Journal of Medical Systems (2019) 43: 235
https://doi.org/10.1007/s10916-019-1372-8

IMAGE & SIGNAL PROCESSING

Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification

M. Jansi Rani[1] · D. Devaraj[2]

Received: 22 March 2019 / Accepted: 5 June 2019 / Published online: 17 June 2019

Abstract
Cancer is a deadly disease which requires a very complex and costly treatment. Microarray data classification plays an important role in cancer treatment. An efficient gene selection technique to select the more promising genes is necessary for cancer classification. Here, we propose a Two-stage MI-GA Gene Selection algorithm for selecting informative genes in cancer data classification. In the first stage, Mutual Information based gene selection is applied which selects only the genes that have high information related to the cancer. The genes which have high mutual information value are given as input to the second stage. The Genetic Algorithm based gene selection is applied in the second stage to identify and select the optimal set of genes required for accurate classification. For classification, Support Vector Machine (SVM) is used. The proposed MI-GA gene selection approach is applied to Colon, Lung and Ovarian cancer datasets and the results show that the proposed gene selection approach results in higher classification accuracy compared to the existing methods.

Keywords Data mining · Cancer data classification · Gene selection · Genetic algorithm · Mutual information

Introduction

Diagnosis of tumor or cancer at the early stage is vital for the treatment of cancer [1]. Development of microarray data [2] samples and the application of microarray data analysis techniques play a major role in diagnosis and treatment of cancer. Microarray data or microarray dataset is a numerical representation of microarray in the form of a data sheet or a data file. Data mining techniques can be applied on this microarray data to extract useful information and can be applied for genetic analysis or research.

Cancer data samples contain large number of genes most of which are either redundant or useless or sometimes both [3]. Since it has high dimensions, cancer data classification is difficult. The number of genes in a microarray data is relatively higher than the number of samples by many folds and this result in the problem of over-fitting of data [4]. To improve the accuracy in classification and detection of cancer samples, suitable genes should be selected. Hence, cancer classification is done in two-stages; gene selection and then classification. Identifying the relevant genes is a critical task as it is difficult to obtain deeper information related to the genes.

Gene selection techniques in general are grouped into three, namely: (1) Filter approach, (2) Wrapper approach and (3) Hybrid approach. The filter-based approaches [5] select genes based on the general characteristics of the data while considering each gene separately. Filter approaches ignore the gene-to-gene dependencies in most cases and no classification or evaluation is done when selecting the genes. On the other hand, the wrapper approaches [6] take into consideration the gene-to-gene dependencies and also use a classification model to evaluate the various gene subsets before selecting the most promising gene subset. The filter approaches are less complex and easier to implement but they provide lesser classification performance. The wrapper approaches provide better classification but suffer from high

This article is part of the Topical Collection on *Image & Signal Processing*

✉ M. Jansi Rani
jansisujan@gmail.com

D. Devaraj
deva230@yahoo.com

[1] School of Computing, Kalasalingam Academy of Research and Education, Krishnankoil, Virudhunagar, India

[2] School of Electronics & Electrical Technology, Kalasalingam Academy of Research and Education, Krishnankoil, Virudhunagar, India

complexity and data over-fitting issue [7]. The hybrid approaches [8, 9] combine the advantages of both filter and wrapper approaches with better performance and less complexity.

Thanh Nguyen et al., [10] proposed a modified Analytic Hierarchy Process (AHP) that uses multiple gene ranking schemes to obtain an aggregate gene ranking that is used for gene selection. Selected genes were fed to the Hidden Markov Model (HMM) for classification. The Modified AHP not only improves the performance of HMM but also six other existing classifiers. The quality of the gene subset taken depends on gene-to-gene dependencies and gene-to-class relevance [11] and different combinations of genes provide different results. To calculate such dependencies and relevance, certain supervised and unsupervised measures have been used [12]. Algorithms such as the Tree Harvesting [13] use unsupervised similarity measures to identify similar genes. Supervised techniques on the other hand use predictive scores such as Wilcoxon Test [14], t-Test Statistics [15], Cox Model Score Test [16], Entropy [15], etc. Gene Clustering [17] and Gene Shaving [18] are some of the most used supervised techniques for estimating the dependencies between genes. Since all these measures rely on the actual microarray data, they are prone to noisy data and outliers.

Use of efficient learning strategies [19] can show deeper characteristics about genes that can be used for efficient classification. Recently, evolutionary algorithms [20] have been applied for gene selection process. Particle Swarm Optimization (PSO) [21], Artificial Bee Colony Algorithm (ABC) [22, 23] and Genetic Algorithm (GA) [23, 24] have been applied for gene selection, Artificial Neural Networks (ANN) [20, 22], and Fuzzy Logic System (FLS), [25] have been used for classification. Genetic Algorithm (GA) is a well-known and most used algorithm for identifying the most suitable solution from a population of given solutions; which in this case is the most promising gene subset [26]. The evolutionary operations of GA can accurately identify the most suitable genes from the microarray data.

Use of swarm optimization algorithms [27] is one of the efficient strategies for gene selection and has given rise to many gene selection techniques in recent times. Swarm optimization works on natural principle of swarm intelligence of nature to converge on the most optimal solution. Swarm optimization algorithms such as Ant Colony Optimization (ACO) [28], Artificial Bee Colony (ABC) [29], Particle Swarm Optimization (PSO) [21], Bat Algorithm (BAT) [30], etc. have been used for gene selection and dimensionality reduction. Swarm optimization algorithms are relatively simple and efficient for gene selection compared to other complex approaches such as the Random Forest (RF) [31], Artificial Neural Networks (ANNs) [22] and Fuzzy Logic System (FLS). But hybrid techniques using swarm intelligence provide promising results in gene selection such as the rough set approach proposed by Suguna et al. [32] that uses the Artificial Bee Colony (ABC) with Fuzzy Logic. Similarly, Shokouhifar et al. [33] proposed a hybrid gene selection strategy by combining the ABC with that of the Neural Networks (NN) that provides better results compared to standard ABC approach. Recent times also focus the use of Deep Learning algorithms for cancer classification as in [34, 35].

Both filter and wrapper approach are used for gene selection, such as the wrapper approach proposed by Nakamura et al. [31] with Bat Algorithm and the filter approach proposed by Xue et al. [36] using Particle Swarm Optimization (PSO). Due to better efficiency in classification, the wrapper approaches have been preferred mostly especially techniques using PSO have been studied largely due to its less complexity and better characterization using lesser number of parameters [37].

Gene selection also features certain interesting hybrid algorithms such as the one proposed by Hala M. Alshamlan et al., [24] that use a unique Genetic Bee Colony (GBC) algorithm for selecting informative genes for classification by combining the Genetic Algorithm (GA) and Artificial Bee Colony (ABC) algorithms. By combining the characteristics of both the algorithms, the GBC algorithm was able to handle both binary-class and multiple-class cancer datasets with better accuracy in both [38]. Advanced Neural Network based learning models also provide efficient gene selection techniques.

The complexity of Genetic Algorithm and the possibility of large population size cannot be handled due to the high dimensionality of microarray data [39]. A less complex solution is proposed here that considers only the most informative genes from the original microarray data for population generation in Genetic Algorithm. For selecting the most informative genes the Mutual Information measure is used [40]. As Mutual Information depends only on the probability distribution of the random variable, it is not affected by noisy data and outliers like the other measures [41].

In this regard, we propose a hybrid two-stage gene selection algorithm in this paper with the aim of obtaining maximum cancer classification accuracy by combining the characteristics of Genetic Algorithm (GA) with the Mutual Information (MI) measure. The first stage of proposed MI-GA algorithm uses Mutual Information to select the set of relevant genes suitable for classification. In the second stage, the Genetic Algorithm is used to identify the most promising gene subset from all candidates obtained using the selected genes from first stage. To further improve the classification accuracy, SVM classifier is used to effectively classify the cancer samples.

The rest of this paper is organized as follows. Section 2 provides the detailed state-of-art in gene selection and classification including recent works. The preliminaries related to the proposed method are discussed in Section 3. The proposed gene selection and classification algorithms are explained in Section 4. Experimental results with three microarray data and

the inferences made are discussed in Section 5. Finally, Section 6 offers concluding remarks with possible future work.

Proposed cancer classification approach

Classification of cancer samples is a critical process as it depends on the combination of genes used for the classification. The proposed cancer classification approach In the first stage, MI based algorithm is applied to select a subset of informative features and GA is applied over this subset to identify the optimal features. Support Vector Machine (SVM) based classifier uses hyperplanes to identify the type of class for each sample within the microarray data. In the proposed cancer classification strategy "No Cancer" is represented by '0' and "Cancer" data is represented by '1'. Figure 1 shows the overall process of the proposed cancer classification.

The obtained gene subset from the MI-GA approach is processed using SVM separately with various kernel functions.

Different microarray cancer data will be difficult to analyze and understand all the patterns using the same kernel function. In order to make the classification more suitable and flexible for different microarray data, multiple kernel functions have been used and the most suitable one for the given cancer microarray data will be identified and chosen for classification.

Proposed two stage gene selection algorithm

Consider a cancer microarray data $M(S, G, C)$ that consists of n samples, d genes and k class labels in such a way that the sample set S is represented as $S = \{s_1, s_2, \ldots, s_n\}$ and the set of all genes in the data is represented as $G = \{g_1, g_2, \ldots, g_d\}$. The set of class labels is represented as $C = \{c_1, c_2, \ldots, c_k\}$ where k = number of class labels within the class. The value of

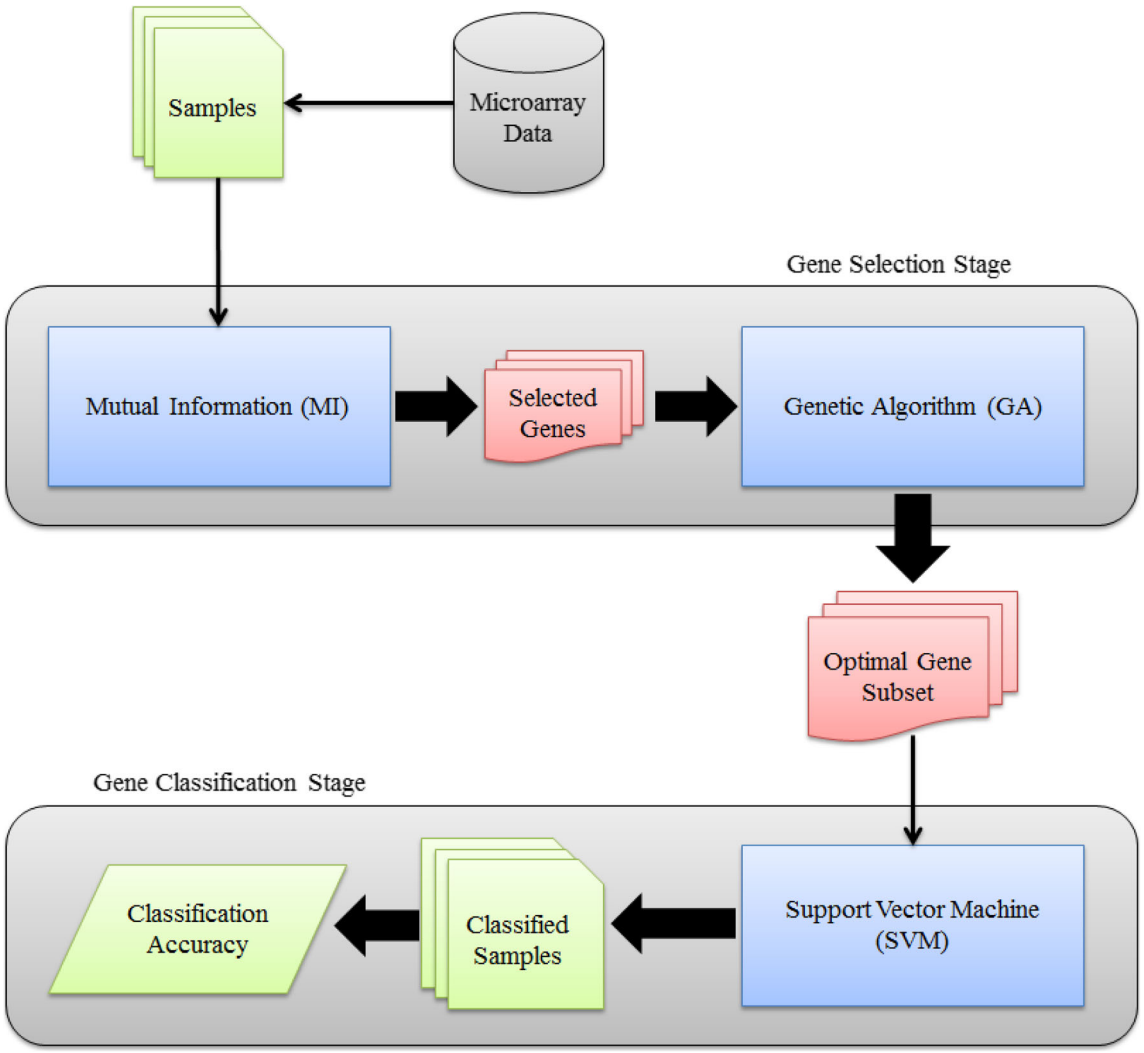

Fig. 1 Process flow of proposed cancer classification approach

the genes inside any given sample S_i is represented as $\{g_1^i, g_2^i, \ldots, g_d^i\}$ where $i = 1, 2, \ldots, n$.

The process flow of the two-stage MI-GA gene selection process is shown in Fig. 2. In the stage1, the important features are selected based on the Mutual Information value. The Mutual Information based gene selection is a filter approach that selects the genes that have a higher information gain with the cancer class. The selected genes are fed to the Genetic Algorithm based gene selection which is a wrapper based approach which identifies the most promising gene subset. Since redundant and useless genes are removed in the first stage of gene selection, the obtained gene subset from second stage contains only the most informative genes.

The objective of the proposed method are: (1) To select the most suitable gene subset from any microarray cancer data, and (2) To classify the cancer samples accurately using the selected gene subset to obtain better classification accuracy in any given microarray cancer data.

The obtained gene subset from the hybrid two-stage gene selection algorithm is then given to the final classification stage. The details of the MI-GA based gene selection are presented in the following sub sections.

Informative gene selection using mutual information

According to probability theory, the Mutual Information between any two random variables is the dependency of one variable over the other in a mutual aspect that shows the amount of information that can be obtained about one variable if the other is known. In a given system with two random discrete variables X and Y, where X is the input variable with N_X possible values $x \in X$ and Y is the output variable with N_Y possible values $y \in Y$. The amount of uncertainty is the measure of Mutual Information that can be related to entropy. Mutual Information between these variables, denoted as $I(X;$

Fig. 2 Process flow of Gene Selection stage

Y) can be expressed using entropy and conditional entropy as given below.

$$I(X;Y) = H(Y) - H(Y|X) \quad (1)$$

Here, $H(Y)$ is the amount of uncertainty in the output variable Y as expressed in Eq. (2) and $H(Y|X)$ is the amount of uncertainty still retaining in the output variable Y if input variable X is known as expressed in Eq. (3). The difference between the two provides the amount of information gained about the output variable Y if the input variable X is known.

$$H(Y) = -\sum_{j=1}^{N_Y} P\left(y_j\right) * log\left[P\left(y_j\right)\right] \quad (2)$$

Here, $P(y_j)$ is the probability of occurrence of the event $Y = y_j$ and $P(x_i)$ is the probability of occurrence of the event $X = x_i$ in the given data. $P(y_j|x_i)$is the conditional probability that denotes the probability of occurrence of the event $Y = y_j$ when given the event $X = x_i$.

$$H(X|Y) = -\sum_{i=1}^{N_X} P(x_i) * \left[\sum_{j=1}^{N_Y} P\left(y_j|x_i\right) * log\left[P\left(y_j|x_i\right)\right]\right] \quad (3)$$

The higher mutual information denotes a higher knowledge about the output variable and lesser uncertainty. In the case of microarray cancer data the genes with higher mutual information provides better information for accurate classification.

Genetic algorithm based feature selection

In the second stage, the process of selecting the important features is formulated as an optimization problem with maximization of accuracy as the objective. This is mathematically stated as

$$Maximize(CC) \quad (4)$$

Where CC is the number of correctly classified samples.

We apply Genetic Algorithm to solve this optimization problem. Genetic Algorithm is a meta-heuristic algorithm that works based on the process of natural selection to obtain optimal and high-quality solutions by applying genetic operators namely, Selection, crossover and mutation to produce offspring. The set of possible solutions to the given problem are called as the initial population and each member of the population is called as chromosome. A chromosome is built using a collection of genes and all chromosomes have the same number and type of genes. Multiple genes are selected from the population that are genetically bred using crossover and mutation with the aim to obtain next generation offspring with better fitness during the optimization process. Crossover and mutation are the two operators that are used for generating the new population. In crossover process, two chromosomes (parents) are combined together to form new chromosomes, called offspring. The crossover operator is applied repeatedly; genes of good chromosomes are appearing in the population. Crossover is usually applied in a GA with a high probability. Mutation plays an important role in GA. Some random changes occurs in the characteristics of chromosomes when applying the mutation operator.

Many generations of GA run is required till a solution with maximum fitness is obtained and this is returned as the most optimum solution for the problem.

The standard Genetic Algorithm requires two criteria to be followed before implementing the optimization process; (1) Genetic Representation of the domain that contains the list of possible solutions and (2) A Fitness Function is needed to evaluate the solutions. The most common representation adopter is using a bit array. The fitness function is formed from objective function that defines the closeness of a solution to achieve the desired goals. The efficiency in designing the fitness function decides the ability of the Genetic Algorithm to converge to the appropriate solution.

A gene subset with high classification accuracy is said to have better fitness. The fitness of gene subset S is calculated using,

$$\text{Fitness(S)} = \frac{\text{Claification Accuracy of dataset M when gene subset S is used}}{100} \quad (5)$$

After computing the fitness value, the three genetic operators namely, selection, crossover and mutation will be applied repeatedly until the condition is reached. The details of the genetic operators are given in [42, 43]. 50 features will be selected through the Mutual Information process and that will be given as input to the Genetic Algorithm. At the end of the Genetic Algorithm process, 10/20 features will be selected. In the result section, these are discussed in detail.

Support vector machine classifier

The Support Vector Machine is a supervised learning algorithm used for the analysis of data during classification. The SVM processes the training dataset or the training data that contains n data points that are in format$\left(\vec{x}_1, y_1\right)$, $\left(\vec{x}_2, y_2\right), \ldots, \left(\vec{x}_n, y_n\right) \forall i = 1, 2, \ldots, n$. Here $\vec{x}_i$ denotes the i^{th} data point and is represented as a p-dimensional real vector that is generated by using the samples from the input dataset. y_i denotes the class to which the vector $\vec{x}_i$ belongs and this is represented as 0 and 1. The goal of SVM is to identify the hyperplane that divides the set of all vectors that belong to $y_i = 0$ and $y_i = 1$ that have the maximum distance between the hyperplane and the

Table 1 Test dataset details

Dataset	Number of Samples	Number of Genes	Training Samples	Testing Samples
Colon Cancer	62	2000	31	31
Lung Cancer	203	12,600	101	102
Ovarian Cancer	253	15,154	126	127

nearest point from both the groups. The hyperplane for a set of vectors $\vec{x}$ is expressed as given in Eq. (6),

$$\vec{w}.\vec{x} - b = 0 \tag{6}$$

Here $\vec{w}$ is the normalized vector to the hyperplane and b represents the bias value that is used to define the offset and angle of the hyperplane. For defining the offset b of the hyperplane with respect to the origin along the normal vector $\vec{w}$ the $b/\|\vec{w}\|$ parameter is used.

The hyperplane is generated by using a specific kernel function that is used to represent the vectors from the input dataset. The most commonly and used kernel functions for learning are: (1) Linear Function– $k(x_i, x_j) = x_i.\ x_j$, (2) Polynomial Function – $k(x_i, x_j) = [x_i.\ x_j + c]^d$, (3) Quadratic Function – $k(x_i, x_j) = [x_i.\ x_j + c]^2$, (4) MLP Function – $k(x_i, y_i) = tanh\,(kx_i.\ x_j + c)$and (5) Radial Bias Function – $k(x_i, x_j) = exp\,(-\gamma|x_i - x_j|^2)$. where K is the kernel function, x_i, x_j are n dimensional inputs, d is the degree of the polynomial. c is a constant that allows to trade off the influence of the higher order and lower order terms. In this paper, the SVM is trained separately with all the 5 kernel functions.

Results and discussion

The proposed approach has been applied for gene selection in three microarray datasets given in Table 1. The split up of data into training and testing data has been given in Table 1. The data has been split in such a way that the training data size is equal to or greater than the testing data size. This is done to make sure there is sufficient amount of data available for training.

The implementation has been done using the MATLAB environment installed in a system with Intel Core i7 Processor, 16 Giga Byte of RAM, 1 Tera Byte of Hard Disk and having 64-bit Windows 10 Operating System.

In Genetic Algorithm (GA), the following values have been used for Mutation and Crossover.

Population size = 20.

Mutation rate = 0.05.

Crossover rate = 0.6.

Total number of generations r = 20.

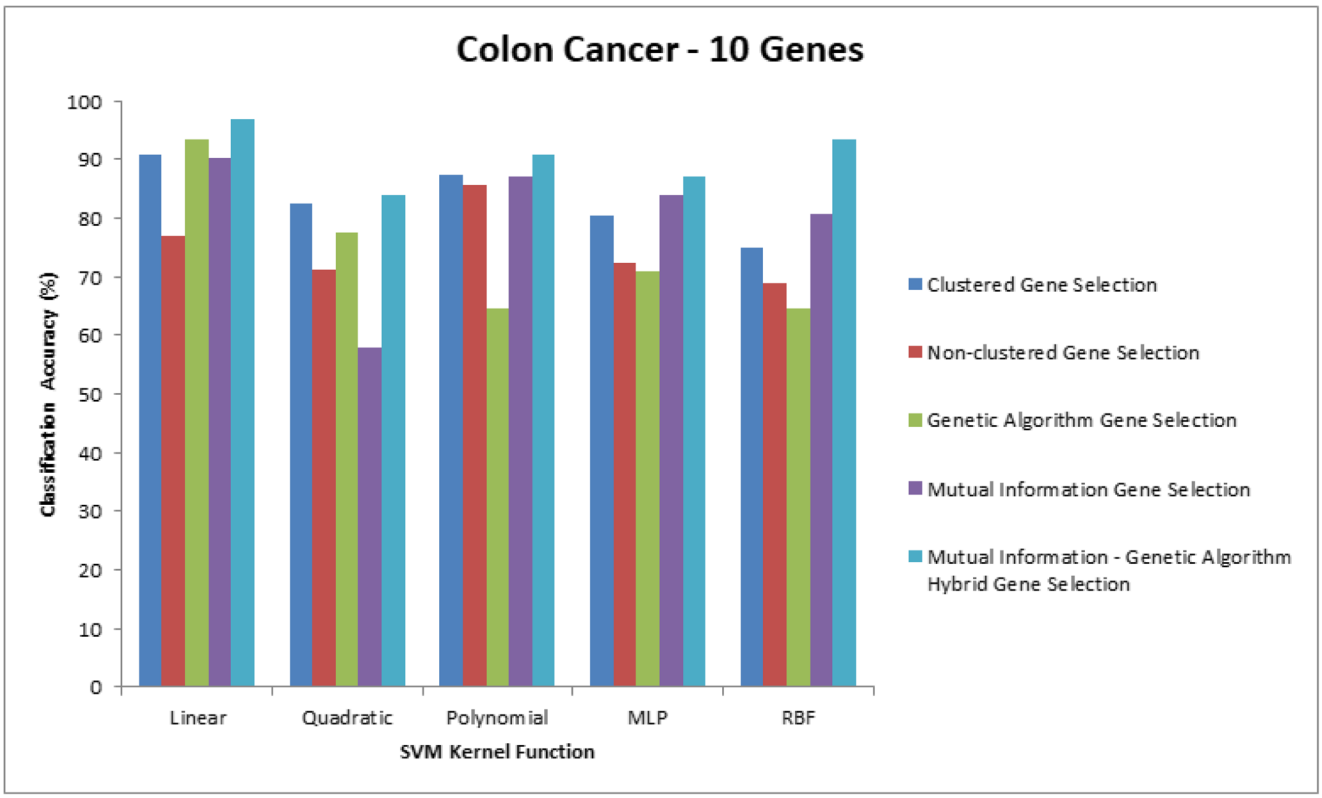

Fig. 3 Comparison of classification accuracy of Colon Cancer dataset using 10 genes

Initially the Colon Cancer dataset is used to test the effectiveness of the proposed MI-GA gene selection algorithm. All five SVM kernel functions have been used for Classification. To obtain the optimal gene subset 50 genes were selected from the MI stage and given to the GA. For comparison gene selection was done using 4 existing approaches and selected genes were given to the classifier. The 4 existing methods used for comparison are Clustered Gene Selection (C-Sel), Non-clustered Gene Selection (NC-Sel), Genetic Algorithm based Gene Selection (GA-Sel) and Mutual Information based Gene Selection (MI-Sel).

Figure 3 shows the result obtained by the various approaches. It can be seen that the proposed two-stage MI-GA gene selection approach produces better classification accuracy in Colon Cancer dataset compared to the existing approaches when using all the five kernel functions in the SVM classifier. By using only 10 genes, the proposed approach was able to achieve a classification accuracy of 96.77% which is much higher than the existing approaches. Next, the proposed MI-GA gene selection approach is executed with 10 genes and 20 genes selected from the MI stage. In this case, the execution time is also recorded along with the classification accuracy. The experiment is conducted for each of the cancer microarray datasets that are mentioned in Table 1. The comparison of classification accuracy and execution time of MI-GA gene selection approach using all five kernel functions and by selecting different number of genes from the MI stage are displayed in Figs. 4, 5 and 6 for the Colon Cancer, Ovarian Cancer and Lung Cancer datasets respectively.

From Fig. 4 it can be obtained that the Colon Cancer microarray dataset obtains maximum classification accuracy of 100% when using the polynomial kernel function. The least accuracy is seen when using the MLP and RBF kernel functions and especially in case of MLP the accuracy is too low. This is because since RBF and MLP are neural network based functions they tend to deviate more in case of random gene values that are not normalized and stays in a large search space and this affects the overall accuracy of the Colon Cancer dataset. In case of execution time, average the MLP kernel function obtains better results compared to the other kernel functions. The Linear kernel function obtains higher execution time in case of Colon Cancer microarray dataset.

From Fig. 5 it can be seen that the linear kernel function obtains a maximum classification accuracy of 100% but still the other kernel functions also provide promising results. The execution time is also least in case of the linear kernel function but the Quadratic kernel function takes more execution time. In general, the proposed MI-GA gene selection performs well

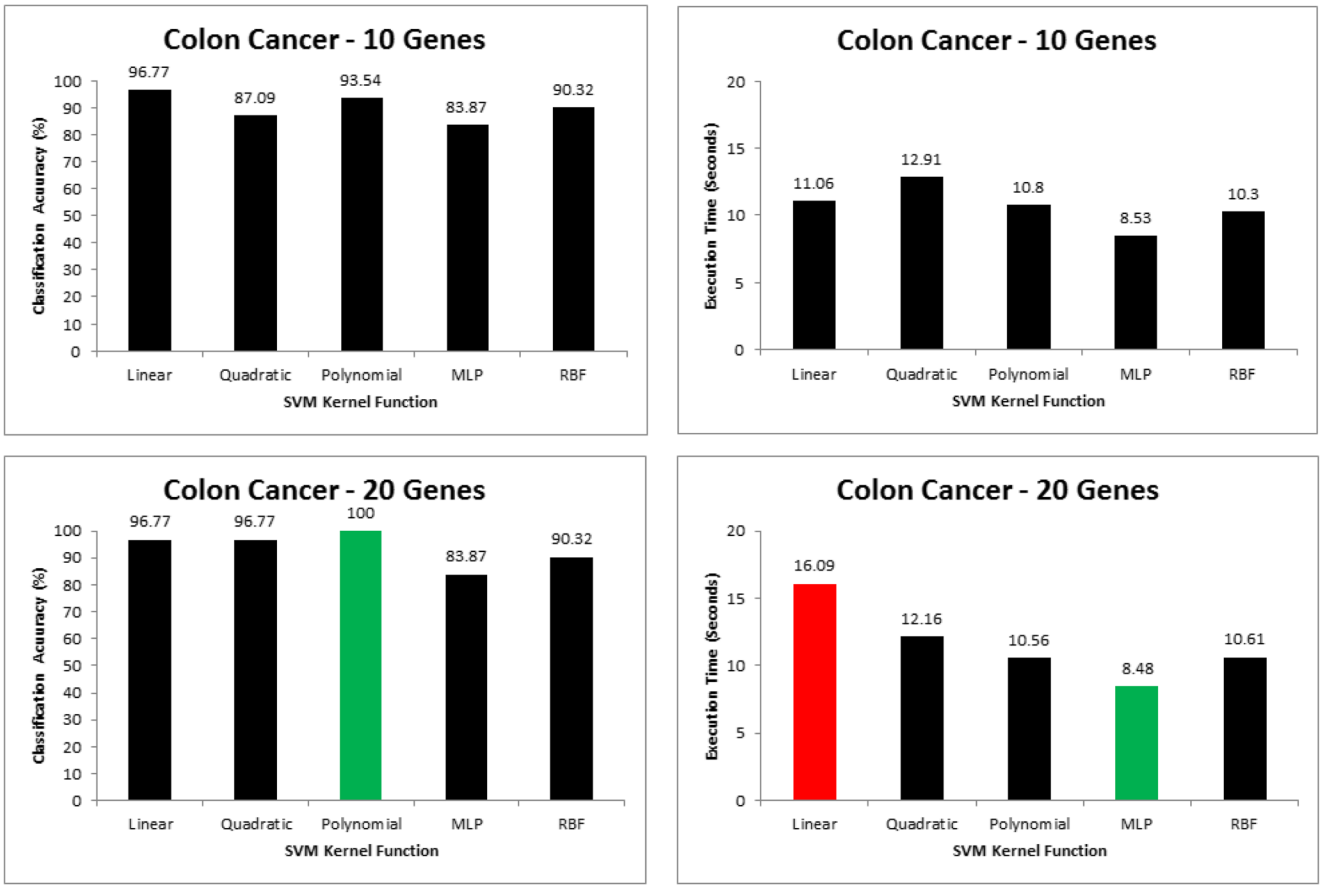

Fig. 4 Colon Cancer classification accuracy and execution time using MI-GA approach

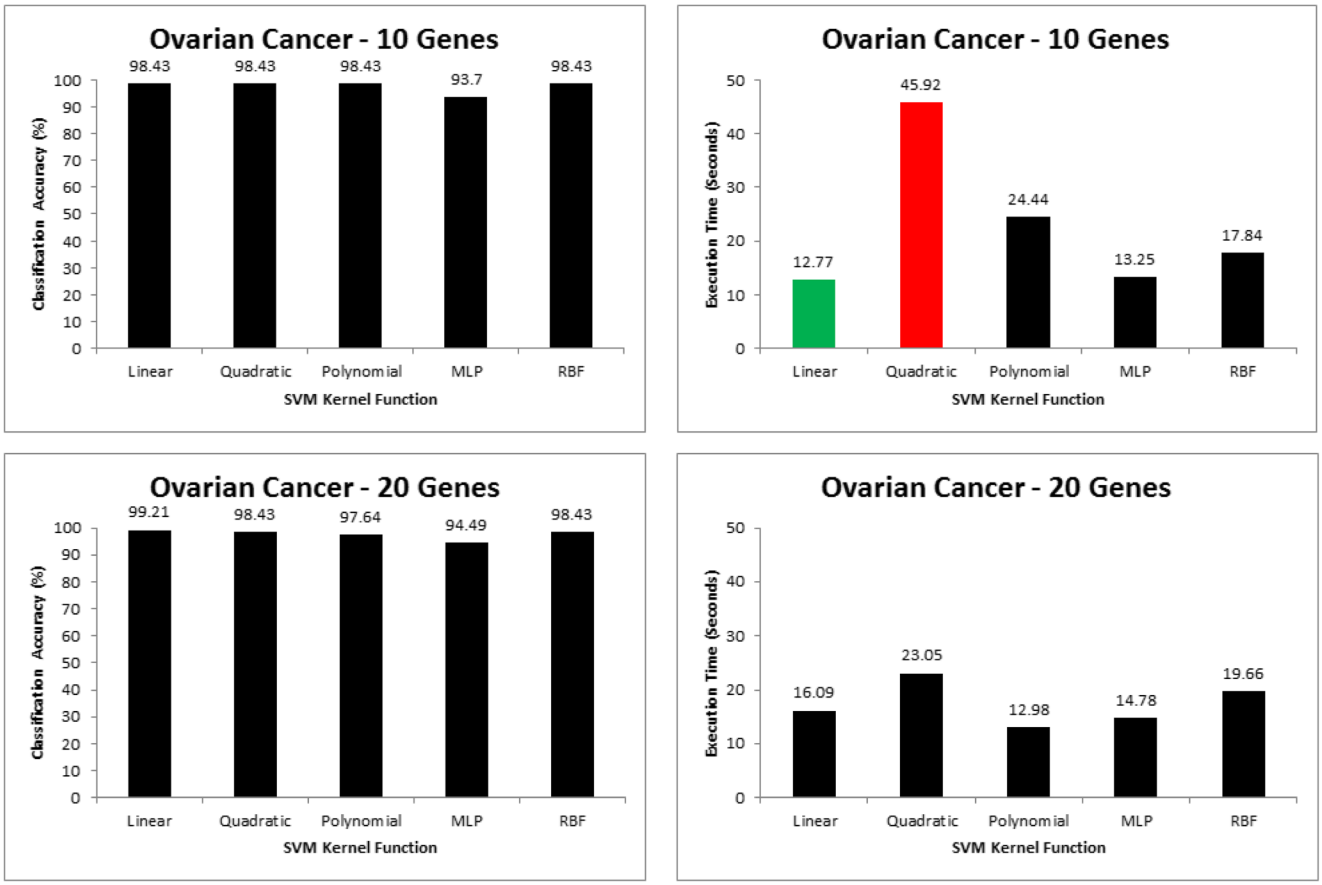

Fig. 5 Ovarian Cancer classification accuracy and execution time using MI-GA approach

in any SVM kernel for the Ovarian Cancer dataset. This is because the gene values in the Ovarian cancer dataset are normalized within the range 0 to 1 and this makes it easier for the SVM kernel to implement the learning process respective of the kernel function used.

Finally as shown in Fig. 6, for the Lung Cancer microarray dataset only obtain a maximum classification accuracy of 81.37% obtained when using Polynomial kernel function in SVM. Optimal results are obtained from Quadratic, RBF and Linear kernel functions also. The performance of MLP kernel function is not up to the mark. The Lung Cancer gene values are spread over a large search space and the values are not normalized. This makes the learning process more complex and reduces the accuracy. But still the proposed MI-GA gene selection approach obtains a maximum classification accuracy of 81.37% for the Lung Cancer dataset that is higher than many existing gene selection approaches. In terms of execution time the RBF and MLP functions perform better but this is not really important if the classification is not satisfactory.

Performance Evaluation

To demonstrate the efficiency of the proposed approach in microarray data classification, additional matrices like True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), True Negative Rate (TNR), Precision, Prevalence, Accuracy, F1 score are evaluated. True Positive is the case in which we predicted 1 and the actual output is also 1. True Negative is the case in which we predicted 0 and the actual output is 0. False Positive is the case in which we predicted 1 and the actual output is 0. False Negative is the case in which we predicted 0 and the actual output is 1.The equations for the above metrics are given below

$$Accuracy = \frac{Number\ of\ correctly\ Classified\ samples}{Total\ number\ of\ Samples} \quad (7)$$

$$True\ Positive\ Rate = \frac{True\ Positive}{False\ Negative + True\ Positive} \quad (8)$$

$$False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive + True\ Negative} \quad (9)$$

$$True\ Negative\ Rate = \frac{True\ Negative}{Falde\ Positive + True\ Negative} \quad (10)$$

$$False\ Negative\ Rate = \frac{False\ Negative}{True\ Positive + False\ Negative} \quad (11)$$

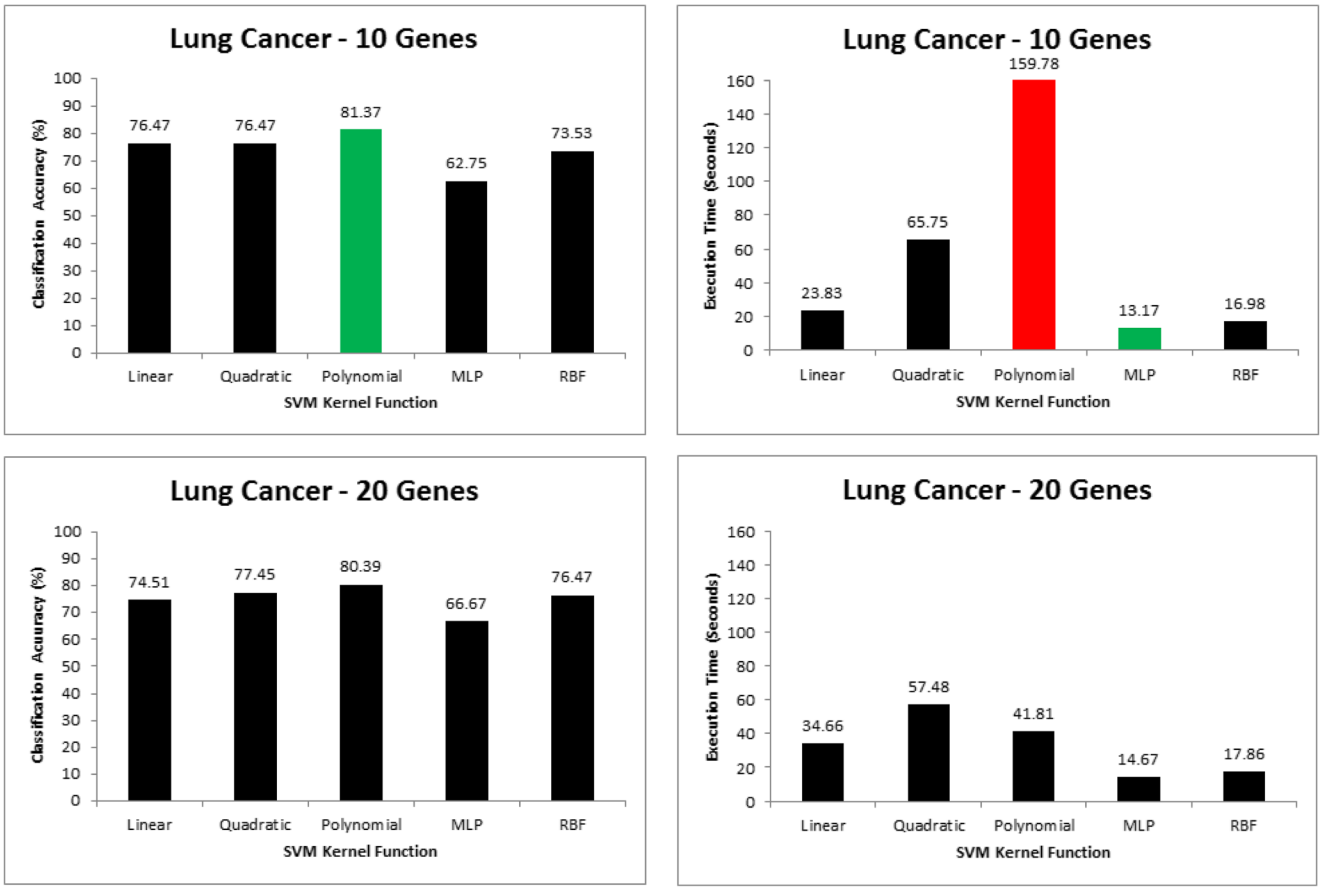

Fig. 6 Lung Cancer classification accuracy and execution time using MI-GA approach

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (12)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (13)$$

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (14)$$

The results are reported in Table 2.

From the table it is clear that the accuracy obtained for different datasets is more than 80% in all cases and for Colon Cancer and Ovarian Cancer more than 95% accuracy is obtained with only 10 and 20 genes. The precision value for them is also above 95% which indicates the accuracy in classification of the samples. This demonstrates the efficiency of the proposed hybrid gene selection algorithm for microarray data classification.

Table 2 Performance Evaluation of MI-GA Gene Selection using Confusion Matrix

Dataset	TPR	FPR	FNR	TNR	Precision	Prevalence	Accuracy	F1 Score
Colon Cancer 10 Genes Linear	0.9500	0	0.05	1.0	1.0	0.6452	0.9677	0.9744
Colon Cancer 20 Genes Polynomial	0.95	0.0909	0.05	0.9091	0.95	0.6452	1.0	0.95
Ovarian Cancer 10 Genes Linear	1.0	0.0435	0	0.9565	0.9759	0.6378	0.9842	0.9878
Ovarian Cancer 20 Genes Linear	1.0	0.0435	0	0.9565	0.9759	0.6378	0.9921	0.9878
Lung Cancer 10 Genes Polynomial	0.8286	0.2188	0.1714	0.7813	0.8923	0.6863	0.8137	0.8593
Lung Cancer 20 Genes Polynomial	0.8286	0.4688	01714	0.5313	0.794	0.6863	0.8039	0.8112

Conclusion

Classification of cancer samples is a complex task since it depends on many aspects such as the type of microarray DNA samples, the type of cancer, the number of genes used for classification and the information available within the gene subset used. We proposed and implemented an efficient gene selection approach for cancer classification in Colon Cancer, Lung Cancer and Ovarian Cancer microarray data. The proposed MI-GA gene selection approach uses a two-stage process that employs Mutual Information based gene selection in the first stage and Genetic Algorithm based gene selection in second stage. Efficiency of the proposed gene selection approach is verified by using the SVM based classifier that uses five variations and each variation uses different kernel functions. For the experimentation purpose, a total of three microarray cancer datasets were taken with different characteristics and class count. Both binary class and multi-class microarray datasets were used. Experimental results show that the proposed MI-GA gene selection approach performs better than the existing approaches in all the datasets and produces maximum classification accuracy. In future, the machine learning techniques such as Genetic Algorithm in combination with Fuzzy Logic and Neural Network can be applied to obtain better classification of microarray data set. This hybridization may help in reducing the complexity of the classification model and getting the better accuracy.

References

1. Reboiro, J. M., Arrais, J. P., Oliveira, J. L. et al., Gene committee: A web-based tool for extensively testing the discriminatory power of biologically relevant gene sets in microarray data classification. *BMC Bioinf.* 15(1):31, 2014.
2. Saber, H. B., and ELLOUMI, M., DNA microarray data analysis: A new survey on Biclustering. *International Journal for Computational Biology (IJCB)* 4(1):21–37, 2015.
3. Kirubakaran, R., Periya Nayaki, A., and Prathibhan, C. M., A survey on data mining in big data. *International Journal of Research and Scientific Innovation* III(IA):37–40, 2016.
4. Algamal, Z. Y., and Lee, M. H., Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional Cancer classification. *ELSEVIER Journal of Computers in Biology and Medicine* 67:136–145, 2015.
5. Ditzler, G., Polikar, R., and Rosen, G., A sequential learning approach for scaling up filter-based feature subset selection. *IEEE Transactions on Neural Networks and Learning Systems* PP(99): 1–15, 2017.
6. Ma, L., Li, M., Gao, Y., Chen, T., Ma, X., and Qu, L., A novel wrapper approach for feature selection in object-based image classification using polygon-based cross-validation. *IEEE Geoscience and Remote Sensing Letters* 14(3):409–413, 2017.
7. Leung, Y., and Hung, Y., A multi-filter-multi-wrapper approach to gene selection and microarray data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7(1): 108–117, 2010.
8. Motieghader, H., Najafi, A., Sadeghi, B., and Masoudi-Nejad, A., A hybrid gene selection algorithm for microarray Cancer classification using genetic algorithm and learning automata. *ELSEVIER, Informatics in Medicine Unlocked* 9:246–254, 2017.
9. Ray, S. S., Ganivada, A., and Pal, S. K., A granular self-organizing map for clustering of gene selection in microarray data. *IEEE Transactions on Neural Networks and Learning Systems* 27(9): 1890–1906, 2016.
10. Nguyen, T., and Nahavandi, S., Modified AHP for gene selection and Cancer classification using Type-2 fuzzy logic. *IEEE Transactions on Fuzzy Systems* 24(2):273–287, 2016.
11. Han, F., Yang, C., Wu, Y.-Q., Zhu, J.-S., Ling, Q.-H., Song, Y.-Q., and Huang, D.-S., A gene selection method for microarray data based on binary PSO encoding gene-to-class sensitivity information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14(1):85–96, 2017.
12. Li, J., Malley, J. D., Andrew, A. S., Karagas, M. R., Moore, J. H. Detecting gene-gene Interactions using a Permutation-based Random Forest Method, SPRINGER, BioData Mining, Volume 9, Issue 14, 2016.
13. Martin, C. W., Tauchen, A., Becker, A., Nattkemper, T. W. A Normalized Tree Index for Identification of Correlated Clinical Parameters in Microarray Experiments, SPRINGER BioData Mining, Volume 4, Issue 2, 2011.
14. Liao, C., Li, S., Luo, Z. Gene Selection for Cancer Classification using Wilcoxon Rank Sum Test and Support Vector Machine, IEEE International Conference on Computational Intelligence and Security, November 2006.
15. Jansi Rani, M., Devaraj, D. A Combined Clustering and Ranking based Gene Selection Algorithm for Microarray Data Classification, IEEE International Conference on Computational Intelligence and Computing Research.
16. Wan, Y-W, Nagorski, J., Allen, G. I., Li, Z., Liu, Z. Identifying Cancer Biomarkers Through a Network Regularized Cox Model, IEEE International Workshop on Genomic Signal Processing and Statistics, November 2013.
17. Paul, A. K., and Shill, P. C., Incorporating gene ontology into fuzzy relational clustering of microarray gene expression data. *ELSEVIER, Biosystems* 163:1–10, 2018.
18. Sheng, J., Deng, H.-W., Calhoun, V., and Wang, Y.-P., Integrated analysis of gene expression and copy number data on gene shaving using independent component analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8(6):1568–1579, 2011.
19. Du, W., Cao, Z., Song, T., Li, Y., Liang, Y. A Feature Selection Method based on Multiple Kernel Learning with Expression Profiles of Different Types, SPRINGER, BioData Mining, Volume 10, Issue 4, 2017.
20. Dashtban, M., and Balafar, M., Gene selection for microarray Cancer classification using a new evolutionary method employing artificial intelligence concepts. *ELSEVIER, Genomics* 109(2):91–107, 2017.
21. Jain, I., Jain, V. K., Jain, R. Correlation Feature Selection based improved-Binary Particle Swarm Optimization for Gene Selection and Cancer Classification, ELSEVIER, Applied Soft Computing, In Press, 2017.
22. Garro, B. A., Rodriguez, K., and Vazquez, R. A., Classification of DNA microarrays using artificial neural networks and ABC algorithm. *ELSEVIER, Applied Soft Computing* 38:548–560, 2016.
23. Alshamlan, H. M., Badr, G. H., and Alohali, Y. A., Genetic bee Colony (GBC) algorithm: A new gene selection method for microarray Cancer classification. *ELSEVIER, Computational Biology and Chemistry* 56:49–60, 2015.
24. Peng, S., Xu, Q., Ling, X. B., Peng, X., Du, W., and Chen, L., Molecular classification of Cancer types from microarray data using

the combination of genetic algorithms and Support vector machines. *ELSEVIER, FEBS Letters* 555(2):358–362, 2003.
25. Nilashi, M., Ibrahim, O., Ahmadi, H., and Shahmoradi, L., A knowledge-based system for breast Cancer classification using fuzzy logic method. *ELSEVIER, Telematics and Informatics* 34(4):133–144, 2017.
26. Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., and Hermann, B., Frieboes; "prediction of lung Cancer patient survival via supervised machine learning classification techniques". *ELSEVIER, International Journal of Medical Informatics* 108:1–8, 2017.
27. Jin, C., and Jin, S.-W., Gene selection approach based on improved swarm intelligent optimization algorithm for tumour classification. *IET Systems Biology* 10(3):107–115, 2016.
28. Yan, Z., Yuan, C., in Biometric Authentication, First International Conference, ICBA 2004, Hong Kong, China, July 15–17 2004. Lecture Notes in Computer Science, ed. by D Zhang, AK Jain. Ant colony optimization for feature selection in face recognition (Springer, Berlin, 2004), pp. 15–17.
29. Karaboga, D., Gorkemli, B., Ozturk, C., Karaboga, N. A comprehensive survey: artificial bee colony (ABC) algorithm and applications. Artif. Int. Rev. (2012).
30. Nakamura, R., Pereira, L., Costa, K., Rodrigues, D., Papa, J., in SIBGRAPI Conference on Graphics, Patterns and Images, BBA: a binary bat algorithmfor feature selection, (OuroPreto, 22–25 2012).
31. Zhou, Q., Zhou, H., and Li, T., Cost-sensitive feature selection using random Forest: Selecting low-cost subsets of informative features. *ELSEVIER, Knowledge-based Systems* 95:1–11, 2016.
32. Suguna, N., and Thanushkodi, K., An independent rough set approach hybrid with artificial bee colony algorithm for dimensionality reduction. *Am. J. Appl. Sci.* 8(3):261–266, 2011.
33. Shokouhifar, M., Sabet, S., in 3rd International Conference on Machine Vision, Hybrid approach for effective feature selection using neural networks andartificial bee colony optimization (IEEE, Piscataway, 2010), pp. 502–506.
34. Guillen, P., Ebalunode, J. Cancer Classification based on Microarray Gene Expression Data using Deep Learning, IEEE International Conference on Computational Science and Computational Intelligence, December 2016.
35. Ahmed M. Abdel-Zaher, Ayman M. Eldeib; "Breast Cancer Classification using Deep Belief Networks", ELSEVIER, Expert Systems with Applications, Volume 46, pp. 139–144.
36. Xue, B., Cervante, L., Shang, L., and Zhang, M., A particle swarm optimization based multi-objective filter approach to feature selection for classification. *Artif. Intell. Rev.* 7458:673–685, 2012.
37. Chen, B., Chen, L., and Chen, Y., Efficient ant colony optimization for image feature selection. *Signal Proc.* 93(6):1566–1576, 2013.
38. Lotfi, E., and Keshavarz, A., Gene expression microarray classification using PCA-BEL. *ELSEVIER, Computers in Biology and Medicine* 54:180–187, 2014.
39. Taguchi, Y-h. Principle Component Analysis based Unsupervised Feature Extraction Applied to Budding Yeast Temporally Periodic Gene Expression, SPRINGER, BioData Mining, Volume 9, Issue 22, 2016.
40. Zhang, L., Qian, L., Ding, C., Zhou, W., and Li, F., Similarity-balanced discriminant neighbor embedding and its application to Cancer classification based on gene expression data. *ELSEVIER, Computers in Biology and Medicine* 64:236–245, 2015.
41. Vanitha, C. D. A., Devaraj, D., and Venkatesulu, M., Gene expression data classification using Support vector machine and mutual information-based gene selection. *ELSEVIER Procedia Computer Science* 47:13–21, 2015.
42. Kaya, M., The effects of a new selection operator on the performance of a genetic algorithm. *ELSEVIER, Applied Mathematics and Computation* 217(19):7669–7678, 2011.
43. Shuai, X., and Zhou, X., A genetic algorithm based on combination operators. *ELSEVIER, Procedia Environmental Sciences* 11, Part A:346–350, 2011.