UDC 575.
https://doi.org/10.2298/GENSR1903805U
Original scientific paper

A NOVEL MICROARRAY GENE SELECTION AND CLASSIFICATION USING INTELLIGENT DYNAMIC GREY WOLF OPTIMIZATION

K.R.UTHAYAN

Department of Information technology, SSN College of Engineering, India

Uthayan. K.R. (2019): *A novel microarray gene selection and classification using intelligent dynamic grey wolf optimization*. - Genetika, Vol 51, No.3, 805-828.

Effective diagnosis of cancer in the medical field is very important to specific treatment. Exact prediction of different cancer types will provide a better treatment and minimization of toxicity in patients. Microarray high dimensionality of gene expression dataand large number of genes against small sample size, noise and repetition in datasets are the main issues which lead to poor classification accuracy. The selection of informative genes and to reduce dimensionality, Gene Selection technique is used in Microarray. In this paper, a novel meta-heurists algorithm based on Grey Wolf Optimization (GWO) and Artificial Intelligence (AI) is combined to design a model for cancer classification. This proposed work consists of two stages. First, a filter method such as Laplacian and Fisher score, are applied to extract the significant subset of features for faster classification and then Intelligent Dynamic Grey Wolf Optimization (IDGWO) is employed to identify the relevant genes. GWO is a swarm-based algorithm selected for gene expression data classification problem, because it makes classification easy about training and testing cancer data. The significant differences between filter methods of datasets are found by using several analyses. The proposed method was applied on five benchmark datasets by considering top 100 ranked genes selected by fisher score in Lymphoma and SRBCT that had a 100% performance using the IDGWO classifier.

Keywords: Fisher Score and Grey Wolf Optimization, Gene Selection, Laplacian, Microarray data.

Corresponding author: Uthayan.K.R, Department of Information Technology,SSN College of Engineering, uthayankr@ssn.edu.

INTRODUCTION

Computer aided cancer disease prediction is crucial to the effective utilization of particular treatments. Over the decade the cancer classification has enhanced in research field and need arises for a totally computerized and less subjective technique for cancer diagnosis. Late investigations have revealed that DNA microarrays could give subjective data for cancer prediction at gene profile expression level as the result of their ability to measure the sample of messenger ribonucleic corrosive (mRNA) transcripts by using large number of subset of genes simultaneously. In this paper, the primary focus is given on gene selection of microarray cancer classification-based algorithms.

As of late, in researching cancer diseases, numerous ways have been opened up by the innovations of microarray utilizing gene expressions (ALSHAMLAN *et al.*, 2014) .The huge number of gene expression levels are quantified in a single chip using Microarray. The microarray comprises of up to 6000 spots and measuring the area of 2cm by 2cm (SALOME *et al*, 2011). The various microarray technologies include Serial Analysis of Gene Expression (SAGE), nylon membrane and illumine bead array (DOSHI *et al*, 2014). Hence microarrays offer a proficient technique for social affair information that can be utilized to decide the articulation example of thousands of qualities.The major test for most classification issues is very high dimensionality of gene profile.

The comparisons made with combinations of various sample sizes and repeated gene features are the primary reasons for the high error rate in classification (BENNET *et al*, 2015). The significant gene is identified to figure out the given relevant cell type with stipulated time under problem specific constraints that are reactive to the specific gene types or tissue samples which relative to cause the cancer disease.

The informative genes are formed to a minimal set of features that are the most prescient to its nearby group by utilizing a classification model (ALSHAMLAN *et al.*, 2014). This will improve the performance of the classifiers to categorize the samples precisely. This improves the gene selection for cancer diagnosis and this would be helpful for early diagnosis and drug discovery. It is substantially less expensive to concentrate on the outflow of just a couple of qualities as opposed to concentrating on a huge number of genes. Likewise, the dimensionality of the features also lessens, and this yields to a reduction in the classification computational cost.

In preprocessing, to get an efficient prediction of cancer, feature selection is an important step for analyzing the data; reduce the dimensionality of the datasets (XIONG *et al.*, 2008). The traditional feature selection models depend on whether the selection process combined with search space techniques or not (ALBA *et al.*, 2007). The feature subset selection and classification are the two phases involved in filter method which is simple and fast to compute. The second method is wrapper strategy which is similar like filter method, combined with a learning algorithm to compute the classification accuracy. From a theoretical perspective, the features in wrapper methods are profitable because the optimization of segregate energy of the long last utilized induction algorithm are done using this method features (CHO *et al.*, 2003).

Wrapper Classification Methods are utilized for upcoming selection of gene expression in cancer prediction (ZHANG *et al.*, 2007) to segregate a kind of tumor, to reduce the number of genes qualities to research if there should arise an occurrence of another patient, and furthermore to aid medicate revelation and early diagnosis. The wrapper model regularly utilizes

developmental methodologies to control their search space. The feature subset contains a set of populations with a valid solution. The fitness value of each population in the subset is evaluated in an iterative fashion to enhance the solutions step by step including all the features subset. Some of the cutting-edge wrapper approaches are Particle Swarm Optimization (PSO) (KAR *et al*., 2015), Ant Colony Optimization (ACO) (YU *et al.,* 2009), Artificial Bee Colony(ABC)algorithm (ALSHAMLAN *et al.*, 2014; BENNET *et al.*, 2015), Distance Sensitive Rival Penalized Competitive Learning with Support Vector Machine (DSRPCL-SVM) (XIONG *et al.*, 2008), Genetic Algorithm with SVM (VUKUSIC *et al.,* 2007) and hereditary programming. Several researches were finished by focusing different aspects, at the same time it has impediments. In this way, there is a requirement for better strategy which for giving appropriate treatment based on cancer cells.

In this work, a novel hybrid metaheuristic method called Intelligent Dynamic Gray Wolf Optimization (IDGWO) in view of Gray Wolf Optimization and some computerized reasoning ideas and strategies is portrayed. Grey Wolf Optimizer (GWO) is a swarm-based algorithm, it is chosen for gene expression data classification problem, since it makes order simple for training and testing cancer data. The proposed strategy essentially comprises a pair of two consecutive steps. As first, a score-based strategy Fisher score (MALINA *et al.,* 1981) and the Laplacian score is utilized to decrease the dimensionality and all the more vitally to give measurably more relevant features subset to subsequent stage. Later, the proposed IDGWO is utilized to locate the relevant genes for cancer classification.

The extraordinary function of Fisher score and its power to noise is pertinent to different applications (XUAN *et al.*, 2007; LIAO *et al.*, 2014). Besides, the high performance of Fisher score for quality determination against other generally utilized techniques, such as Z-score, information gain, andT-test (CHEN *et al.*, 2005). Regardless, every technique has its own attributes that influence the dependability of conclusive outcomes. Further, to identify the predictive genes in cancer datasets, Laplacian discriminant analysis (NIIJIMA *et al.,* 2009) demonstrates its aggressive execution. The Laplacian score is an unsupervised strategy that depends on the hidden structure of a dataset. This trademark spurred to use and research it as a preprocessing venture notwithstanding that is an unsupervised strategy. The proposed IDGWO is benchmarked in blend of both Laplacian and Fisher score. Beforehand, a comparison based on dissimilarity of selected top *M* genes is performed.

MATERIALS AND METHODS

Literature Survey

In the zone of bioinformatics, Classification is the important data mining strategies. This can be found in cancer classification which is as of late tended to by numerous analysts uncommonly in the wake of rising of microarrays. RUIZ *et.al* (2006) tended to the gene selection issue under a classification framework. The point is to generate a classifier that accurately samples the classes (diseases or phenotypes) of new unlabeled samples, utilize them for the classification task. There are two stages in the algorithm, named BIRS (Best Incremental Positioned Subset). It depends on the measurable importance of including a quality of ranked-list to the last subset. For high dimensional datasets, BIRS turns out a fast technique that gives performance in prediction accuracy, at distinguishing important qualities, as well as concerning the computational cost.

HUERTA *et al.* (2006), proposed an approach for the classification of high dimensional Microarray data by combining Genetic Algorithm with Support Vector Machines (SVM). This approach is related to the fuzzy based prefiltering method based preprocessing tool which allows reducing to a great extent the data dimensionality by grouping similar genes. The GA utilizes documents to record high quality solutions. Selecting predictive genes and for final gene selection and classification are utilized by the GA joined with an SVM classifier. The approach is surveyed on two surely understood cancer datasets to consider the behavior of the model. The classification of high dimensional Microarray Data was looked out by the ALBA *et al*. (2007) for the utilization of a Particle Swarm Optimization (PSO) and a Genetic Algorithm (GA) (both enlarged with Support Vector Machines SVM). The utilization of an adapted initialization method has demonstrated an awesome effect on the execution of proposed algorithms, since it presents an early arrangement of acceptable solutions in their development procedure.

SALEM *et al.* (2011) introduced a feature reduction method which joins F-statistics and entropy techniques. Two hybrid cancer classification systems can be brought up by the output of the integrated gene selection technique by encouraging it into two unique classifiers. The two frameworks were a result of incorporating the proposed gene selection technique once with SVM bringing about the First CS and some other times with KNN bringing about the Second CS. To begin with CS and Second CS were connected to two open microarray datasets, leukemia dataset and lymphoma dataset. The First CS was ended up being an intense framework which can be adjusted to any microarray gene expression dataset.

In cancer classification, gene expression profiling based on Microarray is an essential one for diagnosis and prognosis purposes. ALSHAMLAN *et al.* (2014) discussed about the execution of Bio-Inspired evolutionary gene selection method such as (GA, PSO, and ACO) in cancer classification with microarray dataset. In this exploration, it is demonstrated that they are reliable and more suitable wrapper gene selection strategy, since it is equipped for scanning for optimal or near-optimal solutions on complex and large spaces of possible solutions.

BENNET *et al.* (2015) proposed an ensemble features selection strategy which is a blend of Recursive Feature Elimination (RFE) and Based Bayes Blunder Filter (BBF) for gene selection and Support Vector Machine (SVM) algorithm for cancer classification. In Bioinformatics and Data mining, classification of cancer is a promising research area. The genes and BFF are ranked by the SVM-RFE which is connected to remove the repetition on top ranked genes. Also, a few gene selection techniques against various classifiers were analyzed. RFE and BBF takes out the morphological and clinical methods, which is also a key part in accurate cancer classification.

JI-GANG *et al.* (2007) proposed another technique based on BBF, to choose significant genes and repeated features in classification of clinical data. GOLUB *et al*. (1999) proposed two different frameworks for arranging the same microarray dataset were by Support Vector Machine as a classifier, once with Locality Preserving Projection procedure (LPP) and the other with F-score ranking feature selection technique. The two frameworks result in useful and powerful classification of gene expression data.

TANG *et al.* (2007) proposed feature selection by SVM-RFE to evade instability. The initial process of SVM-RFE includes pre-filtering process, eliminate irrelevant, redundant and noisy genes from the informative genes and also it includes multiple iterations to generate gene subsets. In sequence, all gene subsets are joined together and dispense with one quality of each

progression. This subset guarantees linear SVM is utilized for classification. Publicly accessible datasets such as ALL/AML, colon cancer and lymphoma are used for execution of cancer classification. In this method the gene subset selection strategy is utilized with ranking method in terms of accuracy and area under ROC to achieve efficient results.

Problem Statement

Due to certain limitations in the microarray dataset such as curse of dimensionality, the modest numbers of irrelevant genes samples create unwanted noise which leads the classification as a challenging task for given sample FU, 2004. In classification and clustering phase, computational complexity and the dimensionality of the gene expression matrix is increased by the irrelevant genes GAN, 2008; GARRO, 2016, GHORAI, 2010; HANCER, 2015. As an outcome, it is very important to include those least correlated genes with the more informative genes, which is a feature selection problem in microarray data analysis. So, a small subset of genes can be identified by the implementation of the microarray data classification for the prediction of cancer.

The main contributions of this research paper are summarized as follows:

a) The number of the selected informative genes, which are selected from the top relevant genes are reduced by the Filtering method.

b) A novel predictive framework based on Intelligent Dynamic Grey Wolf Optimization (IDGWO) is presented to search the highly relevant genes.

c) The designed framework IDGWO is the best technique when it is applied to cancer application and to achieve efficient classification performance.

Feature reduction for the microarray cancer disease classification-based optimization algorithms, consist of two steps: Initially the filtering methods were employed to reduce the number of genes and to select the top statistically relevant genes. Secondly IDGWO was employed to identify the highly relevant genes. The cancer datasets were loaded in which training and testing samples was separated from which top M genes are statically accessed by filter methods. Then, the IDGWO framework is applied to search the relevant genes for cancer classification. The detailed description of initial gene ranking, gene selection, Genetic Algorithm (GA) and IDGWO are explained below.

Gene Ranking

The selection of the relevant subsets is promoted by the ranking methods without losing the informative gene while reducing the search space for genetic algorithm. Being difficult to apply optimization method directly to high dimensional datasets DEEGALLA, 2007; DESSÌ, 2009; DESSÌ, 2013, reduced feature subset provides the possibility of putting into optimization usually effective against small or middle scale datasets, for microarray data classification.

The GWO prefers Gene Ranking to reduce the dimensionality of highly irrelevant genes and to identify the relevant genes for cancer prediction. By using the M rank method, the genes are separately ranked and M ranked set of genes contains the relevant genes in descendant order. The*T* top-ranked genes from each set are selected to reduce the dimensionality, where *T* denoted as fixed threshold.

Gene Selection

Hundreds and thousands of genes (features) are in Cancer Microarray. Over fitting (For an instance it is easy identify a decision function which classifies the accurate training data but inappropriate test data) may be caused by the classification of relevant genes from a high dimensional space, with increased execution time. Gene Selection techniques should be introduced to find the highly informative subset and relevant genes by searching through the space of features SANDEN, 2008; STATNIKOV, 2008; WANG, 2011; WANG, 2005.

The ranking method consists of two approaches are filter and wrapper for YANG**,** 2006 cancer classification. The filter method is an efficient ranking method is employed to reduce the number of genes from huge number of gene set. In a wrapper method, a classifier is used as the evaluation criterion. By using the feature ranking methods, each gene is identified based on specific criterion and selecting a subset of genes above or below a specific threshold and eradicate the irrelevant genes according to gene characteristics.

To reduce the time complexity of the data mining applications, feature subset selection is the important task in pre-processing OLYAEE, 2013. In microarray data classification, redundant genes and irrelevant features leads to high dimensionality problem, these limitations are eliminated using the filtering techniques. The efficiency of microarray data analysis is improved by reducing the number of irrelevant genes. In this framework, Fisher-Score XUAN, 2007 and Laplacian MOLER, 2000 scoring methods were selected based on their performance.

Fisher score uses the supervised approach for ranking the genes, i.e., to identify the informative gene, it needs the class labels. Furthermore, the high performance of Fisher score for gene selection is highly robustness to noise such as T-test CHEN *et al.* (2005), information gain and Z-score. Laplacian discriminant analysis showed its competitive performance for identifying predictive genes in cancer datasets. The Laplacian score is an unsupervised method that relies on the underlying structure of the dataset. Moreover, a comparison based on the dissimilarity of selected top M genes using these two scoring methods was performed. Both scoring methods are separately used in the initial feature selection method LEE and LEU (2011).

Genetic Algorithm

Genetic Algorithm (GA) belongs to the class of evolutionary algorithm mainly it is employed to find the optimal solution to a set of solutions KURKURE *et al.* (2015). In this paper, the relevant genes for cancer classification are found out by GA. The solution gets improved on the generations based on the mechanisms of natural selection according to the principle of "survival of the fittest" introduced by Charles Darwin. The basic stages in GA such as encoding, selection, cross over and mutation operators. Encoding is a process of representing individual genes by using binary string with 1s and 0s. An initial population is created consisting of randomly generated rules. Each rule is represented as a string of bits. The string consists of binary bits are 1 to represent selection of feature else 0 to drop that feature.

Cross Population means selecting two parents in selection. Crossover is the process of generating a child from two parent taken from the population. Mutation is adopted to avert the algorithm from trapping in a local minimum population DASHTBAN *et al* (2017). For binary representation, a simple mutation inverts the value of each gene with a small probability. The population is evaluated and tested for each generation for termination criteria. If the termination

criterion is not satisfied, then population is repeated again by GA operators and then reevaluated. The steps for GA for cancer classification are given below.

Input: Chromosome with n features (cancer dataset)
Output: Single image with cancer detection
Step 1 - Initialize the population randomly with chromosomes.
Step 2 - Initialize N (number of samples in the training set).
Step 3 - Apply the Crossover to the selected bestchromosome.
Step 4 - Apply Mutation for each chromosome to generate new population.
Step 5 -Evaluate the fitness.
Step 6 - Select best fit chromosome (relevant cancer genes).
Step 7 - If the condition not satisfied repeat the step 3.

GA are sensitive to initial parameters such as mutation can significantly influence the search space, take more time for convergence, computationally expensive and not guaranteed to find an optimal solution. Due to worse fitness functions certain optimization problems cannot be resolved by genetic algorithm, which makes the bad chromosome blocks and good chromosomes crossover. To overcome these flaws optimization algorithm is employed to identify the relevant genes for cancer classification. Grey Wolf optimizer (GWO) is a swarm-based algorithm, it is selected for gene expression data classification problem, because it makes classification easy about training and testing cancer data.

Intelligent Dynamic Grey Wolf Optimization

Artificial Intelligence (AI) concepts and techniques are employed to find the minimum number of relevant genes for cancer classification (samples) from a large number of genes by meta-heurists algorithm. The selected benchmark dataset was loaded for the evaluation and missing data are imputed using the KNN imputation. Then the training data was split into two sub groups for the learning process, namely training and test samples. Figure 1 shows the process of IDGWO cancer classification. The training data are used only for constructing a classifier and evaluating individuals during the evolutionary process, while the test sub-data are used to assess the final solutions. Immediately after partitioning, the dataset was minimized concerning top selected genes by statistical scoring. The meta-heurists algorithm is employed on feature selection subsets.

Grey Wolf Optimization a current meta-heuristic algorithm proposed by MIRJALILI *et.al* based on the social hierarchy and hunting mechanism inspired by grey wolves. The power dispatch problem which is imposing the valve point effect and generator constraints can be solved by the Grey Wolf Optimization.

Grey Wolf Optimization is generally a mathematical approach whose arrangement convergence is controlled by the leadership order and chasing mechanism of Grey wolves. It demonstrates that each search space as a multilevel decision system and it doesn't require slope for search path. The grey wolf system comprises of four types, which includes alpha, beta, delta and omega are utilized for reenacting the leadership hierarchy. Alongside the three fundamental advances includes in GWO are chasing, searching prey, encircling prey and attacking prey. Grey

wolves are thought to be peak predators which imply that they are at the highest point of the food chain order and preferred to live in a pack. The group size is 5-12 on average.

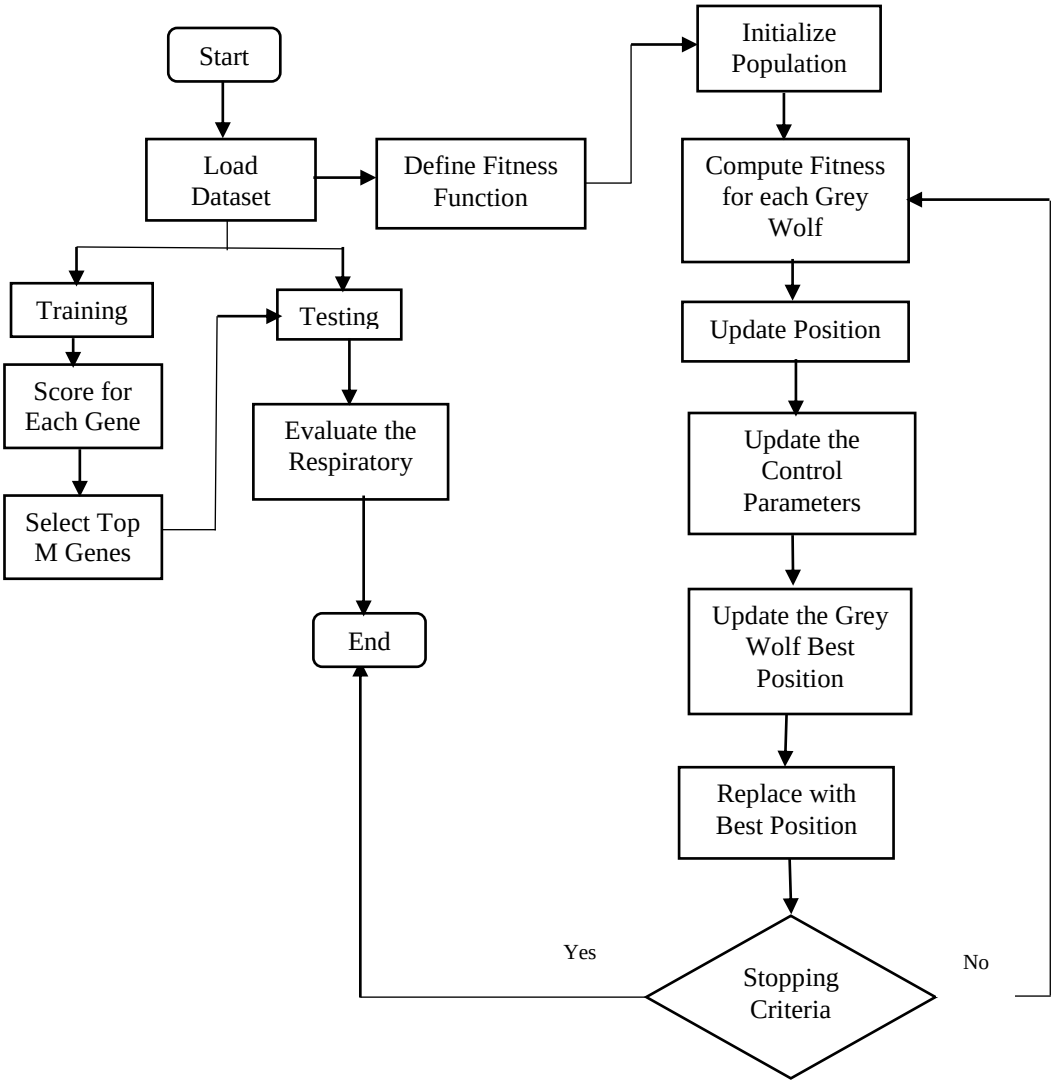


Figure 1. Flowchart for IDGWO

The alpha is the pioneer (leader) includes male and a female and mainly it performs in choosing from the hunting, resting spot, and time to wake. Likewise, been seen, in which an alpha takes after other wolves and its behavior and the alpha wolves should be a friend of the pack. Strangely, the alpha isn't really the strongest member of the pack however the best

regarding dealing with the pack. This demonstrates the association and discipline of a pack is considerably more vital than its strength.

The beta is the second level in the grey wolves' hierarchy which is subordinate wolves that assistance for basic leadership or other packs activities of alpha. The beta wolf is most likely the best possibility to alpha for some situation if any of the alpha wolves passes away or turns out to be exceptionally old. The beta wolf takes after the alpha and requests the other lower level wolves and appeared in Figure 2. Beta assumes the part of a counsel to the alpha and furthermore a trained for the pack. The beta supports the alpha's orders for the pack and it gives some comments to the alpha.

The lowest position wolf is omega and permitted to eat at last. Mostly it dependably needs to present the procedure to all other dominant wolves. It might appear that omega isn't a significant individual in the pack, yet it has been watched that the entire pack confront internal battling and issues if there should arise an occurrence of losing the omega. This is because of the brutality and dissatisfaction with all wolves by the omega(s). This helps fulfilling the whole pack and it keeps up the dominance structure. In the event that a wolf isn't an alpha, beta, or omega is called subordinate. Delta wolves ought to submit information about alphas and betas, however delta can command the omega. Elders are the accomplished wolves that used to be either alpha or beta. Hunters should help the alpha and beta for hunting and furthermore give nourishment to the pack. The caretakers are responsible for charge of nurturing the weak, sick and injured wolves in the pack.

In addition to the social hierarchy of wolves, the significant social behavior of grey wolves is group hunting. The fundamental stages of grey wolf hunting are as follows,

- Tracking, chasing, and searching the prey.
- Pursuing, encircling and harassing the prey until it stops moving.
- Attack of the prey.

Mathematical Model and Algorithm

The mathematical model of grey wolf hierarchy is tracking; encircling and attacking prey are given below.

A. Social Hierarchy

For outlining GWO, the mathematically model is utilized by the social order for wolves it trusts in the fittest arrangement, for example, the alpha (α), second and third are named as beta (β) and delta (δ). The remaining wolves are assigned as omega (Ω). In the GWO calculation the chasing (enhancement) are guided by α, β and δ. The Ω wolves take after the staying three wolves.

B.Encircling Prey

During hunting, the encircling behavior grey wolves' prey are designed in order to mathematically model encircling behavior the following equations are considered as

$$D = \left(C.Xp\left(t\right) - X(t)\right)$$ -----Eq(1)

$$X(t+1) = Xp(t) - A \quad \text{-----Eq(2)}$$

Where t is the current iteration, *A* and *C* are coefficient vectors X_p (t) represents the position vector of the victim. The vectors *A* and *C* can be calculated as below.

$$\vec{A} = 2\vec{a}.r1 - \vec{a} \text{ ----------Eq(3)}$$

$$\vec{C} = 2.r2$$

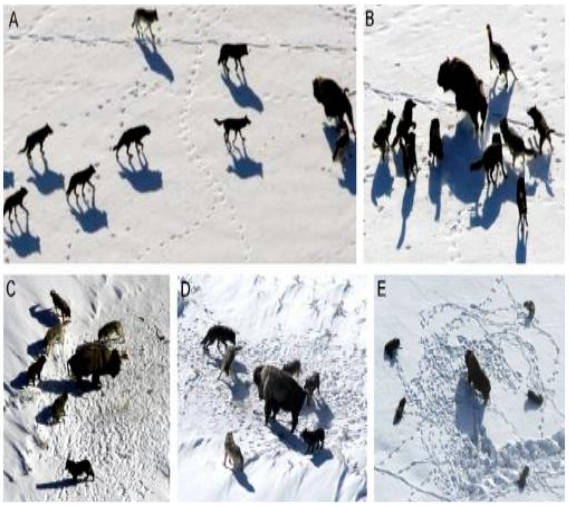


Figure 2. Hunting Behavior of Grey Wolves

where *a* includes that it linearly decreased from 2 to 0 over the course of iterations and r1 and r2 are random vectors are lies in the range [0, 1].

A two-dimensional position vector and the possible neighbors are illustrated with Eq (3).In the Figure 2, a grey wolf in the position of (X, Y) can update its position according to the prey position (X*, Y*).

Different spaces around the best place can be reached with respect to the current position by adjusting the value of A and C For instance, (X*-X, Y*) can be reached by setting A = (1, 0) and C = (1, 1). The possible updated positions of a grey wolf in 3D space are depicted in Figure 2. The random vectors r1 and r2 permits wolves to reach the prey location and update its position in any random location of search space by using equations.

C.Hunting

Grey wolves will make out the prey area and enclose them. Typically, the chase is control by the alpha. The beta and delta additionally take part in chasing process. Be that as it may, in a pursuit space there is no thought regarding the area of the optimum (prey)

(2017). In order to mathematically simulate the hunting behavior of grey wolves and the alpha (best candidate arrangement) beta, and delta have better learning about the potential area of prey is appeared in Figure 3. Along these lines, by saving the initial three best arrangements acquired up until now and the other search agents (including omegas) are required to update their locations as indicated by the position of the best search. The accompanying models are displayed and appeared in Eq 4.

$$\overrightarrow{D_\alpha} = |C1X\alpha - X|$$
$$\overrightarrow{D_\beta} = |C2.X\beta - X| \text{-------Eq(4)}$$
$$\overrightarrow{D_\delta} = |C2.X\delta - X|$$

In encircle scheme the final position in a random search space are represented by the positions of alpha, beta, and delta in the search space and other wolves are update their positions randomly around the space.

$$\overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{A_1}.(\overrightarrow{D_\alpha})$$
$$\overrightarrow{X_2} = \overrightarrow{X_\beta} - \overrightarrow{A_2}.(\overrightarrow{D_\beta}) \text{----------Eq(5)}$$
$$\overrightarrow{X_3} = \overrightarrow{X_\delta} - \overrightarrow{A_3}.(\overrightarrow{D_\delta})$$
$$\vec{X}(t+1) = \frac{X1+X2+X3}{3} \text{ ------Eq(6)}$$

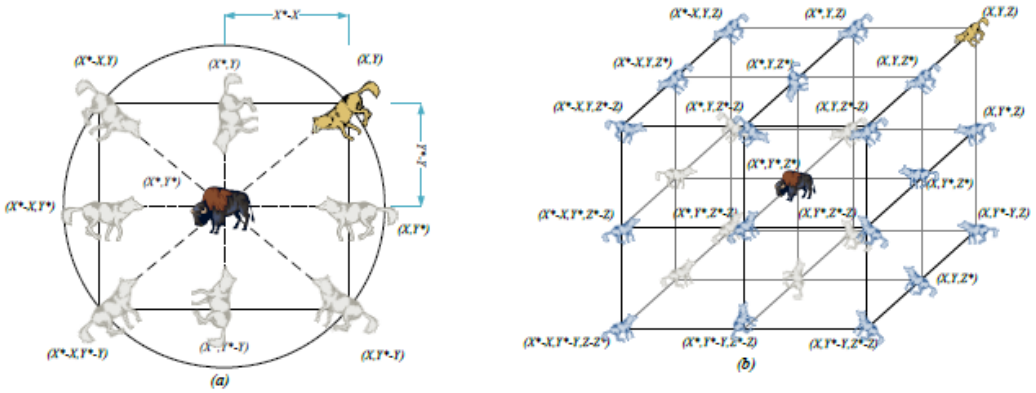


Figure 3.2D and 3D Position Vectors with Possible Next Locations

D. Attacking Prey

As appeared in the Figure 3 the prey is hunted by the grey wolves when it stops. The mathematical model describes about searching for prey by diminishing the value of a. Note that the variance scope of A is decreased by diminishing the estimation of a. As such, A is a random value in the interim [-a, a] where a is diminished from 2 to 0 through the course of iterations. At

the point random values of are in [-1, 1] the following position of a search agent can be in any position between its present position and the position of the prey Figure 5 demonstrates that |A|<1 forces the wolves to attack towards the prey. With the operators proposed up until this point, to update the alpha, beta and delta location; the GWO allow its search agents according to the position and attack towards the prey. In any case, the GWO algorithm is level to stagnation in local solutions with these operators. The reality of the matter is that the encircling technique demonstrates investigation to some degree, however GWO needs more operators to underscore exploration.

E. Search for Prey

The position of the alpha, beta, and delta helps the Grey Wolves to searches its prey shown in Figure 6. The alpha, beta and delta will diverge for seeking the prey and they will merge to attack prey EL BAKRAWY (2017). In order to generate the mathematical model dissimilarity by using the random values A more prominent than 1 or not exactly - 1 to oblige the search operator to diverge from the prey. This accentuates exploration and permits the GWO to search globally demonstrates that |A| > 1 powers the Grey wolves to separate from the prey to ideally locate afitter prey.

The C vector is the impact of approaching to moving towards prey in nature. The Vector C does the obstruction in nature which is showed up in the hunting ways of wolves and it keeps them from rapidly and conveniently moving towards prey. The C vector can be likewise considered as the impact of obstructions to moving toward prey in nature. Based on wolf location LI *et al.(*2017), it can arbitrarily give the prey a weight and make it harder and more distant to go after wolves, or vice versa.

The GWO favors the exploration of C vector which includes the random values between 0 and 2 which provides random weights for prey in order to emphasize ($C > 1$) or deemphasize ($C < 1$) the prey distance defined in Eq 3. This supports GWO to show a more random behavior throughout optimization, favoring exploration and local optima avoidance. The *C* is not linearly decreased with respect to *A*. During initial iterations is it required toprovidearandomvaluein order to emphasize exploration which is useful in case of local optima stagnation, especially in the final iterations. The overall flowchart for the grey wolf optimization is shown in Figure 4**.**

To perceive how GWO is theoretically able to tackle optimization issues, a few focuses might be noted as.

- The hunting technique enables the candidate solutions for find the right position of the prey.
- The adaptive values ensure the Exploration and exploitation for example, a and A.
- The adaptive values of parameters A permit the GWO to progress easily amongst exploration and exploitation.
- With diminishing A, the one portion of the iterations are committed to exploration ($|A| \geq 1$) and another half are devoted to exploitation ($|A| < 1$).

The Two primary parameters of GWO, A and C should be balance.

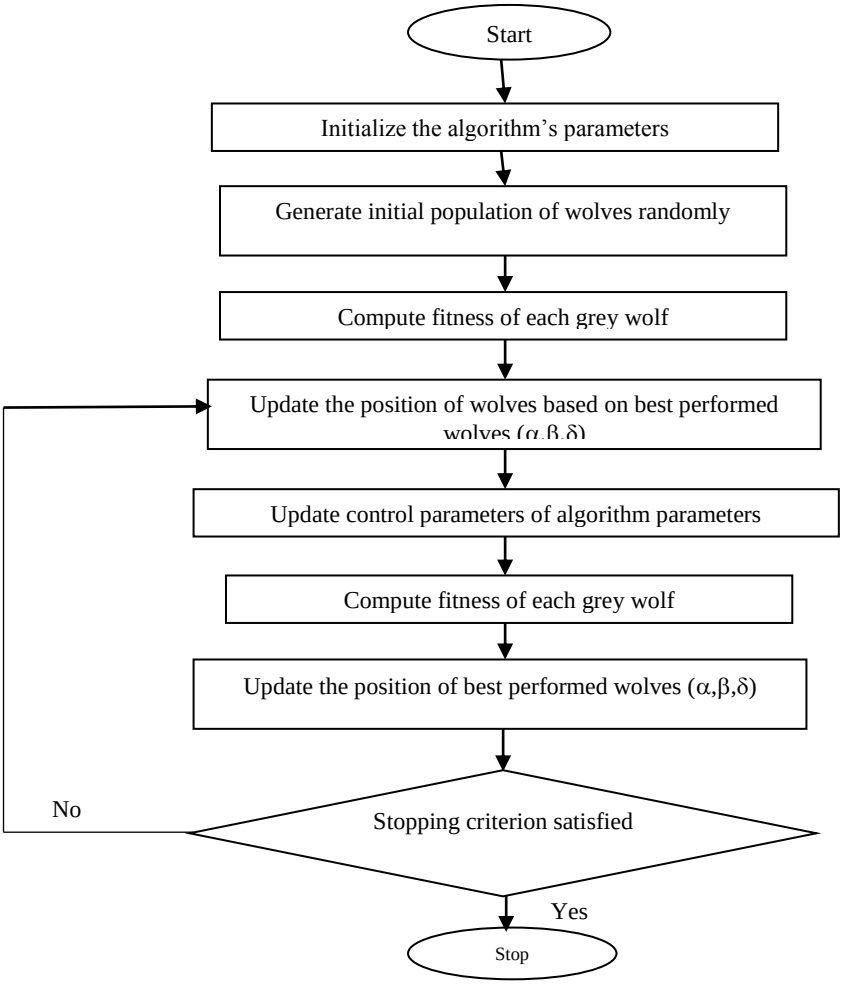
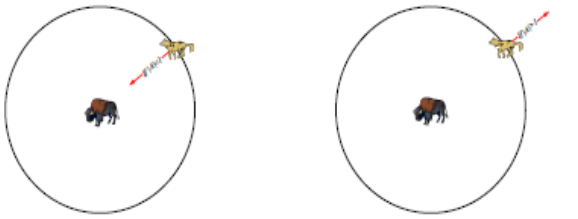


Figure 4. Flow chart for GWO



i

Figure 5. Attacking Prey with Searching Prey

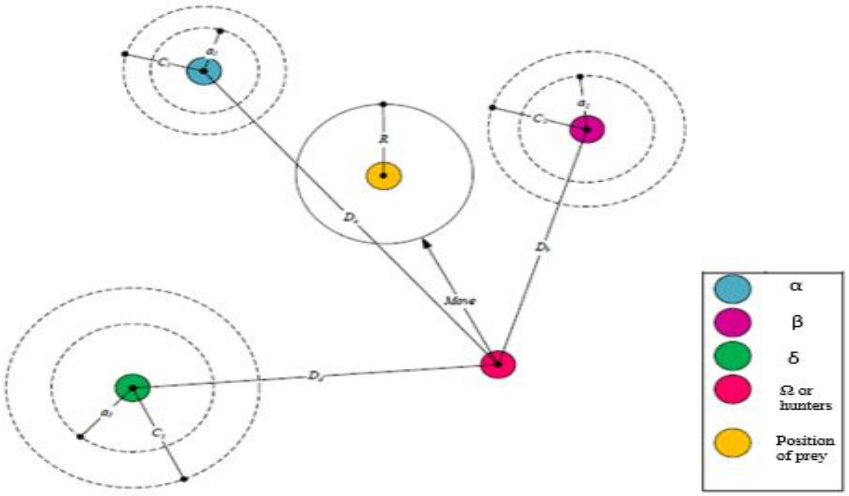


Figure 6. Position Updating in GWO

Furthermore, once the meta-heuristics process was completed, the relevant subset genes are identified from the huge number of collected genes and used for further process. Cross Validation is a factual strategy utilized in evaluating and comparing the learning algorithms during the training process of the classifier where its tasks is to partition the trained dataset into two portions; one is utilized for training and the other is utilized for validation. One thought is the nature of collected genes ought to be measured over the testing tests or utilizing Leave-One-Out-Cross-Validation (LOOCV) where it utilizes one sample for test and all samples for training SALEM et al.(2011).

RESULTS AND DISSCUSSION

The performance evaluation of proposed Intelligent Dynamic Grey Wolf Optimization is simulated by MATLAB tool under windows environment. After initial feature selection using either Fisher-score or Laplacian-score, top genes were selected. Then, a new dataset was obtained using the filtered genes. First, the original split in test/train data was loaded. Then, the proposed method was performed independently on each dataset using either of the filter methods separately.

Dataset Description

In this IDGWO framework, five publicly available high-dimensional micro-array datasets are used, such as Lymphoma SALEM *ET AL*. (2011), Leukemia HUERTA *et al.* (2006), SRBCT WANG *et al*.(2007), MLL HUERTA *et al.* (2006) and Prostate HUERTA *et al.* (2006) cancer datasets to evaluate the proposed method for gene selection and cancer classification.

Table 1. Statistics of microarray caner datasets

Datasets	Number of Class	Number of Genes	Number of Samples
Lymphoma	3	4026	62
Leukemia	2	7129	72
SRBCT	4	2308	83
MLL	3	12582	57
Prostate	2	5966	102

Filter Methods

Table 2.Comparison of Filter Methods

Datasets	Top genes	Filter Methods							
		Fisher Score				Laplacian Method			
		K-NN	SVM	GA	IDGWO	K-NN	SVM	GA	IDGWO
Lymphoma	5	48.2	56.8	48.6	56.5	39.8	85.6	68.2	68.5
	10	56.3	63.5	59.6	57.8	45.8	89.5	81.8	69.8
	50	64.8	72.9	72.5	88.6	52.6	94.5	63.6	72.8
	100	72.5	85.6	88.6	100	65.8	100	68.2	75.6
	200	85.6	88.6	100	100	72.6	100	81.8	78.6
Leukemia	5	45.6	36.5	54.3	65.8	65.8	59.1	68.2	77.3
	10	59.6	42.6	56.3	75.6	78.6	68.2	68.2	77.3
	50	69.8	59.8	65.9	82.6	88.5	77.3	71.8	95.5
	100	86.9	63.5	71.6	89.6	100	77.3	73.6	95.5
	200	88.6	78.5	78.6	95.6	100	88.2	78.2	95.5
SRBCT	5	49.6	75.6	75.8	78.5	55.6	51.1	27.8	90.5
	10	52.8	83.6	79.6	81.6	57.8	51.1	70.8	92.5
	50	72.3	95.6	81.6	85.6	59.6	51.1	84.7	95.6
	100	88.6	100	85.3	95.6	62.8	63.8	91.7	100
	200	92.6	100	89.5	100	65.8	74.5	93.6	100
MLL	5	72.6	65.8	75.6	87.6	67.8	78.2	91.6	89.0
	10	85.6	72.6	89.6	89.6	69.8	68.1	94.6	89.5
	50	92.6	82.6	92.5	91.5	70.5	83.3	98.5	89.7
	100	95.6	89.6	95.6	92.8	71.6	81.9	100	86.0
	200	100	95.3	98.6	96.8	75.8	84.7	100	91.2
Prostate	5	82.6	81.5	79.6	78.6	82.3	89.0	55.1	75.6
	10	86.9	85.6	82.6	82.6	89.0	91.2	70.6	76.8
	50	89.6	92.5	85.9	89.5	91.2	89.7	71.3	78.6
	100	95.8	97.6	92.6	91.6	89.7	86.0	83.1	82.6
	200	98.9	99.5	95.6	98.6	86.0	96.2	81.6	85.9

The quality of top selected genes by filter methods such as Fisher score and Laplacian for each dataset are evaluated and shown in table 2. The top genes are selected in the order of 5 subsets such as 5, 10, 50,100 and 200. The proposed method is compared with K-Nearest Neighbor (K-NN) SALEM *et al*. (2011), Support Vector Machine (SVM) [6], Genetic Algorithm (GA) DASHTBAN *et al*. (2017). The IDGWO was performed independently for each dataset. The obtained results after each run were evaluated using Leave-One-Out-Cross-Validation (LOOCV) which is the most promising evaluation criteria to identify the best genes.

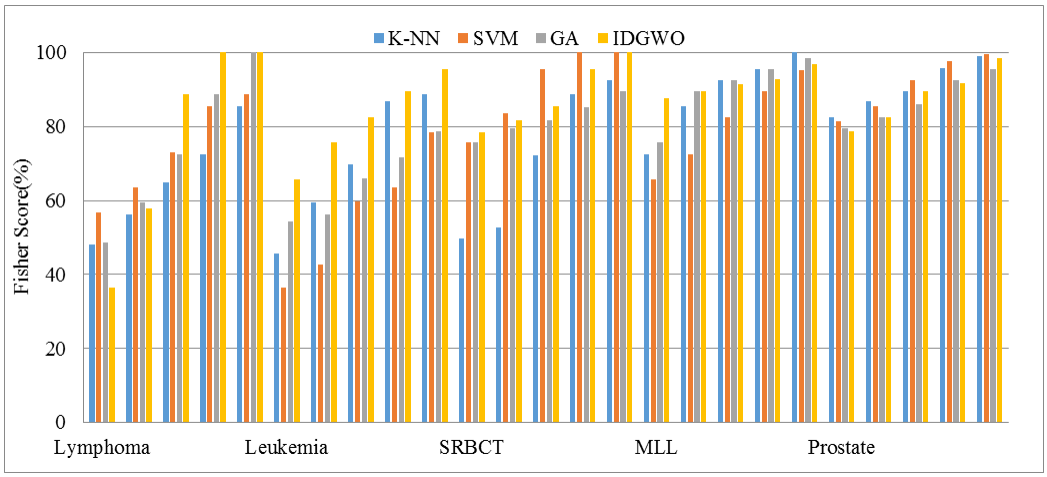


Figure 7. Comparison of Fisher Score for Various Classifiers

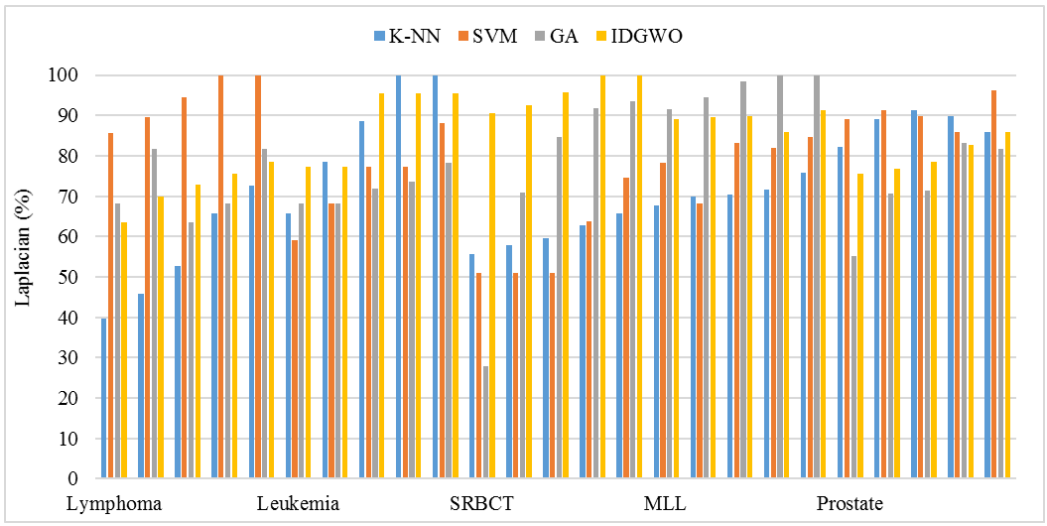


Figure 8. Comparison of Fisher Score for Various Classifiers

Figure 7 and 8 shows that the quality of top selected genes by the filter methods over five subsets of top genes for each dataset which were cross validated by three classifiers using LOOCV validation. Hence, top genes were selected after some trial and error in a way that could be more comparable with the results of other methods at first glance, the best subset among five subsets, considering all datasets, is the top 100 genes selected by Fisher score in Lymphoma and SRBCT that had a 100% performance using the IDGWO classifier. It seems that the best performance of the filter methods was evaluated by SRBCT and MLL dataset.

Table 3. P-Value For datasets

Methods	Datasets				
	Lymphoma	Leukemia	SRBCT	MLL	Prostate
F-test	0.003	0.001	0.03	0.07	0.09
T-test	0.68	0.24	0.23	0.08	0.17

Table 3 denotes the p-values to find the differences in means and variances of IDGWO with various datasets. F-test was used to determine whether the results of IDGWO and other methods of unequal variance or not. The resulting p-values show that there was a relatively significant difference between the variances between two groups of every dataset.

Table 4. Comparison of significant Difference for Ranking Methods

Methods	Datasets				
	Lymphoma	Leukemia	SRBCT	MLL	Prostate
F-Statistics	1.0	14.8	1.1	z	2.5
p-value	0.49	~0	0.42	~0	0.03
Variance pair	(137,139)	(325,22)	(348,386)	(954,43)	(83,207)

Table 4 shows that significant difference between the variance in the performance of two ranking methods of Leukemia, MLL and Prostate. However, it was not convincing enough to reject the null hypothesis in the case of SRBCT and lymphoma. Moreover, the testing difference between ranking methods of significance level of 0.01 leads to alternative assumption of the case of Prostate.

Table 5.Comparison of tests for scoring methods

Methods	Datasets				
	Lymphoma	Leukemia	SRBCT	MLL	Prostate
T-statistics	−1.9	−3.3	−1.3	−4.6	−3.4
p-value	2.03	2.09	2.03	2.09	2.05

Table 5 shows the Comparison of the t-statistics with two-tail critical values revealed a statistically significant difference between the quality of the selected genes by Fisher and

Laplacian. Afterwards, the t-test was performed for comparing the means of the observed performance over each dataset, with the identified assumption (whether the numbers had equal or unequal variance took from f-test). This test was also performed under significance level of 0.05. This test was performed over all observed performances of the proposed method to see where the IDGWO per-formed significantly better when a particular filter method had been used.

Table 6.Comparison of Significance Test of IDGWO

Methods	Datasets				
	Lymphoma	Leukemia	SRBCT	MLL	Prostate
F-test	sig	no	no	sig	no
T-test	no	no	sig	no	sig

Where ‘sig’ denotes significant difference and no denotes no difference. Table 6 shows the results of significance test for comparing the means and variances of IDGWO's results between Fisher and Laplacian for each dataset using t-test and f-test, respectively. It is worth reminding that the performance of two filtering methods to select top genes was overtly different to Prostate with only one common gene. Overall, these results show that the choice of filter method can significantly affect evolutionary search results. Nonetheless, it does not mean that a filter method performs better in all datasets even so it could be said that it performs relatively better in some datasets.

Table 7.Accuracy comparison for Proposed Methods

Datasets	Filter Methods							
	Fisher Score				Laplacian Method			
	K-NN	SVM	GA	IDGWO	K-NN	SVM	GA	IDGWO
Lymphoma	44	57	68	72	47	58	62	68
Leukemia	57	67	75	84	55	64	72	75
SRBCT	54	62	79	89	59	66	78	83
MLL	68	75	86	91	68	78	82	89
Prostate	78	81	86	95	75	83	85	93

Table 7 shows the accuracy comparison of various classifiers such as K-NN, SVM, GA and proposed IDGWO with respect to fisher score and Laplacian method among various cancer datasets.

Figure 9 and 10 shows the accuracy comparison of proposed methods for two scoring methods such as fisher score and Laplacian method. The proposed IDGWO achieves high classification accuracy of 95% and 93% for both scoring methods when compared to other

standard techniques. The obtained result used to identify the relevant genes from large number of genes for cancer classification.

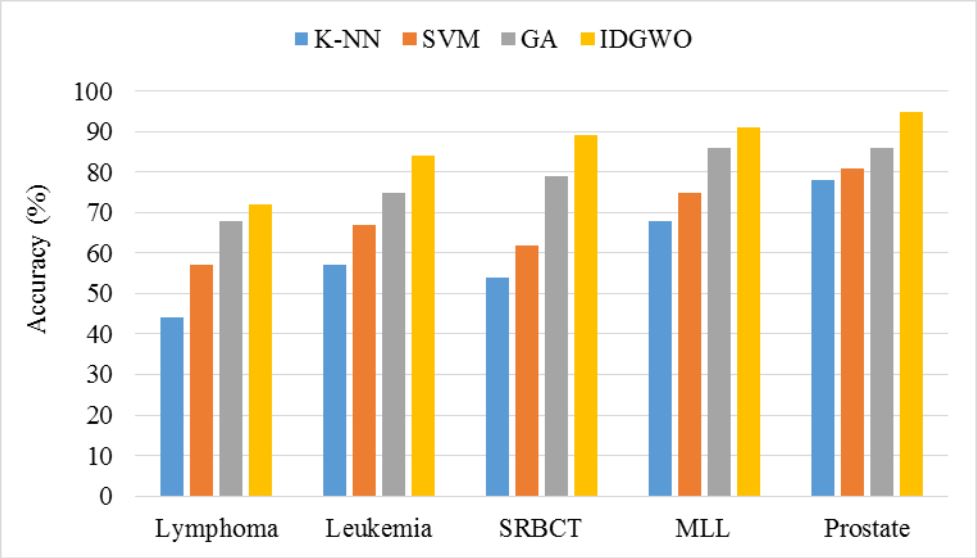


Figure 9. Accuracy Comparison of Fisher Score

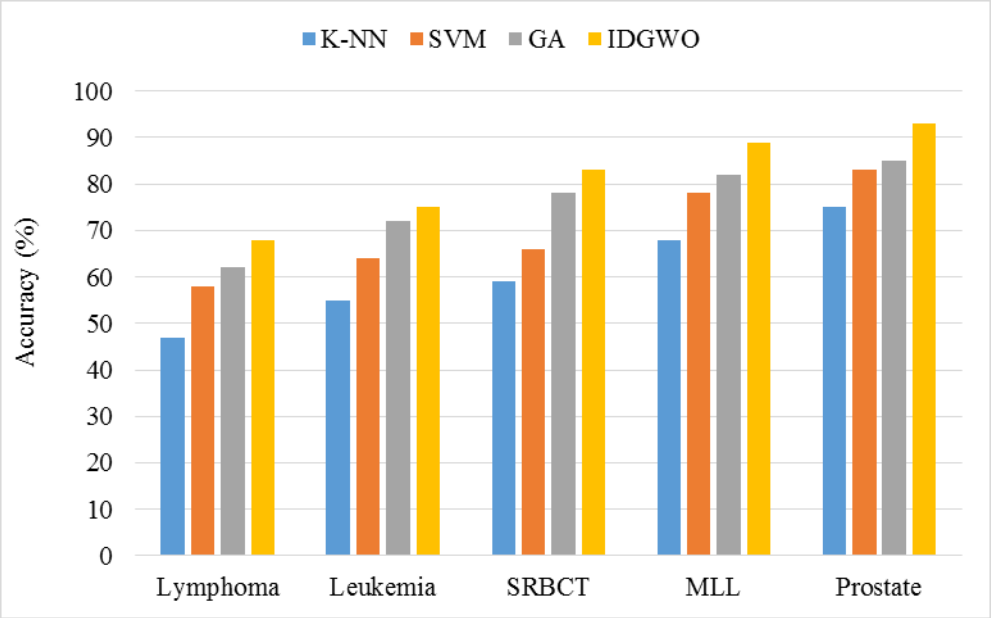


Figure 10. Accuracy Comparison of Fisher Score

CONCLUSION

The cancer diagnosis is the challenging task in the medical field. By calculating the accurate tumor kinds, treatment and toxicity can be offered on the patients in a greater value. The selection of relevant gene for identifying the cancer is evaluated by optimization techniques. Gene selection of microarray dataset suffers from the high dimensionality and noise genes from the irrelevant genes. It is achieved by employing a feature selection methods and meta-heurists algorithm for cancer classification.

A novel algorithm, called Intelligent Dynamic Grey Wolf Optimization (IDGWO), based on the concepts of grey wolf Optimization and artificial intelligence, for gene selection and cancer classification in microarray data, was proposed. Firstly, a ranking method was used to reduce the dimensionality and to provide significant genes and forward to the metaheuristic algorithm. Next, Intelligent Dynamic Grey Wolf Optimization (IDGWO) method was used along with two different ranking methods, the Fisher-score and the Laplacian score was proposed to find the relevant genes. Grey wolf optimizer method will be used for selecting the appropriate features, the genes are divided into feature and it is divided into subsets.

The proposed algorithm analyzes this occurrence and provides a way to investigate important genes for classification. It was shown that the choice of filter method could significantly affect the obtained results from some datasets. Both scoring methods were compared with the similarity and dissimilarity upon different datasets. The proposed IDGWO achieves high classification accuracy of 100% for both scoring methods when compared to other standard techniques. It shows efficient results when compared to the other standard techniques and achieves high classification accuracy.

Received, July 21th, 2018
Accepted May 18th, 2019

REFERENCES

ALBA, E., J.GARCIA-NIETO, L.JOURDAN, AND E.G. TALBI (2007). Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In Evolutionary Computation, 2007. CEC 2007. IEEE Congress on (pp. 284-290). IEEE.

ALSHAMLAN, H.M., G.H. BADR, Y.A. ALOHALI (2014). The performance of bio-inspired evolutionary gene selection methods for cancer classification using microarray dataset. International Journal of Bioscience, Biochemistry and Bioinformatics, 4(3), p.166.

BENNET, J., C. VAN GANAPRAKASAM, N. KUMAR (2015). A Hybrid Approach for Gene Selection and Classification using Support Vector Machine. International Arab Journal of Information Technology, Vol. 12, No. 6A, 2015.

CHEN, D., Z. LIU, X. MA, D. HUA (2005). Selecting genes by test statistics. BioMed Research International, 2005(2), pp.132-138.

CHO, S.B., H.H., WON (2003). Machine learning in DNA microarray analysis for cancer classification. In Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19 (pp. 189-198). Australian Computer Society, Inc.,

CHU, F., L. WANG (2005). Applications of support vector machines to cancer classification with microarray data. International journal of neural systems, *15*(06), pp.475-484.

DASHTBAN, M., M., BALAFAR (2017). Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics, *109*(2), pp.91-107.

DEEGALLA, S., H.BOSTRÖM (2007). Classification of microarrays with knn: Comparison of dimensionality reduction methods. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 800-809). Springer, Berlin, Heidelberg.

DESSÌ, N., B. PES (2009). An evolutionary method for combining different feature selection criteria in microarray data classification. Journal of Artificial Evolution and applications, 2009, p.3.

DESSÌ, N., E.PASCARIELLO, B. PES (2013). A comparative analysis of biomarker selection techniques. BioMed research international, 2013.

DOSHI, J., M.CHINDHE, , Y.KHARCHE, , S.GHOSH, VALADI (2014). Simultaneous Gene Selection and Cancer Classification using Chemical Reaction Optimizatio J.n. In Proceedings of the World Congress on Engineering (Vol. 1).

EL BAKRAWY L.M. (2017). "Grey Wolf Optimization and Naive Bayes Classifier Incorporation for Heart Disease Diagnosis", Australian Journal of Basic and Applied Sciences, *11*(7), Pp.64-70.

FU, L.M. AND C.S. FU-LIU (2004). Multi-class cancer subtype classification based on gene expression signatures with reliability analysis. FEBS letters, *561*(1-3), pp.186-190.

GAN, Z., T.W. CHOW, H D.UANG (2008). Effective gene selection method using Bayesian discriminant-based criterion and genetic algorithms. Journal of Signal Processing Systems, *50*(3), pp.293-304

GARRO, B.A., K.RODRÍGUEZ, R.A. VÁZQUEZ (2016). Classification of DNA microarrays using artificial neural networks and ABC algorithm. Applied Soft Computing, *38*, pp.548-560

GHORAI, S., M A.UKHERJEE, S.SENGUPTA, P.K. DUTTA (2010), December. Multicategory cancer classification from gene expression data by multiclass NPPC ensemble. In Systems in Medicine and Biology (ICSMB), 2010 International Conference on(pp. 41-48). IEEE.

GOLUB, T.R., D.K. SLONIM, P. TAMAYO, C. HUARD, M. GAASENBEEK, J.P. MESIROV, H. COLLER, M.L. LOH, J.R., DOWNING, M.A. CALIGIURI, C.D., BLOOMFIELD (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, *286*(5439), pp.531-537.

HANCER, E., B. XUE, D.KARABOGA, M. ZHANG (2015). A binary ABC algorithm based on advanced similarity scheme for feature selection. Applied Soft Computing, *36*, pp.334-348

HUERTA, E.B., B.DUVAL, J.K.HAO (2006). A hybrid GA/SVM approach for gene selection and classification of microarray data. In Workshops on Applications of Evolutionary Computation (pp. 34-44). Springer, Berlin, Heidelberg.

KAR, S., K.D. SHARMA, M. MAITRA (2015). Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Systems with Applications, *42*(1), pp.612-627.

KURKURE M, A THAKARE, S. GUDADHE (2015). Genetic Candidate Group Search Approach for Post Clustering Content based Image Retrieval. International Journal of Computer Applications. 1;*132*(16).

LEE, C.P., Y.LEU (2011). A novel hybrid feature selection method for microarray data analysis. Applied Soft Computing, 11(1), pp.208-213.

LI Q., H.CHEN, H.HUANG, X ZHAO., Z CAI., C.TONG, W.LIU, X.TIAN (2017). "An Enhanced Grey Wolf Optimization based Feature Selection Wrapped Kernel Extreme Learning Machine for Medical Diagnosis", Computational and Mathematical Methods in Medicine.

LIAO, B., Y. JIANG, W, LIANG, W. ZHU, L.CAI, C Z.AO (2014). Gene selection using locality sensitive Laplacian score. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), *11*(6), pp.1146-1156.

MALINA, W., (1981). On an extended Fisher criterion for feature selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, (5), pp.611-614.

MIRJALILI S., S.M. MIRJALIL,I D A. LEWIS (2014). "Grey Wolf Optimizer". Advances in Engineering Software, 69, Pp.46-61.

MOLER, E.J., M.L. CHOW, I.S.MIAN (2000). Analysis of molecular profile data using generative and discriminative methods. Physiological Genomics, *4*(2), pp.109-126.

NIIJIMA, S., Y. OKUNO (2009). Laplacian linear discriminant analysis approach to unsupervised feature selection. IEEE/ACM Transactions on Computational Biology and Bioinformatics, *6*(4), pp.605-614.

OLYAEE, S., Z.DASHTBAN, M.H. DASHTBAN (2013). Design and implementation of super-heterodyne nano-metrology circuits. Frontiers of Optoelectronics, *6*(3), pp.318-326.

RUIZ, R., J.C RIQUELME, J.S. AGUILAR-RUIZ (2006). Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognition, *3*9(12), pp.2383-2392.

SALEM, D.A., R.A.AA, A.SEOUD, H.A.LI (2011). A new gene selection technique based on hybrid methods for cancer classification using microarrays. International Journal of Bioscience, Biochemistry and Bioinformatics, *1*(4), p.261.

SALOME, J., R., SURESH (2011). An effective classification technique for microarray gene expression by blending of LPP and SVM. Medwell Journals: Asian Journal of Information Technology, *10*(4), pp.142-148.

SANDEN, S.V., D.LIN, T. BURZYKOWSKI (2008). Performance of gene selection and classification methods in a microarray setting: A simulation study. Communications in Statistics—Simulation and Computation®, *37*(2), pp.409-424.

STATNIKOV, A., L.WANG, C.F, ALIFERIS (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC bioinformatics, *9*(1), p.319.

TANG, Y., Y.Q. ZHANG, Z. HUANG (2007). Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), *4*(3), pp.365-381.

VUKUSIC, I., S.N. GRELLSCHEID, D T. WIEHE (2007). Applying genetic programming to the prediction of alternative mRNA splice variants. Genomics, *89*(4), pp.471-479.

WANG L., F.CHU, W. XIE (2007). Accurate Cancer Classification using Expressions of Very Few Genes, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), *4*(1), Pp.40-53.

WANG, Y., X. CHEN, W. JIANG, L, LI, W. LI, L.YANG, M. LIAO, B. LIAN, LV, Y. S.WANG, S. WANG (2011). Predicting human microRNA precursors based on an optimized feature subset generated by GA–SVM. Genomics, *98*(2), pp.73-78

WANG, Y., .V. TETKO, IM.A, HALL, E. FRANK, A. FACIUS, K.F MAYER,. D H.W. MEWES (2005). Gene selection from microarray data for cancer classification—a machine learning approach. Computational biology and chemistry, *29*(1), pp.37-46.

WEI Y., N NI., D LIU., H.CHEN, M.WANG, Q.LI, X. CU,I Y H.E (2017). "An Improved Grey Wolf Optimization Strategy Enhanced SVM and Its Application", Predicting the Second Major. Mathematical Problems in Engineering.

XIONG, W., Z.CAI, J. MA (2008). A DSRPCL-SVM approach to informative gene analysis. Genomics, proteomics & bioinformatics, *6*(2), pp.83-90.

XUAN, J., Y. WANG, Y. DONG, Y. FENG, B. WANG, J. KHAN, M. BAKAY, Z. WANG, L. PACHMAN S. WINOKUR, D Y.W., CHEN (2007). Gene selection for multiclass prediction by weighted fisher criterion. EURASIP Journal on Bioinformatics and Systems Biology, pp.3-3.

YANG, K., Z. CAI, J.LI, D G. LIN (2006). A stable gene selection in microarray data analysis. BMC bioinformatics, *7*(1), p.228.

YANG, P., D Z. ZHANG (2007). Hybrid methods to select informative gene sets in microarray data classification. AI 2007: Advances in Artificial Intelligence, pp.810-814.

YU, H., G. GU, H. LIU, J.SHEN, J. ZHAO (2009). A modified ant colony optimization algorithm for tumor marker gene selection. Genomics, proteomics & bioinformatics, *7*(4), pp.200-208.

ZHANG, J.G, D H.W., DENG (2007). Gene selection for classification of microarray data based on the Bayes error. BMC bioinformatics, 8(1), p.370.

NOVA *MICROARRAY* GENSKA SELEKCIJA I KLASIFIKACIJA KORIŠĆENJEM *INTELLIGENT DYNAMIC GREY WOLF* OPTIMIZACIJU

K.R.UTHAYAN

Department za Informacionu tehnologiju, SSN Koledž inžernjerstva, Indija

Izvod

Efikasna dijagnoza kancera je veoma važna za specifični tretman. Tačna predikcija različitih tipova kancera obezbeđuje bolji tretman i minimizuje toksičnost kod pacijenata. *Microarray* visoka dimenzionalnost podataka ekspresije gena i veliki broj gena nasuprot maloj veličini uzorka, šumu i ponavljanjima u setu podataka su glavni uzroci koji vode lošoj pouzdanosti klasifikacije. Selekcija informativnih gena i redukcija dimenzionalnosti, tehnika genske selekcije se koristi u microarray. U ovom radu, novi *meta-heurist* algoritam zasnovan na *Grey Wolf Optimization* (GWO) i veštačka intelegencija (AI) su kombinovane da dizajniraju model za klasifikaciju kancera. Ovaj rad ima dve faze. Prvo, filter metod kao što je Laplacian i Fisher skor, su primenjeni da ekstraktuju značajan podset svojstava za bržu klasifikaciju zatim *Intelligent Dynamic Grey Wolf Optimization* (IDGWO) je primenjen za identifikaciju odgovarajuće gene. GWO je *swarm*-zasnovan algoritam izabran za klasifikaciju podataka za ekspresiju gena, pošto omogućava lakšu klasifikaciju oko testiranja kancera. Značajna razlika između filter metoda su nađene korišćenjem nekoliko analiza. Predložen metod je primenjen na pet grana seta podataka uzimajući top 100 rangiranih gena izabranih pomoću fišer skora kod limpoma i SRBCT koje imaju a 100% performance korišćenjem IDGWO klasifikatora.

Primljeno 21.VII.2018.
Odobreno 18. V. 2019.