IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 38-45
www.iosrjournals.org

Evolution of A Dual-Level Gene Selection Technique That Integrate A Novel Composite Approach Using The Cuckoo Optimization Algorithm And Harmony Search For Cancer Classification

*R.Kalai vani[1], S.Subhasini[2]
[1](Assistant professor,Department of IT, Sankara college of Science and commerce,Coimbatore,TN, India)
[2](Research scholar, Sankara college of Science and commerce,Coimbatore,TN,India

Abstract: *For each cancer type, only slight genes are descriptive. The gene selection job remains a challenging one. To conquered this problem, we propose a dual-level gene selection Technique called MRMR-COA-HS. In the first level, the minimum redundancy and max-imam relevance (MRMR) feature selection is used to choose a subset of relevant genes. The preferred genes are then fed into a covering setup that incorporate a new algorithm, COA-HS, using the support vector machine as a classifier. The technique was enforced to four microarray datasets, and the performance was analyzed by the leave one out cross-acceptance method. Provisional performance analysis of the pro-posed method with other evolutionary algorithms recommended that the proposed algorithm powerful outperforms other technique in choosing a less number of genes while preserving the largest classification accuracy. The techniques of the used genes were further investigated, and it was validated that the selected genes are biologically relevant to each cancer type.*

Keywords
1. *Gene selection*
2. *Minimum redundancy and maximum relevance (MRMR)*
3. *Evolutionary algorithms*
4. *Cuckoo optimization algorithm*
5. *Harmony search algorithm*

(COA-HS)

I. Introduction

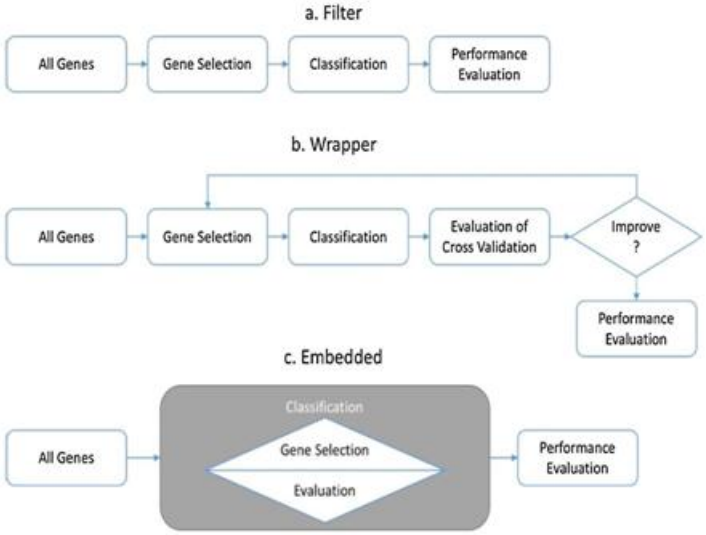
Over the last two decades, the beginning of DNA microarray method has contributed opportunities for personalized medicine by assessed the expression stages of thousands of genes simultaneously [1]. Microarray technology has recently been used to deter-mine subtypes of certain cancers based on particular in the expression stages of key genes [2–4]. This technique gives detailed data on the genetic makeup of any particular can-cer patient, thereby potentially improving the accuracy of treat-ment decisions made by physicians [5].

During microarray a, the number of genes is meaningful and It is higher than the number of samples [6,7] and classification to a next level of accuracy is challenging because of the phenomenon of dimensionality [8,9]. To avoid these problems, gene preferred techniques have been proposed in which only the most meaningful genes are selected and used for classification purposes [10–13]. There are number of advantages to this technique of reducing the number of genes and selecting only genes that are more predictive during classification. By having fewer genes, not only is the processing time for classification significantly decreased, but the chance of misclassification is also reduced.

Furthermore, using a huge number of genes as input into the classifier can cause the classifier to be over-fitted [14]Gene selection methods can be divided into three main approaches based on their interaction with the classifier, namely, filter methods, wrapper methods and embedded methods [14,15]. Filter methods analyze the similar no of genes by examining only the general features of the data and avoiding the impact of selected genes on the classification performance [16]. Wrapper gene selection prefers a search procedure in the space of possible gene subsets. The selected genes are then evaluated based on their competence to improve classification accuracy [17–19]. In the enclose gene selection technique, feature selection is linked to the classification level; however, this connection is much powerful than in the wrapper technique. This is because gene selection in incorporated technique is added in the classifier construction and the classifier is assessed to give a criterion for feature selection [20,21] (see Fig. 1). More recently, evolutionary algorithms improved for gene selection have been used within the framework of wrapper techniques [22,23]. Finally, a new gene selection approach by means of mixed based on data clustering was proposed, recommending that optimization-based clustering could select more informative genes to develop classification accuracy [24].Every gene

selection approach has merits and demerits [14]. For example, although the filter technique is simple and computationally efficient, its performance lags behind other.

Fig. 1. Featured selection methods.



Approaches, since the classifier works independently and is not associated in gene selection [25]. Conversely, the wrapper and embedded technique, which include the gene selection process into the classification task, can improve higher classification accuracy but suffer from scalability problems due to their high computational costs and are not practical for big datasets [26,27].High classification accuracy is, of course, of the importance for personalized medicine. However, biomarker is also an area of ongoing research, where it is important to diagnose a few number of genes to spot patterns (e.g., selecting small genes that are all separately expressed among different samples) [28,29]. Therefore, in this study, the main tasks were to select the optimum number of the most effective genes that can perfect distinguish between two cancer types. The gene selection assess was performed in two stages. Minimum redundancy and maximum relevance (MRMR) feature selection [30] was first used to select a subset of the most relevant and least redundant genes. The selected genes were then fed into a wrapper setup that joins the proposed COA-HS optimization algorithm with a support vector machine (SVM) as a classifier. The SVM was performed as the classifier in this work, as its classification performances has been proven and established by a number of comparative assessments with other algorithms [31–33]. Two-level gene selection combines the advantages of both the filter and wrapper methods of gene selection. The techniques were applied to four microarray datasets and the improvement was analyzed by the leave one out cross-validation (LOOCV) method.

1.1 Microarray data

Microarray data for four cancer stages (leukemia, prostate, lym-phoma, and colon) were utilized in this study. Gene expression data for leukemia [1] and prostate cancer [34] were accessed from the Broad Institute (www.broadinstitute.org). Gene types of data for lymphoma [35] were assessed from the Lymphoma/Leukemia Molecular Profiling Project (llmpp.nih.gov). A gene expression dataset for colon cancer [36] was utilized from the Princeton University Gene Expression Project (http://genomics-pubs.prince-ton.edu). Primary information based to the datasets used in this study, including the number of genes, samples and the two classes for each dataset, is provided in Table 1.

II. Methodology

2.1 Pre-Processing level

The general methodology is illustrated in Fig. 2. The data were divided into nine states. After this pre-processing level, the top 100 genes were selected using MRMR. The selected genes were fed into a wrapper setup consisting of the COA-HS algorithm and the SVM classifier to choose the minimum number of genes that gives 100% accuracy. Finally, the classification development of the selected genes was measured in terms of accuracy via the LOOCV method. To evaluate the performance of the COA-HS, the results were checked to those established from other evolution-ary algorithms, such as the genetic algorithm (GA), the particle swarm optimization (PSO) algorithm, the harmony search (HS) algorithm, and the cuckoo optimization algorithm (COA). The codes are needed in this study were written using Mat lab 2014a.

2.2 Discretization

The gene expression data were first divided to less noise and to develop the accuracy of the classification results [37]. The expression value of each gene was divided into a nine-state variable based on the mean value (l) and standard devi-ation (r) for that gene. For every gene, the nine states proposed whether the gene was not expressed (state zero) or expressed and how much it was over-expressed (states +1 to +4) or under-expressed (states1 to 4). Table 2 details the different states uti-lized in the data discretization.

3.2. First-stage gene selection using minimum redundancy maximum relevance (MRMR)

Gene expression data are typically available in a matrix format (see Fig. 3), where each row represents a gene and each column represents a sample. The last row generally represents the class label for each sample. The class label (C_i) for a two-class classification task is defined by either 1 or 2.The goal of feature selection in a classification task is to find a subset of features that best characterize the statistical signifi-cance of the classification task [38]. MRMR, a filter method of gene selection, identifies those genes that produce more data with respect to the class label of the samples. In this research, mutual data was used for MRMR to mentioned the relevancy and redundancy of genes and target classes. In this context,

Table 1 Basic information related to the microarray datasets used in this study.

Microarray dataset	Number of genes	Number of samples	Class 1		Class 2	
Leukemia[a]	7129	73		(ALL)	25	(AML)
Prostate[b]	12,600	102	50 (Normal)		52	(Cancerous)
Lymphoma[c]	4026	47		(Germinal center B-DLCL)	23	(Active B-DLCL)
Colon[d]	7457	62		(Normal)		40 Cancerous)

1. Golub et al. [1].
2. Singh et al. [32].
3. Alizadeh et al. [35].
4. Alon et al. [36].

Fig. 2. Schematic of the general methodology used for gene selection.

Table 2 Discretization of gene expression data.

Data	State	Data	State
l < d < l 1/2 r	0	l < d < l + 1/2 r	0
l 1/2 r < d < l r	1	l + 1/2 r < d < l + r	1
l r < d < l 3/2 r	2	l + r < d < l + 3/2 r	2
l 3/2 r < d < l 2r	3	l + 3/2 r < d < l + 2r	3
d < l 2r	4	d > l + 2r	4

Fig. 3. Gene expression matrix.

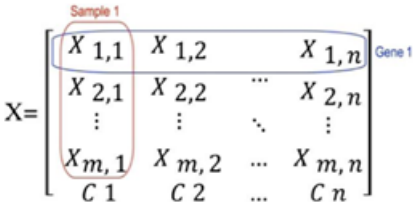


Fig 3.3. Second-stage selection using evolutionary algorithms

Basically, optimization is the of searching the accurate solution across all possible solutions. Population-relevant to optimization algorithms is started to choose a random set of solutions (initial population) that is enhanced via an iterative process. For each iteration, a cost function is organized to quantify the outcome of the optimization task. Since, in this study, the problem with the classification of microarray accurate data is developing higher accuracy using the minimum number of selected genes, the cost technique was designed as follows:Where Accuracy is the accuracy of the SVM classifier portion by the LOOCV method, NOG is the number of selected genes and a is a coefficient. Since efficient data was more important than the number of characteristics, a was set to 1000 to provide the accuracy more weight in the cost function. Therefore, by

minimizing the cost function, the number of selected genes was minimized while the accuracy was maximized. In the LOOCV method, one sample is treated as a test sample data while the remaining samples are utilized to train the SVM and the accuracy is calculated. If there are N samples, this function is indefinite N times, each time with a various sample, and the average accuracy is calculated for the selected genes. SVM was used for the classification of selected genes, as the SVM classifier is a powerful classification algorithm and has been demonstrated to exhibit excellent development in a variety of biological classification tasks [39]. The LOOCV technique was chosen, as it can overcome data over-fitting [40]. The performance of the classifier was utilized LOOCV in terms of accuracy, sensitivity and specificity. Accuracy indicates the ratio of all correctly classified samples. Sensitivity mentioned the ratio of correctly classified positive samples and specificity denotes the ratio of correctly classified negative samples.In this study, a new composite optimization algorithm, COA-HS, was improved by joining the recently invented COA [41] and HS algorithms. The results were compared with the PSO, GA, HS, and COA algorithms. PSO is an optimization algorithm that stems from the simulation of birds flocking [42–45]. The GA algorithm is another well-known evolutionary algorithm that was first introduced by John Holland in 1975 [46–48]. To fully describe the newly proposed COA-HS algorithm, the details of both COA and HS techniques will initial be explained about the following two sections.

3.3.1. Cuckoo optimization algorithm (COA)

COA is a population-related optimization algorithm that was proposed by Rajabion in 2011 [41] and was inspired by the life of a cuckoo bird. The cuckoo's performance in laying eggs is indiudual in the sense that a cuckoo never constructs its own nest when laying eggs and rests other birds' nests to lay its eggs. In doing so, if the cuckoo's eggs resemble to the host's eggs, it is likely that the cuckoo's eggs will hatch and become mature cuckoos. If the cuckoo's eggs are invented by the host bird, the alien eggs will be destroyed. In the COA algorithm, each egg in a nest indicates a potential solution andeach cuckoo represents a successful new solution. The objective of the COA is to search the nest with the highest probability of an egg's survival. Therefore, the more eggs that survive after being worked in a host nest, the highest the phase of profit assigned to that nest. When the time comes for the migration of the newly matured cuckoos, they move towards the best nest that has the highest survival rate and lay eggs within a radius of it; this is known as the egg laying radius (ELR) and can be calculated by Eq. (8).

Velocity= Number of current cuckoo's eggs

Total number of eggs

Since, in nature, the development of each population is controlled, in the COA algorithm, a parameter,Nmax, Its gives the limit to the maximum number of cuckoos that can live in each phase. The process of hatching and the migration of cuckoos towards a better nest is repeated 100 times (100 iterations) to search the best solution. In other words, the cost function is minimized through 100 iterations of the COA. A flowchart of the COA algorithm is illustrated in Fig. 4.

3.3.2. Harmony search (HS) algorithm

The harmony search (HS) algorithm is a music-inspired optimization algorithm [49]. In jazz, musicians improvise their instruments' pitch to find a perfect harmony, which can be achieved.

III. Flowchart of The Cuckoo Optimization Algorithm

There are three levels. The first level is to play a pitch from memory. The second level is to play a random pitch within the accept-able range of available pitches. Finally, the third option is to play a pitch adjacent to a pitch in their memory. In the HS algorithm, these levels are, respectively, referred to as harmony memory (HM), the pitch adjustment rate (PAR) and the harmony memory consideration rate (HMCR). The HS algorithm follows a number of steps, as outlined below:

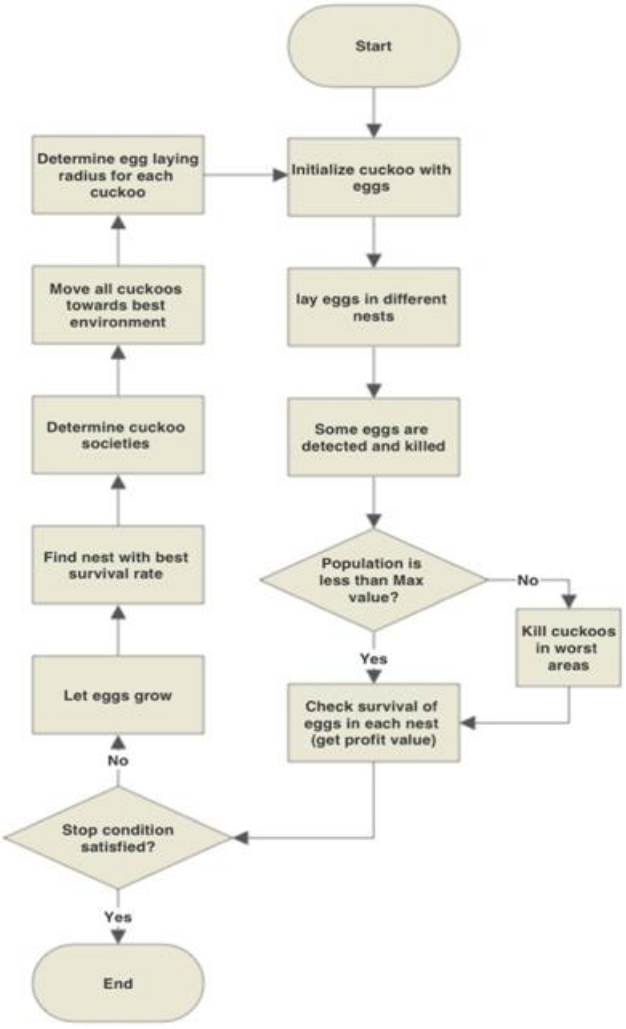
1. Initialize the HM
2. Initialize the HMCR
3. Improvise a new harmony memory
4. Update the HM
5. Check the stopping criterion.

The HS algorithm has been successfully used for various optimization problems, such as feature selection [50], discrete design variable [51] and continuous optimization problems [52].

3.3.3. Proposed algorithm (COA-HS)

In this section, the details of the proposed algorithm, COA-HS, for finding and selecting the most informative genes toImproving the classification accuracy are explained. A flowchart of COA-HS is illustrated in Fig. 5. As discussed, in the COA algorithm (Section 3.3.1), each egg in a nest include a solution and each cuckoo represents a new solution. So that, In the gene expression analysis, a solution to a gene. The COA-HS algorithm starts with the initialization of the cuckoos. At the initial iteration, after the initial population-laid eggs, the profit values of the eggs are calculated by performing the cost function. These solutions (eggs) are then fed into the HS algorithm. HS is utilized to develop the chance of determine the untouched area of the solution space, which COA alone is unable to explore. In the solution space, the untouched area can be assessed by the improvisation process through HMCR and PAR, which are set to 0.8 and 0.3, respectively. In a result, a improvement solution can be acquired by preventing premature convergence of COA.

During the first level approach, once the solutions from COA are implemented into the HS algorithm, the HS iterates 50 times, after which the profit value for the solutions recommended by the HS are calculated via the cost function. The profit values of the solutions suggested by the COA and HS are then compared and the solutions (eggs) with the improved profit value are chosen to survive. After that, these eggs grow and become cuckoos. The survival rate of each cuckoo is calculated and all cuckoos move towards the nest with the highest survival rate and lay eggs within the ELR of the best nest (best position). In other words, the space of solutions is refined towards the best solution. This concludes one iteration of the COA-HS algorithm and the process is repeated 50 times. Each time the cuckoos lay eggs in a further improved position, these outputs in finding a good solution related on finding the cost function.



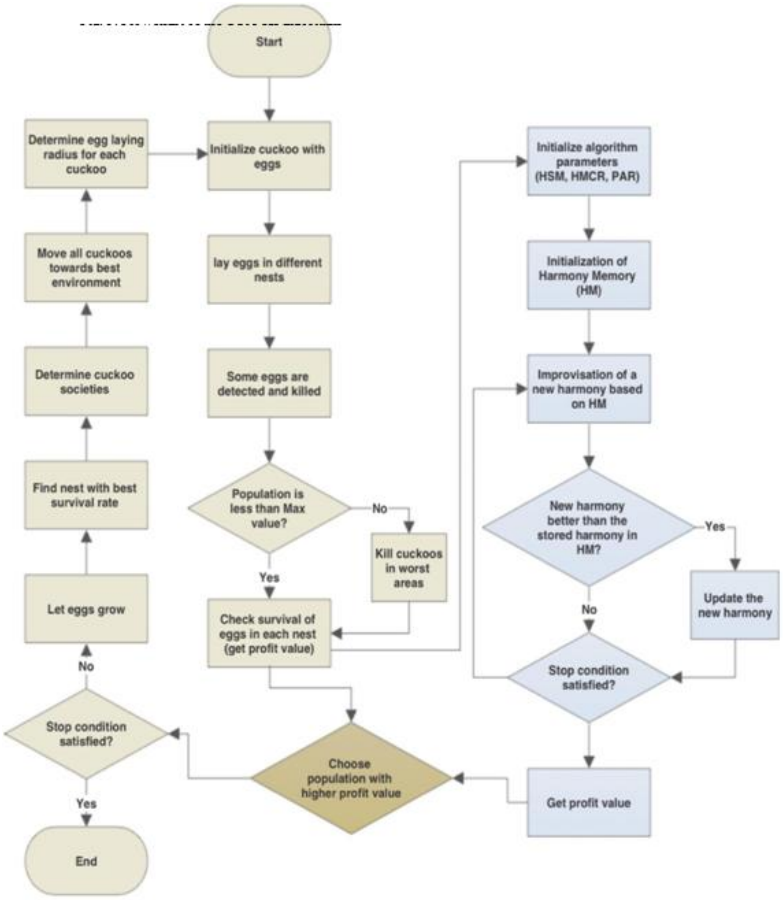


Fig.5.FlowchartoftheCOA-HSalgorithm

IV. Results

To evaluate the lesser number of genes that can best distinguish between two classes of cancer, the number of candidate genes was first decrease to 50 using MRMR. These 50 genes were then fed into our proposed algorithm, COA-HS, to find the better genes while using the highest accuracy. The accuracy of the SVM classifier was measured using the LOOCV method.Fig. 7 illustrates the accuracy of the SVM classifier for the top 50 genes selected via MRMR. In general, as the number of genes increases, the accuracy increases. However, in some instances, the classification accuracy was decreased as the number of genes increased. For example, in the case of the prostate cancer dataset, the classification accuracy for the first 8 genes was 97%, but the accuracy was decreased as the number of genes increased, achieving values of 91–93% when 90–100 genes were utilized.

Table 4

Most informative genes selected using the COA-HS algorithm that achieve 100% classification accuracy for all microarray datasets.

Dataset	Predictive gene
Leukemia	MPO myeloperoxidase
	Oncoprotein 18 (Op18) gene
	Proteasome iota chain
	ARHG ras homolog gene family, member G (rho G)
	LYN V-yes-1 yamaguchi sarcoma viral related oncogene
	OBF-1 mRNA for octamer binding factor 1
Lymphoma	UG Hs.120716 ESTs
	UG Ha. 1 69081 ets variant gene 6 (TEL oncogene)
	MCL1 myeloid cell differentiation protein

	Prostate	(40508_at) Glutathione S-transferase alpha 4	
		(37639_at) Hepsin	
		(769_s_at) Annexin A2	
		(1980_s_at) Non-metastatic cells 2, protein (NM23B)	
		(41661_at) Cadherin, EGF LAG seven-pass G-type receptor 1	
	Colon	H. sapiens RON mRNA for tyrosine kinase	
		H. sapiens mRNA for metallothionein (HUMAN)	
Fig. 10. Cost minimization over 100 iterations for prostate cancer.		H. sapiens pterin-4a-carbinolamine dehydratase (PCBD)	
		RETINOL-BINDING PROTEIN I, CELLULAR	
		Human galactokinase (galK) mRNA, complete cds	
		accuracy, which was in each case fewer than 100%. For the lym-phoma dataset, selecting 2 genes by the COA-HS algorithm lead to 100% accuracy. However, other algorithms required a markedly higher number of genes to obtain 100% accuracy. Table 4 lists the genes selected by the COA-HS algorithm that provided 100% classification accuracy of the SVM classifier for each of the datasets. To investigate the relevancy of each selected gene to the cancer type, their functions were investigated and are described below. MPO Myeloperoxidase is the hallmark enzyme of the myeloid lineage. The diagnosis of acute myeloid leukemia (AML) is easy if more than 3% of blast cells are confirmed to be cytochemically	

Table 3

Accuracy (AC), sensitivity (SE), specificity (SP), and number of selected genes (#Genes) via 5 optimization algorithms when combined with the SVM as a classifier for four microarray datasets.

	Leukemia		Prostate		Lymphoma		Colon	
	#Genes	AC/SE/SP	#Genes	AC/SE/SP	#Genes	AC/SE/SP	#Genes	AC/SE/SP
GA	24	100/100/100	28	98.04/91.8/100	22	100/100/100	14	95.16/84.6/100
PSO	16	100/100/100	19	98.04/91.8/100	10	100/100/100	11	96.42/85.8/100
HS	25	100/100/100	33	98.04/91.8/100	21	100/100/100	12	95.16/84.6/100
COA	15	100/100/100	12	99.07/95.2/100	7	100/100/100	12	96.77/87.3/100
COA-HS	6	100/100/100	5	100/100/100	3	100/100/100	5	100/100/100

V. Conclusion

A dual-level gene selection process that uses MRMR and the COA-HS algorithm was proposed to minimize the number of genes that gives 100% accuracy in cancer classification. To this end, MRMR was first used to decrease the number of genes to 50 so that the perfect time for an optimization algorithm would be decreased. The 50 several of genes were collected from UCI Repository then used as inputs for the second stage of gene selection, during which COA-HS was integrated with the SVM classifier and acted as a wrapper gene selection method. The LOOCV technique was implemented to evaluate the improvement of our proposed technique, and the results were verified to those other optimization algorithms, such as PSO, GA, HS, and COA.The cost minimization plots (Fig. 8–11) illustrate that the COA-HS other optimization algorithms in

reaching a better global minimum for all of the cancer datasets examined in this study. Furthermore, as seen in Table 3, the COA-HS reached 100% accuracy with the minimum number of genes for all datasets com-pared to the other algorithms. The functions of the selected genes were further investigated, and it was confirmed that the selected genes are biologically relevant to each type of cancer. Therefore, The dual-level gene selection via COA-HS can select highly informative biomarker genes to achieve 100% accuracy in distinguishing a second-class cancer classification task.

References

[1]. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537, http://dx.doi.org/ 10.1126/science.286.5439.531.

[2]. E. Blaveri, J.P. Simko, J.E. Korkola, J.L. Brewer, F. Baehner, K. Mehta, et al., Bladder cancer outcome and subtype classification by gene expression, Clin. Cancer Res. 11 (2005) 4044–4055, http://dx.doi.org/10.1158/1078-0432.CCR-04-2409.

[3]. Z. Cai, R. Goebel, M.R. Salavatipour, G. Lin, Selecting dissimilar genes for multi-class classification, an application in cancer subtyping, BMC Bioinformatics 8 (2007) 206, http://dx.doi.org/10.1186/1471-2105-8-206.

[4]. R. Wesolowski, B. Ramaswamy, Gene expression profiling: changing face of breast cancer classification and management, Gene Expr. 15 (2011) 105–115, http://dx.doi.org/10.3727/105221611X13176664479241.

[5]. H. Hijazi, C. Chan, A classification framework applied to cancer gene expression profiles, J. Healthc. Eng. 4 (2013) 255–283, http://dx.doi.org/10.1260/2040-2295.4.2.255.

[6]. A. Antoniadis, S. Lambert-Lacroix, F. Leblanc, Effective dimension reduction methods for tumor classification using gene expression data, Bioinformatics 19 (2003) 563–570, http://dx.doi.org/10.1093/bioinformatics/btg062.

[7]. J. Cao, L. Zhang, B. Wang, F. Li, J. Yang, A fast gene selection method for multi-cancer classification using multiple support vector data description, J. Biomed. Inform. 53 (2015) 381–389, http://dx.doi.org/10.1016/j.jbi.2014.12.009.

[8]. A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 153–158, http:// dx.doi.org/10.1109/34.574797.

[9]. C.E. Gillies, M.R. Siadat, N.V. Patel, G.D. Wilson, A simulation to analyze feature selection methods utilizing gene ontology for gene expression classification, J.Biomed.Inform.46(2013)1044–1059, http://dx.doi.org/10.1016/j.

[10]. jbi.2013.07.008.