Interdiscip Sci Comput Life Sci (2017) 9:1–13
DOI 10.1007/s12539-017-0219-6

ORIGINAL RESEARCH ARTICLE

An Agent-Based Clustering Approach for Gene Selection in Gene Expression Microarray

Juan Ramos[1] · José A. Castellanos-Garzón[1,2] · Alfonso González-Briones[1] · Juan F. de Paz[1] · Juan M. Corchado[1,3]

Received: 18 July 2016 / Revised: 26 January 2017 / Accepted: 7 February 2017 / Published online: 9 March 2017

Abstract Gene selection is a major research area in microarray analysis, which seeks to discover differentially expressed genes for a particular target annotation. Such genes also often called informative genes are able to differentiate tissue samples belonging to different classes of the studied disease. Despite the fact that there is a wide number of proposals, the complexity imposed by this problem remains a challenge today. This research proposes a gene selection approach by means of a clustering-based multi-agent system. This proposal manages different filter methods and gene clustering through coordinated agents to discover informative gene subsets. To assess the reliability of our approach, we have used four important and public gene expression datasets, two Lung cancer datasets, Colon and Leukemia cancer dataset. The achieved results have been validated through cluster validity measures, visual analytics, a classifier and compared with other gene selection methods, proving the reliability of our proposal.

Keywords Gene selection · Filter method · Multi-agent system · Clustering · Classification · Machine learning · Visual analytics · DNA-microarray

1 Introduction

Colorectal cancer (CRC) is the third most common type of malignancies worldwide and the second cause of cancer death among adults [1, 2]. CRC is originated by cell damage accumulation and mutations in genes affecting a series of major signalling pathways. About 5% of all cases of this disease are caused by a hereditary syndrome [3]. Summarizing, the incidence and severity of this disease understood as a health problem is beyond doubt, demanding further research for better detection [4]. At its origin, CRC is a benign adenomatous polyp, then it gradually progresses to an adenoma before turning into an invasive cancer [2]. Thus, advances leading to understanding molecular processes taking place in CRC result essential for a suitable knowledge management on the part of both doctors and researchers [4].

Meanwhile, Lung cancer is one of the most common types of malignancies worldwide and one of the most frequent causes of death in developed countries, constituting 27% of all cancer deaths [5]. Thus, early diagnosis is essential for the patient's survival. Unfortunately, most patients are diagnosed at an advanced stage of the disease, in which

This research has been co-financed by the European Social Fund (Operational Programme 2014-2020 forCastilla y Len, EDU/128/2015 BOCYL).

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

Juan Ramos
juanrg@usal.es
http://bisite.usal.es

Alfonso González-Briones
alfonsogb@usal.es

Juan F. de Paz
fcofds@usal.es

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

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

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

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

they have already developed metastases [6]. Such an event is a consequence of the lack of early symptoms, which do not appear until the disease is in a critical condition. Hence, this proliferative syndrome presents a high risk of metastasis which binds to the absence of effective treatments. This has led researchers to develop classifiers based on microarray technology which are able to support metastasis diagnosis and prognosis toward different organs [7].

Leukemia cancer is also one of the most common types of malignancies worldwide, it starts in the stem cells of the bone marrow making blood cells. Leukemia is basically a cancerous disorder of the blood cells for which the cells are not properly formed. Blood cells cannot be divided or reproduced as somatic cells do because they do not have DNA. There exist many kinds of leukemias, each with its own peculiar signs. The leukemia dataset has been taken from a collection of leukemia patient samples reported in [8]. It contains gene expressions corresponding to acute lymphoblast leukemia (ALL) and acute myeloid leukemia (AML) samples from bone marrow and peripheral blood.

The molecular complexity involved in cancer constitutes a major problem for clinical research. This is one of the main reasons why these types of cancer require a further molecular characterization for better understanding of mechanisms affected in tumor cell invasion [9]. Major research efforts are aimed at the discovery of new biomarkers as well as an early diagnosis and identification of specific mutations [10]. In this context, DNA-microarrays provide a means of identification for new genes being key in the genesis and development of diseases [11]. However, the exploration of these large datasets looking for a small subset of significant genes is a crucial but very difficult issue. The use of data mining techniques along with information visualization technology can help to cope with this problem for improving the data analysis process [12].

Feature/gene selection involves an important research topic in gene expression data, dealing with the gene discovery relevant for a particular target annotation. Those genes are called informative genes or differentially expressed genes since they are able to differentiate samples from different populations [13]. They are the basis for developing classifiers in the study of disease diagnosis and prognosis [14, 15]. Although a wide number of methods have been proposed to face the gene selection problem, there is not a single method able to solve all the underlying issues. Multi-agent systems (MAS) are an alternative for building analysis software in Bioinformatics. MAS are endowed with learning and adaptation, allowing us the deployment of autonomous and proactive softwares [16]. Hence, their ability to adapt to the environment, facing highly complex systems.

In consequence with all of the above, this paper proposes a clustering-based MAS for gene selection from gene expression data. The MAS coordinates tasks as gene filtering, gene clustering and cluster visualization components, which cooperate with each other to reach a common goal, a subset of informative genes from an input dataset. We want to stress that the main novelty of the current approach which makes it different to other proposals in the literature is that it includes a MAS and the calculation of cluster boundary gene-points for the discovery of informative genes. Finally, the remainder of this paper has been divided into the following sections: Sect. 2 deals with the existing approaches related to gene selection and their difference with respect to our proposal. Section 3 explains the components and the operation of the MAS in the gene selection process. Section 4 develops a case study on three public datasets and explains the results. Section 5 states the conclusions of this research, whereas references are listed at the end of this paper.

2 Related Work

Gene selection (GS) can be generically defined as the process of extracting gene subsets whose expression level values are representative of a particular target feature, i.e., clinical or biological annotation [13–15]. GS is a very active research area in the analysis of gene expression microarray, which is contributing to the development of the field as a result of involved data mining and machine learning techniques. Particularly, GS from microarrays is addressed to identify/discover those genes which are expressed differentially according to a determined target disease (namely informative genes).

GS methods have been divided into the following four categories: filters, wrappers, embedded and ensemble. Filter methods have been directed to discriminate or filter features/genes based on the intrinsic properties of the dataset. They do this by estimating their relevance scores to state a cut-off schema where an upper/lower bound is imposed to choose features with the best scores [13, 27]. Wrapper methods use a classifier to find the most discriminant feature subset by minimizing an error prediction function [21, 28, 29]. Embedded methods are similar to wrapper but additionally they interact with the learning model, which reduces the runtime taken by wrapper methods [13, 22, 30]. Ensemble methods are relatively new and recombine results from different FS techniques to achieve a more stable feature subset, since small perturbations in the training set can have effects on the results of a GS method applied individually [23, 24, 31].

According to the reviewed literature, filter methods have been widely used in the GS process complex in comparison with the remaining methods. However, the application of a single standard method to find informative genes, i.e., assigning relevance indices to genes using some of the statistical tests and then, ranking them to select the top *k* genes is not the best option since they are often highly correlated [32, 33]. Hence, we propose a multi-agent system (MAS) developing successive filtering of genes by applying different techniques of GS and data mining. In this sense, Table 1 lists the features of methods used in the GS process in comparison with our proposal, which has also been added to the table. This table outlines ten of the main methods used in GS and describes their proposals as well as main features. Column main features describes the category of the method (filter, wrapper, embedded or ensemble) and whether the method is simple or compound by several techniques. As shown in this table, none of the proposals use MAS or cluster boundary genes for GS as done in our approach.

3 An Agent Approach for Gene Selection

The goal of using agents for gene selection in this research has been to automate and plan the different tasks involved in the gene selection process conducted by our proposal. Such tasks are usually executed sequentially and manually in a computer system by the user. This implies that the user should run every operation and modify the statistical parameter values according to the used statistical test. The possibility of committing errors is greater and we would waste time, especially when the dataset is large. In that sense, our approach takes advantage from MAS, i.e., task automation, behavior, extensibility and flexibility. Hence, the paradigm that best adjusts to the automation process of our proposal, in the

Table 1 Comparative table with the main features of gene selection approaches

Proposal	Main features	Explanation
Methods GS1 and GS2 in [17]	Filter (simple methods)	GS1 and GS2 use two gene scoring functions which incorporate the means and the variations of the expression values of genes in the samples belonging to a common class
Three methods in [18]	Filter and clustering (simple methods)	Two clustering-based methods and a correlation-based method are defined. Statistical tests are applied to each similar gene group
Entropy-based method in [19]	Filter (compound method)	This method maximizes the relevance and minimizes redundancy (entropy) of selected genes. The Battitis's greedy algorithm and simulated annealing have also been used in this process
Hybrid approach in [20]	Filter (compound method)	A hybrid approach merging Genetic Ant Colony Optimization (GACO) and FBA is proposed to identify genes to be knocked out
Random forest method in [21]	Wrapper (simple method)	At each iteration, the method builds a new forest after discarding those genes with the smallest variable importance; the selected set of genes is the one that yields the smallest OOB error rate
Embedded approach in [22]	Embedded (compound method)	This method works in two stages: first, it makes pre-selection leading to a reduced gene subset space. Second, it carries out a search ensured by a specialized Genetic Algorithm which uses (among other things) a SVM classifier
Random forest method in [23]	Ensemble (compound method)	This method uses random forest, bagging, boostrap for gene selection and classification
Modified AHP in [24]	Ensemble (compound method)	This method builds a hierarchy of factors for gene selection from different tests of gene ranking and a fuzzy system with genetic algorithms for classification
Unsupervised feature selection in [25]	Ensemble (compound method)	This approach partitions the initial feature set into clusters guided by a new measure called maximal information compression index. After that a single feature is selected from each cluster. The propose of making clusters is to minimize information loss and redundancy
Attribute clustering for grouping [26]	Ensemble (compound method)	This method applies a correlation conducted method to obtain clusters whose attributes show high correlation and interdependence to reduce the search dimension for a reduced attribute set
Our proposal, a MAS	Ensemble (compound method)	A multi-agent system manages the gene selection process through different agent layers applying different approaches: ranking methods, cluster analysis, visual analytics and boundary gene computation as a novel approach for gene selection

range of current techniques is a MAS. Moreover, the employment of MAS, additionally allows for the inclusion of solutions in future versions; such functions as case-based reasoning (CBR) allows to solve learning complex problems on the basis of past solutions. This learning adjustment would not be possible without the extensibility capabilities of MAS.

This section explains the MAS proposed for gene selection, which consists of several layers responsible for carrying out different gene filtering and clustering tasks. The MAS deals with two filtering processes (through agents) before applying clustering techniques. Once the data have been clustered by means of agents operating as hierarchical clustering methods. Other agents are executed, they are in charge of visually exploring the resulting dendrograms and applying statistical techniques of cluster validation to choose the most suitable clustering. Finally, the boundary genes for each cluster from the selected clustering are computed and assumed as informative genes. Note that boundary genes are data points that are located at the margin of densely distributed data, and are very useful in data mining applications, representing a subset of the population that possibly belongs to two or more classes [34]. Awareness of these points is also useful in classification tasks, since they can potentially be misclassified [35]. In consequence, boundary points are good candidates to be informative genes.

Finally, we want to stress that an important feature of our MAS is its ability to add, change or remove components, such as, the filter, clustering, boundary point and classifier methods making our proposal extensible.

3.1 Multi-Agent System

The agent model pursued by our approach consists of four filtering processes and a cluster analysis process, which are performed by the MAS being able to leverage its skills, such as adaptation, scalability and cooperation between agents. The use of MAS in gene expression analysis has already been used in other studies with satisfactory results [36]. JADE (Java Agent DEvelopment Framework) has been used to design and implement our MAS. So the strategy followed to reach an informative gene subset consists of four linked layers: workflow layer, filtering layer, cluster analysis layer, and boundary gene layer as shown in Fig. 1.

Workflow is the main layer of the MAS and has a single agent (MA agent) which is in charge of organizing the information flow of the remaining layers. It also states the order for each agent activity, collects information on settings and repeats sequences performed for gene expression analysis, making it possible to automate repetitive analysis tasks. The remaining layers are explained below:

1. Filtering layer: This is the initial layer applied to the target dataset. This layer is responsible for carrying out two gene filtering processes to reduce noise in the input dataset. Therefore, the layer (FA agent) coordinates the agents for data normalization, significance test by relating genes to the studied disease and the objective function, which combines significance with variance to capture those genes whose variation of their expression levels is meaningful with respect to the rest, whereas high significance is also kept. The Mann–Whitney test has been the significance test applied as a nonparametric test, which states the null hypothesis relating samples to the same population, whereas the alternative hypothesis relates samples to different populations [37]. Thus, once the Mann–Whitney agent has been applied, genes with p value under 0.05 are filtered out towards the next process. Note that such genes are who reject the null hypothesis and in consequence, they have the greatest statistical significance. The following filtering process selects genes with high variation of their expression levels at the same time that high significance for the context is kept. Hence, we combine the variance agent to measure such variations with the significance (p value) assigned to genes into an objective function (score function agent) to filter out those relevant genes. Then, the score given to a gene g holds the following objective function:

$$\text{Score}(g) := \alpha_1 \times \text{Significance}(g) + \alpha_2 \times \text{Variance}(g), \tag{1}$$

where α_1 and α_2 are scalars which can be defined as $\alpha_1 = -1$ since it is in real interval [0, 1], whereas $\alpha_2 = \frac{1}{\text{maxvar}}$. maxvar is the maximum gene variance in the dataset. In consequence, the larger the values of function Score the higher the gene relevance. This means finding small values for Significance against big values for Variance as a maximization process. Then, by assigning a score to each gene based on this function and defining a threshold to filter out those genes with high score, we achieve a reduced dataset as a result of this layer to the workflow layer.
2. Cluster analysis layer: This layer has a dataset as input, filtered by the filtering layer and is responsible for choosing a clustering favorable for the next layer. To do this, three agents representing hierarchical clustering methods commonly used in cluster analysis of DNA-microarray data have been coordinated through the clustering agent (CA). This last agent is in charge of comparing the results (dendrograms) of the three clustering agents to finally selected the most suitable clustering as the end result. Then, the CA agent first uses global cluster validity measures (through the cluster validity agent, CVA for short) to compare the

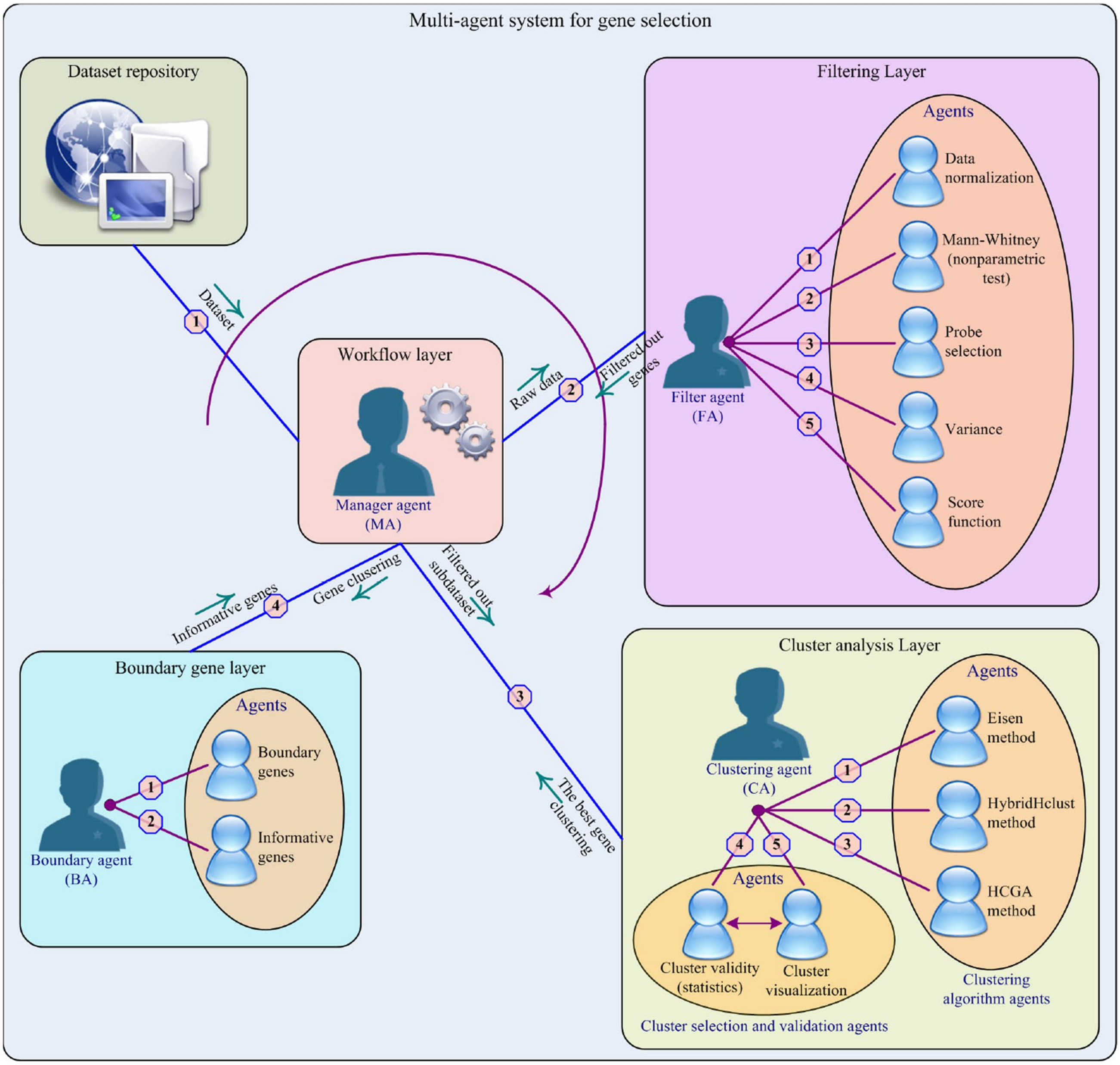

Fig. 1 Multi-agent system for gene selection. There are four linked layers performing specific tasks in the gene selection process. That is, linked processes for gene filtering, cluster analysis and validity, visual analytics and boundary gene calculation. The workflow layer (MA agent) is responsible for managing the information flow with the rest of agent layers (FA, CA and BA agents)

quality of the dendrograms built by its agents. Second, through the cluster visualization agent (CViA), the user visually selects a clustering for each dendrogram with the help of the CVA agent. This last agent assists the CViA agent by assigning a score to each clustering of the visualized dendrogram (based on the clustering homogeneity and separation measures), which allows the user to have an additional statistical criterion of the clustering quality in the dendrogram. Finally, the CA agent selects the best clustering among the ones selected from the user using local cluster validity measures given by the CVA agent. The result of this layer is passed to the workflow layer. Note that in this layer, the variance agent is given apart of the Score agent, although both cooperate to reach the goal of the objective function of the Score agent. This is so done because the variance agent can change its role without affecting the goal of the Score agent. That is, instead of carrying out the variance, it could run another statistical indicator.

- used clustering methods: The *Eisen* clustering method carries out an agglomerative hierarchical clustering in which each cluster is represented by the mean vector for data in the cluster [38]. Furthermore, this method has been one of the first methods bringing a visualization coupling heatmap with dendrogram. The *HybridHclust* method is a divisive hierarchical clustering, which is applied to the data with constraint that mutual clusters cannot be divided. Within each mutual cluster, the divisive strategy is re-applied to yield a top-down hybrid in which mutual cluster structure is retained [39]. Meanwhile, HCGA is an agglomerative hierarchical clustering based on genetic algorithms as the search method. Hence, it uses the evolutionary force to alter and recombine dendrograms from generation to generation to achieve the most favorable dendrograms [40].

3. Boundary gene layer: In this layer, the clustering selected from the cluster analysis layer is processed by the boundary gene agent (BGA) to compute the boundary genes for each of its clusters. Moreover, the informative gene agent (IGA) is responsible for converting the resulting boundary gene clustering to a set of informative genes, formatting it to make it understandable to the user and carrying out classification tasks as well as a last gene filtering process. This agent uses a k-nearest neighbor classifier (kNN), which is one of the simplest but effective classification models [41]. Both agents (BGA and IGA) are coordinated by the filter agent, which returns the informative gene subset to the workflow layer. The BGA agent uses the *ClusterBoundary* algorithm to compute boundary genes from clusters as defined in [12]. Although due to the importance of boundary genes, we consider them informative genes, the IGA agent adds a last filtering process from the boundary gene set. IGA applies a greedy strategy to remove non-significant genes for kNN. This strategy consists of removing those boundary genes that improve or remain the same accuracy of the classifier when they are not included in the classification process. Such a strategy allows us to reduce the size of the final informative gene set and improve the accuracy of the classifier. To conclude, we want to stress some concepts of boundary points and important features of the boundary point algorithm used by BGA, i.e., ClusterBoundary. As explained at the beginning of this paper, boundary points are data points located at the margin of densely distributed data and possibly belonging to two or more classes [34]. ClusterBoundary is based on the boundary definition in terms of theoretical notions from metric spaces. This way, boundary points focus on the set of points at the closure of a cluster that do not belong to the interior of the cluster. In consequence, ClusterBoundary computes cluster boundary points in a four-staged approach: (1) carry out a search for extreme points of the cluster. At each iteration of the algorithm, the cluster boundary is incrementally built from the extreme points. (2) Compute the centroid from the extreme points, which will be the center to built a ball to remove the interior points of the cluster. (3) Compute the mid-points between each extreme point-pair of the cluster. (4) Determine the radius of the a ball with the center already computed in stage (2). The goal of these four stages is to iteratively remove the interior points of the cluster while its extreme points incrementally create the boundary.

4 Case Study

This section outlines the results of applying our approach to four public datasets, two datasets for Lung cancer and two datasets for Colon and Leukemia cancer. The first dataset of Lung cancer (Lung-dataset#1, repository NCBI, http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3627), which comes from an Affymetrix Human Genome U133 Plus 2.0 Array, discloses a comparison study of two non-small cell lung cancer histological subtypes: adenocarcinomas (AC) and squamous cell carcinomas (SCC). The results provide insight into the molecular differences between AC and SCC [42]. The size of the dataset is determined by 54,675 gene probes against 58 tumor tissue samples, which are divided into 18 tissue samples for SCC and 40 ones for AC. The second dataset of Lung cancer (Lung-dataset#2, repository NCBI at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37745) comes from the same chip as the one used for Lung-dataset#1, but outlining a gene expression matrix with 54,613 gene probes × 172 tissue samples, which are divided into two classes: 66 squamous samples and 106 adeno samples. The dataset above has been used in this paper to validate the results achieved on Lung-dataset#1 since both datasets are related to the same disease.

Meanwhile, colorectal cancer (CRC-dataset) is available at http://genomics-pubs.princeton.edu/oncology/affydata/index.html. Gene expression in 40 tumor and 22 normal colon tissue samples from 40 patients have been processed on an Affymetrix oligonucleotide array complementary for more than 6500 human genes [43]. Finally, a gene expression matrix with 2000 gene probes × 62 tissue samples has been achieved. Leukemia cancer (Leukemia-dataset) is available at http://cilab.ujn.edu.cn/datasets.htm. This dataset has a matrix with 7129 genes × 72 tissue samples which have been divided into 49 samples of ALL (acute lymphoblast leukemia sample) and 23 samples of AML (acute

Table 2 Parameter settings of HCGA for Lung-dataset#1, CRC-dataset and Leukemia-dataset

Parameter	Lung-dataset#1	CRC-dataset and Leukemia-dataset
	Value (or interval)	Value (or interval)
Crossover probability	[0.60, 0.75]	[0.50, 0.65]
Mutation probability	[0.10, 0.20]	[0.05, 0.15]
Number of individuals	30	20
Number of generations	$[10^3, 10^6]$	$[10^3, 10^5]$

myeloid leukemia sample). The three datasets have been normalized by columns to mean 0 and variance 1. Once the case study has been described, we are going to show the results reached in each layer after applying the MAS (to each dataset) given in Fig. 1, namely:

1. Filtering layer: This layer includes a data normalization process (normalization agent) before applying the Mann–Whitney test agent. Then, after applying the test, the probe selection agent ranks the current dataset in ascending order through the p value of each gene-probe and selects those gene probes whose p value is <0.05. In this case, we have achieved a new dataset with 13,141 probes from 54,675 probes of Lung-dataset#1, 387 probes have been filtered out from 2000 probes of CRC-dataset and 3979 genes have been filtered out from 7129 genes of Leukemia-dataset.

   After that, the variance and Score function agents are run on the probes of the new dataset to achieve a score for each probe. Next, the current dataset has been ranked in descending order of the values given by Score (probe selection agent). At this point, a threshold to make the filtering cutoff, based on the Score function, has been fixed in the midpoint between the maximum and minimum values reached by Score. Thus, probes with score above the midpoint are filtered out to form a new dataset from Lung-dataset#1 with 999 probes. Note that those probes present the greatest expression level variation against 58 tumor tissue samples at the same time that their statistical significance (the p value) is high, too. Applying the same process to Leukemia-dataset, 527 genes has been filtered out from 3979 genes. Finally, this process has not been applied to CRC-dataset because it has been reduced sufficiently in the above task by the Mann–Whitney test agent.
2. Cluster analysis layer: This layer has three reduced datasets as its input, i.e., Lung-dataset#1 with 999 probes, CRC-dataset with 387 probes and Leukemia-dataset with 527 genes. Besides, this layer involves three main processes of cluster analysis, which consist of setting the clustering method agents to use, running them on the dataset and comparing their results to select a single clustering according to the used cluster validity measures and the dendrograms of the selected methods. Then, each of these processes has been performed in the following way:

   – Settings: The Euclidean distance between data has been used for all methods. In the case of HCGA, we must choose a fitness function and prefix values of a parameter set. So such a fitness function is based on the tradeoff of cluster homogeneity and separation to be defined as:

   –

$$f_{\mathrm{d}}(\mathfrak{G}) = \frac{1}{|\mathfrak{G}| - 1} \sum_{i=1}^{|\mathfrak{G}|-1} f_{\mathrm{c}}(\mathfrak{C}_i), \tag{2}$$

   – where $\mathfrak{G}$ is a dendrogram, $\mathfrak{C}_i$ is the clustering of level i in $\mathfrak{G}$ and f_{c} is the recurrent fitness function to evaluate a clustering of $\mathfrak{G}$, which is defined as,

$$f_{\mathrm{c}}(\mathfrak{C}_{i+1}) = \frac{S_1^*(\mathfrak{C}_{i+1})}{g - k + 1} - \frac{\mathcal{H}_1^*(\mathfrak{C}_{i+1})}{k - 1} + \max \mathfrak{D}, \tag{3}$$

   where $S_1^*(\mathfrak{C}_{i+1})$ and $\mathcal{H}_1^*(\mathfrak{C}_{i+1})$ are separation and homogeneity for clustering $\mathfrak{C}_{i+1}$, respectively, being defined in [40], $k = |\mathfrak{C}_i|$ and $g = \binom{k}{2}$, being the number of distances among the clusters of $\mathfrak{C}_{i+1}$. $\max \mathfrak{D}$ is the maximum distance from proximity matrix $\mathfrak{D}$ of the current dataset. Once the fitness function has been introduced, the HCGA parameters have been initialized as listed in Table 2. The crossover and mutation operators are given by default from the method [40].

   – Method comparison: To compare the results of the three methods, we have used the cluster validity measures, homogeneity (Homog), separation (Separ) and silhouette width (SilhoW) [12], which have been applied by the cluster validity agent to the dendrograms of each method. Keep in mind that the smaller the homogeneity value the higher the cluster quality, whereas the bigger the separation and silhouette width value the higher the cluster quality. Tables 3, 4 and 5 list the scores reached by each method on each dataset with respect to the used

Table 3 Comparison of global cluster validity based on separation and homogeneity for methods HybridHclust, Eisen and HCGA applied to Lung-dataset#1

Method	Homog	Separ	SilhoW
HybridHclust	$\underline{6.240} \pm \underline{0.074}$	10.490 ± 0.038	0.077 ± 0.005
Eisen	8.728 ± 0.108	$\underline{13.018} \pm \underline{0.213}$	-0.026 ± 0.011
HCGA	6.435 ± 0.120	10.657 ± 0.070	$\underline{0.085} \pm \underline{0.003}$

Table 4 Comparison of global cluster validity based on separation and homogeneity for methods HybridHclust, Eisen and HCGA applied to CRC-dataset

Method	Homog	Separ	SilhoW
HybridHclust	2.261 ± 0.104	9.992 ± 0.428	0.236 ± 0.236
Eisen	04.698 ± 0.148	20.186 ± 1.012	0.500 ± 0.019
HCGA	5.859 ± 0.153	17.747 ± 1.781	0.027 ± 0.074

Underline values represent the best score reached by the measures used in each column ofthe tables

Table 5 Comparison of global cluster validity based on separation and homogeneity for methods HybridHclust, Eisen and HCGA applied to Leukemia-dataset

Method	Homog	Separ	SilhoW
HybridHclust	1.681 ± 0.022	3.862 ± 0.020	0.226 ± 0.008
Eisen	2.023 ± 0.013	4.357 ± 0.013	0.139 ± 0.014
HCGA	2.081 ± 0.021	4.338 ± 0.006	0.193 ± 0.011

Underline values represent the best score reached by the measures used in each column ofthe tables

Table 6 Comparison of local cluster validity based on separation and homogeneity for each selected clustering of HybridHclust, Eisen and HCGA in Lung-dataset#1

Method	Cluster	Homog	Separ	SilhoW
HybridHclust	9	6.748	10.728	0.116
Eisen	39	8.191	11.853	−0.040
HCGA	8	7.192	11.034	0.156

Underline values represent the best score reached by the measures used in each column ofthe tables

Table 7 Comparison of local cluster validity based on separation and homogeneity for each selected clustering of HybridHclust, Eisen and HCGA in CRC-dataset

Method	Cluster	Homog	Separ	SilhoW
HybridHclust	11	2.353	9.400	0.206
Eisen	18	4.418	19.965	0.495
HCGA	19	5.510	9.592	0.061

Underline values represent the best score reached by the measures used in each column ofthe tables

cluster validity measures. The scores are the result of computing the mean cluster validity values from applying each measure to each clustering of the dendrograms. The standard error has also been shown for each score and the best scores for each measure have been stressed. Note that these tables perform the first overview of overall quality, although we need to visually explore each dendrogram (with the help of the statistical information given by the cluster validity agent) from each table to finally select a single clustering for each dataset. With the results underlined in Table 3, we cannot yet decide which is the best method. However, in Table 4, it appears to be more clear what method performed better, i.e., the Eisen method, although from the local point of view (to a clustering level), this result can change.

According to the above, we have carried out a visual inspection (using heatmap, dendrogram and scatterplot visualizations given by the cluster visualization agent, CViA for short) based on cluster validity measures from the cluster validity agent (CVA). Those measures are applied to all achieved dendrograms to select the most suitable clustering for each method of each table. Note that both agents, CViA and CVA, cooperate to assist the user in the process of clustering selection. After that, a single clustering is selected for each dataset using the same cluster validity measures on the resulting clusterings. Tables 6, 7 and 8 show the results for each dataset. The Cluster column represents the number of clusters for the clustering selected from each dendrogram. The number of clusters is an additional criterion to finally select a single clustering for each dataset. According to Table 6, the clustering selected from Lung-dataset#1 has been that of the HCGA method, which presents the best silhouette width and whose separation is very similar to the one of the Eisen method, which reached the best separation. As for Table 7, the clustering selected from CRC-dataset has been that of the Eisen method, which presents the best scores for separation and silhouette width.

Supporting the results shown in Tables 6, 7 and 8, Figs. 2 and 3 display the clusterings selected for each dataset. Figure 2 shows visualizations of dendrograms with heatmap for each dataset, whereas Fig. 3 shows the selected clusterings in form of 3D-points representing genes (3D-scatterplots for each dataset). Points with the same color represent genes of the same cluster. The graphics given in Fig. 3 have been achieved by reducing the dimensionality of the gene space to three features for each gene. Principal component analysis (applied correlation analysis) has been used to linearly project genes into the first three principal components [44]. Finally, note that this layer combines three processes of result validation to select the most suitable clustering for each dataset, namely global cluster validity, visual cluster validity and local cluster validity.

– Boundary point layer: Once the clusterings representing each dataset have been achieved, the next task is to run the boundary gene agent (BGA) to compute the cluster boundary genes for each clustering. After that, the informative gene agent (IGA) is in charge of for-

Table 8 Comparison of local cluster validity based on separation and homogeneity for each selected clustering of HybridHclust, Eisen and HCGA in Leukemia-dataset

Method	Cluster	Homog	Separ	SilhoW
HybridHclust	12	1.857	3.948	0.276
Eisen	8	2.040	4.427	0.450
HCGA	10	2.039	4.428	0.490

Underline values represent the best score reached by the measures used in each column ofthe tables

matting the resulting boundary gene clustering, evaluating and filtering those genes by means of a classifier. In this case, kNN (*k*-nearest neighbors, [45]) has been used for that propose. Both agents (BGA and IGA) are coordinated by the boundary agent (BA) to reach common goal for this layer, the subset of informative genes. Once the computation of boundary genes has been performed, 76 gene probes belonging to 63 genes have been achieved for Lung-dataset#1, CRC-dataset has achieved 46 boundary genes and Leukemia-dataset has achieved 57 boundary genes. These three gene subsets can already be considered sets of informative genes. However, one of the roles of the IGA agent is to apply a kNN-based greedy strategy to reduce the gene number given in the subsets above without losing their predictive ability. In that sense, each subset above has been reduced as follows: Lung-dataset#1 has been reduce to 4 genes, CRC-dataset has been reduced to 19 genes and Leukemia-dataset has been reduced 3 genes.

To evaluate the significance of such gene subsets, we have evaluated the accuracy of the kNN classifier for each found subset. Table 9 lists the accuracy reached for each gene subset, before and after of applying the greedy strategy of gene reduction (or gene filtering). Because there are no available test data from selected datasets to evaluate the classifier, we have adopted methodology stratified tenfold cross-validation [46]. As shown in this table, the three informative gene subsets reached high accuracy through the kNN classifier, which proves the significance of the subsets to be used in classification tasks. As shown in this table, the accuracy for each boundary gene subset was improved and the number of genes was decreased after applying the kNN-based greedy strategy of the IGA agent. This proves that in the boundary gene subsets are the informative genes.

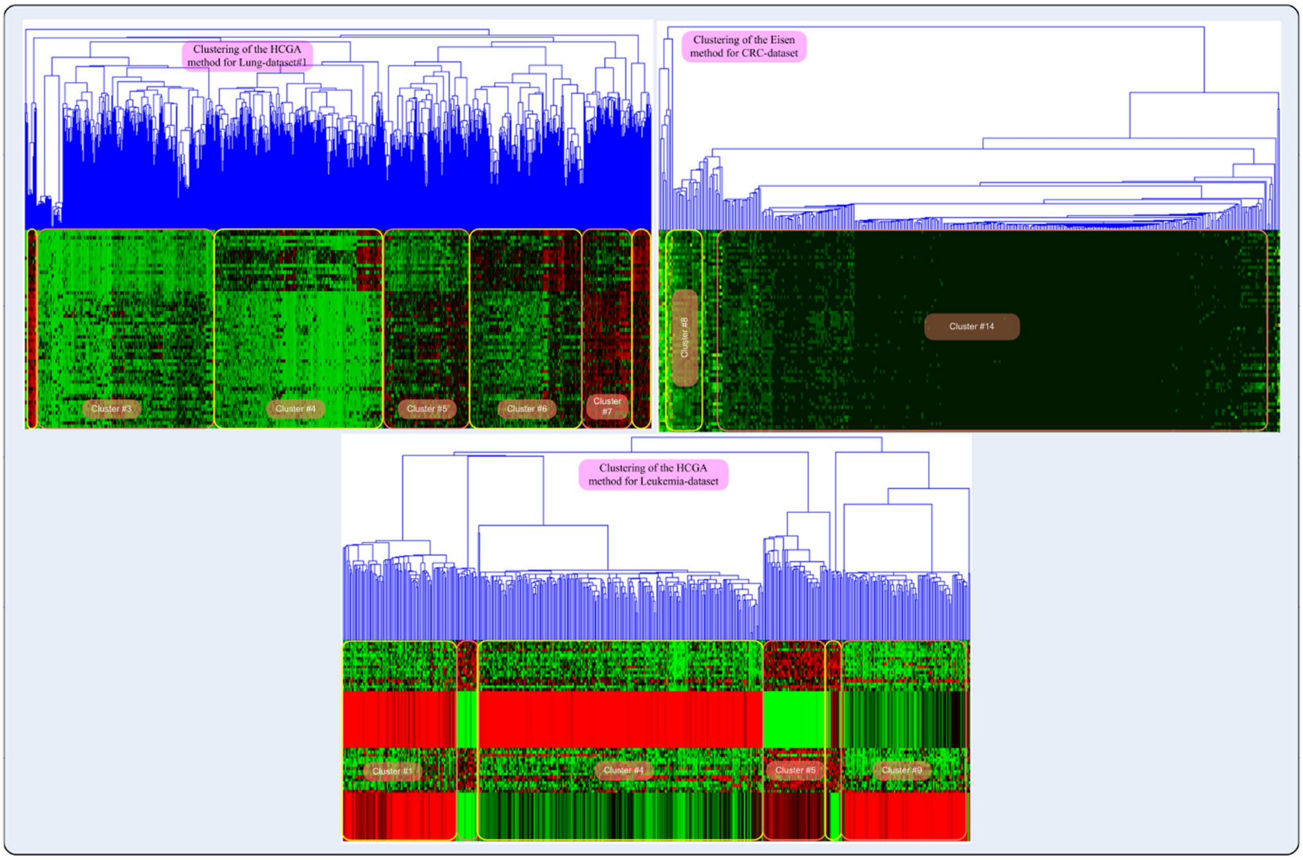

Fig. 2 Dendrograms on heatmaps representing the selected clustering for each dataset. The Eisen method has been selected for CRC-dataset, whereas the HCGA method has been selected for Lung-dataset#1 and Leukemia-dataset

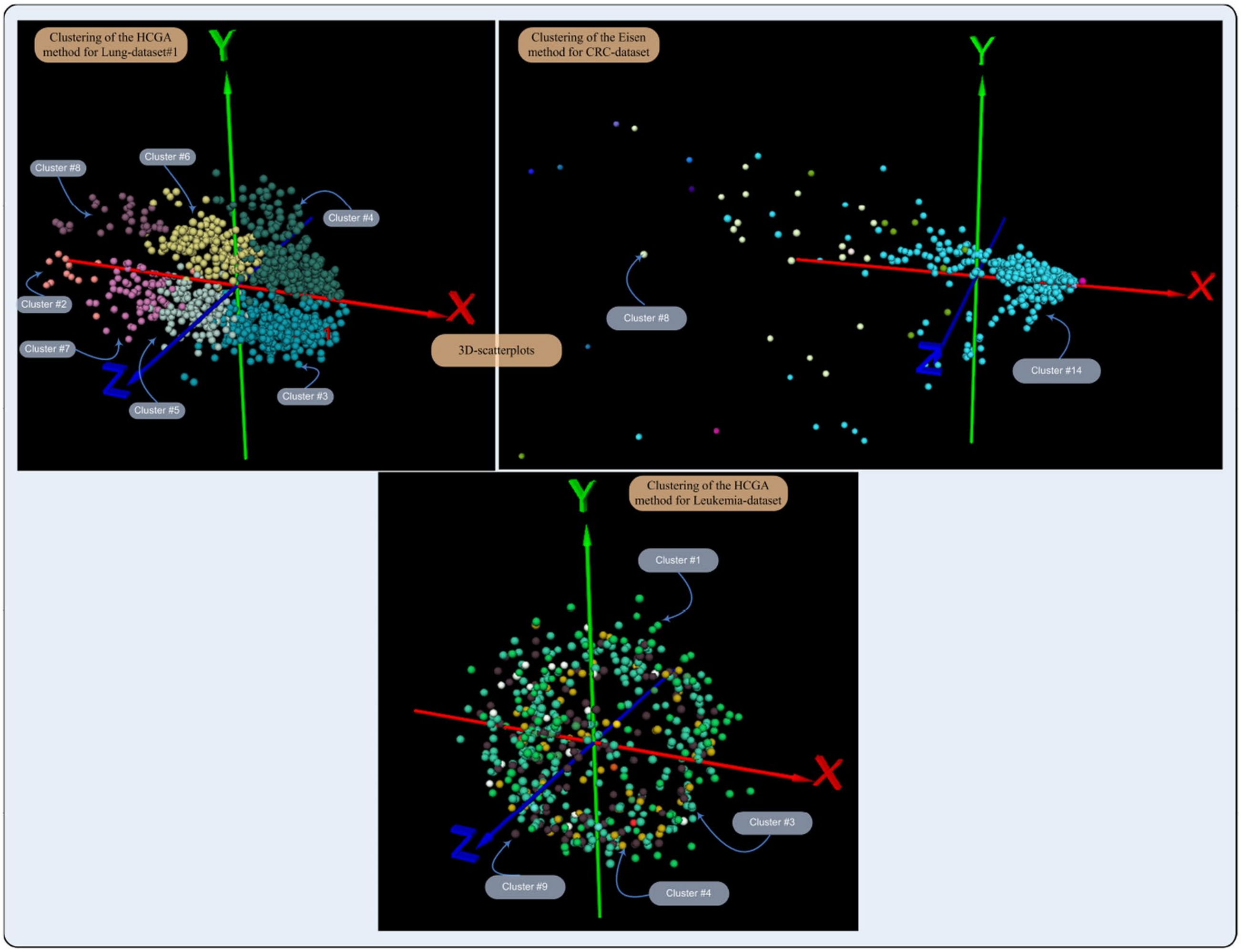

Fig. 3 3D-Scatterplots (principal component analysis) representing the selected clustering for each dataset and clustering methods Eisen and HCGA. Each gene cluster is represented by gene-points with the same color

In accordance with the above, Table 10 displays the results reached on Lung-dataset#2 by choosing (from it) the same boundary genes and genes filtered from such boundary genes given in Table 9 for Lung-dataset#1. The goal of this table is to validate the results reached on Lung-dataset#1 through another different dataset of the same cancer, i.e., Lung-dataset#2. Then, the second row of this table shows two results: the accuracy (kNN classifier) of the 76 boundary genes identified in Lung-dataset#2, which have been discovered in Lung-dataset#1. The other result assesses the four genes identified in Lung-dataset#2, which have been filtered from the 76 boundary genes given in Lung-dataset#1. The last row of the table lists the results reached when the gene filtering strategy used in Table 9 (kNN-based greedy strategy) is applied to the gene sets given in the second row (76 and 4 genes).

Table 9 Accuracy of the kNN classifier applied to the three informative gene subsets found for Lung-dataset#1, CRC-dataset and Leukemia-dataset

Informative genes	Dataset	Number of genes	*K*	Accuracy (%)
Boundary genes	Lung-dataset#1	76	7	87.930
	CRC-dataset	46	9	85.484
	Leukemia-dataset	57	16	90.278
Genes filtered from boundary genes	Lung-dataset#1	4	1	98.276
	CRC-dataset	19	4	90.322
	Leukemia-dataset	3	13	97.222

As shown in Table 10, the accuracy of the boundary genes discovered from Lung-dataset#1 increased its value in Lung-dataset#2, which proves that the boundary genes found by our approach are meaningful beyond the dataset selected to study the same disease. In the case of the 4 genes discovered for Lung-dataset#1, we have that they have shown a decrease in accuracy for Lung-dataset#2.

Table 10 This table selects the same genes discovered by our approach from Lung-dataset#1 (Table 9) in Lung-dataset#2 to evaluate their accuracy through the kNN classifier

Meaning	Number of genes (Lung-dataset#2)	*K*	Accuracy (%)
Boundary genes discovered in Lung-dataset#1 that have been identified in Lung-dataset#2	76 4	3 31	96.512 88.372
Genes filtered from the genes identified above for each case of Lung-dataset#2	16 3	1 18	97.674 90.698

Furthermore, the same filtering process used in Table 9 is also applied to such genes identified in Lung-dataset#2

Nevertheless, this resulting accuracy has not been low. On the other hand, the results shown at the end of this table show that the accuracy of the gene selected from both, the 76 boundary genes and 4 genes given in the second row from Lung-dataset#2 increased. All of this confirms our hypothesis that boundary genes contain a reduced set of informative genes.

To conclude this section, Table 11 shows a comparison of our approach (MAS) with respect to five recent gene selection methods. The accuracy reached by the selected genes of each method is listed along with the number of neighbors (*K*) used in the classification process of kNN, number of genes achieved for each method, name of the methods and the dataset used in each case. The gene selection methods used are: propOverlap given in [47, 48], Boruta in [49, 50], kofnGA in [51, 52], SDA in [53, 54] and Spikeslab in [55, 56]. For the case of kofnGA which is a genetic algorithm, its main parameters have been initialized as follows, Lung-dataset#1: population size = 100, for all datasets, number of generation = 10,000, the fitness function used for all datasets has been correlation between the genes, the remaining parameters have been initialized as stated by the method, for all datasets. CRC-dataset: number of generation = 5248, whereas for Leukemia-dataset, the number of generation was initialized to 5000. The remaining methods have been initialized with their default values. As shown in Table 11, our proposal reached the best results for Lung-dataset#1 and CRC-dataset. For the case of Leukemia-dataset, the accuracy reached by our proposal was close to the one of the methods that achieved the best results. Furthermore, our method along with propOverlap achieved the lowest number of genes.

5 Conclusions

This paper has presented a MAS for gene selection and tissue samples classification from DNA-microarray data. The main goal of this approach has been to automate the processes of clustering selection, visual and analytical cluster validity, gene filtering and classification through a MAS.

Table 11 Comparative table of our proposal (MAS) with respect to five methods of gene selection which have been executed on the three used datasets

Dataset	Gene selection method	Number of genes	*K*	Accuracy (%)
Lung-dataset#1	propOverlap	1824	2	89.654
	Boruta	30	1	98.276
	kofnGA	100	2	68.965
	SDA	40	1	96.552
	Spikeslab	74	4	91.378
	MAS	4	1	98.276
CRC-dataset	propOverlap	550	8	80.644
	Boruta	16	3	83.871
	kofnGA	20	2	70.968
	SDA	20	3	88.710
	Spikeslab	51	3	88.710
	MAS	19	4	90.322
Leukemia-dataset	propOverlap	2	1	97.222
	Boruta	58	1	98.610
	kofnGA	30	2	70.832
	SDA	10	4	98.610
	Spikeslab	93	4	94.443
	MAS	3	13	97.222

The accuracy of the kNN classifier for each method is listed along with its parameter *K*, number of genes and name of each method

Underline values represent the best score reached by the measures used in each column ofthe tables

Within this approach, the practical goal has been to target the selected genes to classification tasks in Lung, Colon, and Leukemia cancer. According to that, we have achieved a subset with 4 informative genes for Lung cancer, a subset with 19 informative genes for Colon cancer and a subset with 3 informative genes for Leukemia cancer. The three subsets have been evaluated in a classifier and compared with other gene selection methods, for which a high accuracy was reached. In the case of Lung cancer, we have used two different datasets representing this disease (Lung-dataset#1 and Lung-dataset#2). Since both datasets represent the same cancer, we have assessed the significance of the informative gene subsets discovered from Lung-dataset#1 in Lung-dataset#2. The results on the second dataset showed good accuracy, which proves that our proposal finds genes meaningful for the studied disease and that such genes are regardless of the dataset used. Therefore, those genes can be used in classification tasks related to the studied disease. Hence, these promising results prove the reliability of our approach.

References

1. Kim S-E, Paik HY, Yoon H, Lee JE, Kim N, Sung M-K (2015) Sex- and gender-specific disparities in colorectal cancer risk. World J Gastroenterol (WJG) 17(21):5167–5175
2. Markowitz S, Bertagnolli M (2010) Molecular basis of colorectal cancer. N Engl J Med 25(361):2449–2460
3. Balaguer F (2014) Cáncer colorrectal familiar y hereditario. Gastroenterología y Hepatología 37:77–84
4. Perea J, Lomas M, Hidalgo M (2011) Molecular basis of colorrectal cancer: towards an individualized management. Revista Española de Enfermedades Digestivas 1(103):29–35
5. Schwartz A, Prysak G, Bock C, Cote M (2006) The molecular epidemiology of lung cancer. Carcinogenesis 28(3):507–518
6. Rothschild SI (2015) Advanced and metastatic lung cancer—what is new in the diagnosis and therapy. PRAXIS 104:745–750
7. Wang KJ, Melani A, Chen KH, Wang KM (2015) A hybrid classifier combining borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer: A case study in taiwan. Comput Methods Progr Biomed 119:63–76
8. Golub T, Slonim D, Tamayo P, Huard C, Gassenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield D, Lander E (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537
9. Martin T, Ye L, Sanders A, Lane J, Jiang W (2013) Metastatic Cancer: Clinical and Biological Perspectives, chap. Cancer invasion and metastasis: molecular and cellular perspective. Landes Bioscience
10. Zappa C, Mousa S (2016) Non-small cell lung cancer: current treatment and future advances. Transl Lung Cancer Res 5(3):288–300
11. Berrar DP, Dubitzky W, Granzow M (2003) A practical approach to microarray data analysis. Kluwer Academic Publishers, New York
12. Castellanos-Garzón JA, García CA, Novais P, Díaz F (2013) A visual analytics framework for cluster analysis of DNA microarray data. Expert Syst Appl (Elsevier) 40:758–774
13. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, de Schaetzen 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
14. Inza I, Larrañaga P, Blanco R, Cerrolaza A (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med, Data Min Genom Proteom (Elsevier) 31:91–103
15. 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
16. Márquez E, Espinosa A, Lemaitre C, Berumen J, Savage J, Leder R (2011) Identification of relevant genes with a multi-agent system using gene expression data. INTECH Open Access Publ 19:425–438
17. Yang K, Cai Z, Li J, Lin G (2006) A stable gene selection in microarray data analysis. BMC Bioinform 7(228):1–16
18. Jaeger J, Sengupta R, Ruzzo W (2003) Improved gene selection for classification of microarrays. Pac Symp Biocomput 8:53–64
19. Liu X, Krishnan A, Mondry A (2005) An entropy-based gene selection method for cancer classification using microarray data. BMC Bioinform 6(76):1–14
20. Mohamed A, Saberi M, Deris S, Omatu S, Fdez-Riverola F, Corchado J (2015) Gene knockout identification for metabolite production improvement using a hybrid of genetic ant colony optimization and flux balance analysis. Biotechnol Bioprocess Eng (Springer) 20(4):685–693
21. Díaz-Uriarte R, Alvarez SD (2006) Gene selection and classification of microarray data using random forest. BMC Bioinform 7:1–3
22. Hernandez J, Duval B, Hao JK (2007) A genetic embedded approach for gene selection and classification of microarray data. In: EvoBIO 2007, lecture notes in computer science (LNCS), vol 4447. Springer, Berlin, pp 90–101
23. Moorthy K, Saberi M (2012) Random forest for gene selection and microarray data classification. In: Knowledge technology, third knowledge technology week, KTW, communications in computer and information science, vol 295. Springer, Berlin, pp 174–183
24. Nguyen T, Khosravi A, Creighton D, Nahavandi S (2015) Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification. PLoS One 3(10):1–23
25. Mitra P, Murthy C, Pal SK (2002) Unsupervised feature selection using feature similarity. IEEE Trans Pattern Anal Mach Intell 24(3):301–312
26. Au WH, Chan K, Wong A, Wang Y (2007) Attribute clustering for grouping, selection, and classification of gene expression data. IEEE/ACM Trans Comput Biol Bioinform (IEEE) 2(2):83–101
27. Guyon I (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
28. Ambroise C, McLachlan G (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA (PNAS) 99:6562–6566
29. Zhou Y, He J (2007) A runtime analysis of evolutionary algorithms for constrained optimization problems. IEEE Trans Evolut Comput 11:608–619
30. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517
31. Haury AC, Gestraud P, Vert JP (2011) The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS One 6(12):e28210. doi:10.1371/journal.pone.0028210

32. Castellanos-Garzón JA, Ramos J (2015) A gene selection approach based on clustering for classification tasks in colon cancer. Adv Distrib Comput Artif Intell J (ADCAIJ) 4(3):1–10
33. Jager J, Sengupta R, Ruzzo W (2003) Improved gene selection for classification of microarrays. In: Pacific symposium on biocomputing (UW CSE Computational Biology Group), PMID: 12603017
34. Xia C, Hsu W, Lee ML, Ooi BC (2006) Border: efficient computation of boundary points. IEEE Trans Knowl Data Eng 18:289–303
35. Jain AK, Murty NM, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
36. González A, Ramos J, De Paz J, Corchado J (2015) Obtaining relevant genes by analysis of expression arrays with a multiagent system. In: 9th international conference on practical applications of computational biology and bioinformatics. Springer International Publishing, pp 137–146
37. Weiss P (2005) Applications of generating functions in nonparametric tests. Math J 9(4):803–823
38. Eisen M, Spellman T, Brown P, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. In: Proceedings of the National Academy of Sciences, vol 95. USA, pp 14863–14868
39. Chipman H, Tibshirani R (2006) Hybrid hierarchical clustering with applications to microarray data. Biostatistics 7:302–317
40. Castellanos-Garzón JA, Díaz F (2013) An evolutionary computational model applied to cluster analysis of DNA microarray data. Expert Syst Appl (Elsevier) 40:2575–2591
41. Tan P, Steinbach M, Kumar V (2006) Introduction to data mining. Addison-Wesley, Reading
42. Kuner R, Muley T, Meister M, Ruschhaupt M, Buness A, Xu E, Schnabel P, Warth A, Poustka A, Snltmann H, Hoffmann H (2009) Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. Lung Cancer 63(1):32–8
43. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745–6750
44. Jolliffe IT (2000) Principal component analysis. Springer, New York
45. 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
46. Flach P (2012) Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press, Cambridge
47. 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
48. 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
49. Kursa M, Rudnicki W (2010) Feature selection with the Boruta package. J Stat Softw 36(11):1–13
50. Kursa M, Rudnicki W (2010) Feature Selection with the Boruta Package. J Stat Softw36(11):1–13. http://www.jstatsoft.org/v36/i11/
51. Wolters M (2015) A genetic algorithm for fixed-size subset selection. R-Package kofnGA, Version 1.2
52. Wolters M (2015) A genetic algorithm for selection of fixed-size subsets with application to design problems. J Stat Softw 68(1):1–18
53. Ahdesmaki M, Strimmer K (2010) Feature selection in omics prediction problems using CAT scores and false non-discovery rate control. Ann Appl Stat 4:503–519
54. 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
55. Ishwaran H, Rao J (2005) Spike and slab variable selection: frequentist and bayesian strategies. Ann Stat 33(2):730–773
56. Ishwaran H, Rao J, Kogalur U (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