selection for biomarker discovery. Comput Biol Chem 34(4):215–225
9. Xue B, Zhang M, Browne W, Yao X (2016) A survey on evolutionary computation approaches to feature selection. IEEE Trans Evol Comput 20(4):606–626
10. Yang P, Hwa Y, Zhou B, Zomaya A (2016) A review of ensemble methods in bioinformatics: including stability of feature selection and ensemble feature selection methods. Bioinformatics 4:296–308
11. Baruque B, Corchado E, Mata A, Corchado JM (2010) A forecasting solution to the oil spill problem based on a hybrid intelligent system. Inf Sci 180(10):2029–2043
12. Guyon I (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
13. Natarajan A, Ravi T (2014) A survey on gene feature selection using microarray data for cancer classification. Int J Comput Sci Commun (IJCSC) 5(1):126–129
14. Shraddha S, Anuradha N, Swapnil S (2014) Feature selection techniques and microarray data: a survey. Int J Emerg Technol Adv Eng 4(1):179–183
15. Tyagi V, Mishra A (2013) A survey on different feature selection methods for microarray data analysis. Int J Comput Appl 67(16):36–40
16. Wang Y, Tetko I, Hall M, Frank E, Facius A, Mayer K, Mewes H (2005) Gene selection from microarray data for cancer classification—a machine learning approach. Comput Biol Chem 29:37–46
17. Liu H, Liu L, Zhang H (2010) Ensemble gene selection by grouping for microarray data classification. J Biomed Inform 43:81–87
18. Bol'on-Canedo V, Sánchez-Marońo N, Alonso-Betanzos A (2012) An ensemble of filters and classifiers for microarray data classification. Pattern Recognit 45:531–539
19. Das A, Das S, Ghosha A (2017) Ensemble feature selection using bi-objective genetic algorithm. Knowl Based Syst 118:124–139
20. Seijo-Pardo B, Porto-Daz I, Boln-Canedo V, Alonso-Betanzos A (2017) Ensemble feature selection: homogeneous and heterogeneous approaches. Knowl Based Syst 123:116–127
21. Badea L, Herlea V, Olimpia S, Dumitrascu T, Popescu I (2008) Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. Hepato-Gastroenterology 88:2015–2026
22. Kota J, Hancock J, Kwon J, Korc M (2017) Pancreatic cancer: stroma and its current and emerging targeted therapies. Cancer Lett 391:38–49
23. Bhaw-Luximon A, Jhurry D (2015) New avenues for improving pancreatic ductal adenocarcinoma (pdac) treatment: selective stroma depletion combined with nano drug delivery. Cancer Lett 369(2):266–273
24. Hidalgo M, Cascinu S, Kleeff J, Labianca R, Löhr JM, Neoptolemos J, Real FX, Van Laethem JL, Heinemann V (2015) Addressing the challenges of pancreatic cancer: future directions for improving outcomes. Pancreatology 15(1):8–18
25. Korc M (2007) Pancreatic cancer-associated stroma production. Am J Surg 194(4):S84–S86
26. Fang Z, Du R, Cui X (2012) Uniform approximation is more appropriate for Wilcoxon rank-sum test in gene set analysis. PLoS One 7(2):e31,505
27. Weiss P (2005) Applications of generating functions in nonparametric tests. Math J 9(4):803–823
28. Berrar DP, Dubitzky W, Granzow M (2003) A practical approach to microarray data analysis. Kluwer Academic Publishers, New York
29. Wolters M (2015) A genetic algorithm for fixed-size subset selection. R-Package kofnGA, Version 1.2
30. Wolters M (2015) A genetic algorithm for selection of fixed-size subsets with application to design problems. J Stat Soft 68(1):1–18
31. Kursa M, Rudnicki W (2010) Feature selection with the Boruta package. J Stat Softw 36(11):1–13
32. Kursa M, Rudnicki W (2016) Wrapper algorithm for all relevant feature selection. Package Boruta, Version 5.1.0. https://m2.icm.edu.pl/boruta/

33. Mahmoud O, Harrison A, Perperoglou A, Gul A, Khan Z, Metodiev M, Lausen B (2014) A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. BMC Bioinform 15(274):1–20
34. Mahmoud O, Harrison A, Perperoglou A, Gul A, Khan Z, Lausen B (2015) propOverlap: feature (gene) selection based on the proportional overlapping scores. R package version 1.0. http://CRAN.R-project.org/package=propOverlap
35. Ahdesmaki AKS (2010) Feature selection in omics prediction problems using CAT scores and false non-discovery rate control. Ann Appl Stat 4:503–519
36. Ahdesmaki M, Zuber V, Gibb S, Strimmer K (2015) sda: shrinkage discriminant analysis and CAT score variable selection. R package version 1.3.7. http://CRAN.R-project.org/package=sda
37. Ishwaran H, Rao J (2005) Spike and slab variable selection: frequentist and Bayesian strategies. Ann Stat 33(2):730–773
38. Ishwaran H, Rao J, Kogalur UB (2013) spikeslab: prediction and variable selection using spike and slab regression. R package version 1.1.5. http://web.ccs.miami.edu/~hishwaran. http://www.kogalur.com
39. Friedman J, Hastie T, Tibshirani R (2008) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22. http://www.stanford.edu/~hastie/Papers/glmnet.pdf
40. Zhou F, Luo Y, Meng Q, Ge R, Mai G, Liu J (2015) Sublasso: gene selection using lasso for microarray data with user-defined genes fixed in model. R-Project, package version 1.0
41. Flach P (2012) Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press, Cambridge
42. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan G, Ng A, Liu B, Yu P, Zhou ZH, Steinbach M, Hand D, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37
43. Vervoort S, Boxtel V, Coffer P (2013) he role of sry-related hmg box transcription factor 4 (sox4) in tumorigenesis and metastasis: friend or foe? Oncogene 32(29):339–409. https://www.ncbi.nlm.nih.gov/pubmed/23246969
44. Hasegawa S, Nagano H, Konno M, Eguchi H, Tomokuni A, Tomimaru Y, Asaoka T, Wada H, Hama N, Kawamoto K, Marubashi S, Nishida N, Koseki J, Mori M, Doki Y, Ishii H (2016) A crucial epithelial to mesenchymal transition regulator, sox4/ezh2 axis is closely related to the clinical outcome in pancreatic cancer patients. Int J Oncol 48(1):145–152. https://www.ncbi.nlm.nih.gov/pubmed/26648239
45. Li Q, Hou L, Ding G, Li Y, Wang J, Qian B, Sun J, Wang Q (2015) Kdm6b induces epithelial-mesenchymal transition and enhances clear cell renal cell carcinoma metastasis through the activation of slug. Int J Clin Exp Pathol 8(6):6334–6344. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525843/
46. Yamamoto K, Tateishi K, Kudo Y, Sato T, Yamamoto S, Miyabayashi K, Matsusaka K, Asaoka Y, Ijichi H, Hirata Y, Otsuka M, Nakai Y, Isayama H, Ikenoue T, Kurokawa M, Fukayama M, Kokudo N, Omata M, Koike K (2014) Loss of histone demethylase KDM6B enhances aggressiveness of pancreatic cancer through downregulation of c/ebp. Carcinogenesis 35(11):2404–2414. https://www.ncbi.nlm.nih.gov/pubmed/24947179