Second International Conference on Artificial Intelligence and its Applications (AIAP'2022)-Online

JANUARY 24-26, 2022 - EL OUED, ALGERIA,

ATTESTATION OF PARTICIPATION

This is to certify that

Ali Dabba

Has presented a paper:

Title: A New Gene Selection Method Based On Moth Flame Optimization

Authors: Ali Dabba, Abdelkamel Tari, Samy Meftali and Rabah Mokhtari

at the Second International Conference on Artificial Intelligence and its Applications, AIAP'22

held in El Oued, Algeria, JANUARY 24-26, 2022.

2nd International Conference on Artificial Intelligence and its Applications (AIAP'2022) University of El Oued

Prof. Brahim Lejdel
Chairman
Conference chair

A New Gene Selection Method Based On Moth Flame Optimization

Ali Dabba[1,2], Abdelkamel Tari[3,4] Samy Meftali[5,6], and Rabah Mokhtari[1,2]

[1] Faculty of Mathematics and Computer Science, Computer Science Department, Mohamed Boudiaf University, M'sila, Algeria,
ali.dabba@univ-msila.dz,
rabah.mokhtari@univ-msila.dz,
[2] Laboratory of Informatics and its Applications of M'sila - LIAM, Algeria,
[3] Faculty of Sciences, Computer Science Department, Abderrahmane Mira University, Bejaia, Algeria,
[4] Laboratory of Medical Computing - LIMD, Algeria,
abdelkamel.tari@univ-bejaia.dz,
[5] Faculty of Science and Technology, Lille University, France,
[6] Research center in Computer Science, Signal and Automatic Control of Lille - CRIStAL,France,
samy.meftali@univ-lille1.fr

Abstract. Cancer classification is an important issue addressed in the Bioinformatics field. In this paper, we present a novel extension of the Moth Flame Optimization Algorithm combined with Mutual Information Maximization (MIM) to solve gene selection problem called Mutual Information Maximization-modified Moth Flame Optimization Algorithm (MIM-mMFOA). The MIM-mMFOA has two phases: the first one is used to solve the difficulty of high-dimensional data, which measures redundancy and relevance of the gene, in order to obtain the relevant gene set. The second phase is dedicated to finding a small gene subset that can be used to classify samples with high accuracy, using a Support Vector Machine (SVM) with Leave One Out Cross Validation (LOOCV) classifier. In order to evaluate the performance of the proposed MIM-mMFOA, we test it on seven Microarray datasets. Experimental results show that MIM-mMFOA achieves a high classification accuracy in comparison to some known algorithms. ...

Keywords: Genes expression, Cancer Classification, Moth Flam Algorithm, Mutual Information, Bio-inspired Algorithms, Bioinformatics

1 Introduction

DNA Microarray is a modern biological research technology for analyzing gene expression. This technology has the ability to measure the expression levels of thousands of genes during important biological processes and provides the ability to diagnose cancer on the basis of gene expression [8].

In literature, several gene selection methods have been proposed and can be arranged into three main categories: filter, wrapper, and embedded methods

[11, 4]. However, we find several approaches that solve this issue, among them but not limited to, GA/SVM [7], GBC [3], GSP [1], PCC-BPSO/GA [6], and mABC [10].

In this paper, we propose a new algorithm called Mutual Information Maximization - modified Moth Flame Optimization Algorithm (MIM-mMFOA). MIM-mMFOA is proposed to investigate and improve the performance of gene selection. First, for high dimensional data, we use a pre-processing (Normalization and MIM) to handle this difficulty. Second, mMFOA uses three procedures, one for the presentation of individual, another for moth movement and the last is a new fitness function (an SVM with LOOCV classifier). Finally, the main objective of MIM-mMFOA is to select the best gene subset from among the predictive gene subsets, which is evaluated with the SVM classifier to provide high classification accuracy. Experimental results showed that the MIM-mMFOA achieves better performance of classification accuracy to solve the gene selection problem in binary class.

The remainder of the paper is organized as follows: The next section describes briefly the proposed method. Section 3 presented experimental results and discussion. Finally, the conclusion is given in Section 4.

2 The proposed algorithm

In this section, we propose a new MIM-mMFOA algorithm for predictive gene selection for cancer classification. This work is based on a hybrid approach between mutual information maximization and Moth-Flame Optimization Algorithm (MFOA), the principle of our proposed algorithm consists of two stages: pre-processing and modified Moth Flame Optimization Algorithm (mMFOA).

2.1 Pre-processing

This stage consists of two phases: Normalization and Mutual Information Maximization.

Normalization In this study, we used a feature scaling to normalize each gene value between $[a, b]$ as shown in the following Eq. 1.

$$X_{new} = (b - a)\frac{X - X_{min}}{(X_{max} - X_{min})} + a \tag{1}$$

Mutual Information Maximization (MIM) We use Mutual Information [15] to reduce the size of the initial problem. Formally, the MI is given by Eq. 2:

$$I(X, Y) = \sum_{x \in S} \sum_{x \in T} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \tag{2}$$

Consider expression data there are n samples and each one has m genes, the data can be represented by the matrix T of dimension $N \times M$.

$I(t, c)$: is mutual information of t of class c, it is calculated as follows .

$$I(t,c) = \log \frac{p(t \setminus c)}{p(t)} = \log \frac{p(t,c)}{p(t) \times p(c)} \approx \log \frac{\alpha \times M}{(\alpha + \beta) \times (\alpha + \delta)} \tag{3}$$

During the application of Eq.3, if the gene expression profile t is irrelevant to the class c, $I(t, c) = 0$. The MIM can be expressed as:

$$MaxMI(t) = \sum_{i=1}^{k} p(C_i \setminus t) \log \frac{p(C_i \setminus t)}{p(C_i)} \tag{4}$$

where k represents the number of classes in the dataset. The principle of the MIM method is each score $MaxMI(G_j)$ of gene G_j $(j = 1, ..., m)$ is calculated independently of all other genes in the same class by Eq.4.

2.2 Modified MFO algorithm for genes expression selection

MFOA is a nature-inspired algorithm that was developed by [9]. In the following steps, we mention the main phases of mMFOA:

Step 1: *Representation of candidate solutions and initialization of population:*

In our work, the moths can fly in two-dimension ($2D$), considered as a candidate solution, the flames are the best position of moths that obtain so far, these both represented by two matrices ($n \times d$), see Eq.5.

$$M = \begin{bmatrix} (m^x_{1,1}, m^y_{1,1}) & (m^x_{1,2}, m^y_{1,2}) & \cdots & \cdots & (m^x_{1,d}, m^y_{1,d}) \\ (m^x_{2,1}, m^y_{2,1}) & (m^x_{2,2}, m^y_{2,2}) & \cdots & \cdots & (m^x_{2,d}, m^y_{2,d}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (m^x_{n,1}, m^y_{n,1}) & (m^x_{n,2}, m^y_{n,2}) & \cdots & \cdots & (m^x_{n,d}, m^y_{n,d}) \end{bmatrix}; F = \begin{bmatrix} (f^x_{1,1}, f^y_{1,1}) & (f^x_{1,2}, f^y_{1,2}) & \cdots & \cdots & (f^x_{1,d}, f^y_{1,d}) \\ (f^x_{2,1}, f^y_{2,1}) & (f^x_{2,2}, f^y_{2,2}) & \cdots & \cdots & (f^x_{2,d}, f^y_{2,d}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (f^x_{n,1}, f^y_{n,1}) & (f^x_{n,2}, f^y_{n,2}) & \cdots & \cdots & (f^x_{n,d}, f^y_{n,d}) \end{bmatrix} \tag{5}$$

Our proposition starts with a random initialization of the moth sets population.

Step 2: *Gene subset selection and fitness calculation for each moth (individual):*

In this step, we used the *Gene_select* function to determine which genes are selected in this individual (subset).

The pseudo-code of *Gene_select* function is given by Algorithm 1.

In this work, we define the fitness value of each individual (moth i) as follows (Eq.6):

$$Fitness_i = \begin{cases} w_1 * Acc_{SVM\_with\_LOOCV}(Moth_i) + w_2 * \frac{t_g - s_g}{t_g} \\ With \quad w_1 + w_2 = 1 \end{cases} \tag{6}$$

$Acc_{SVM\_with\_LOOCV}$ is the accuracy of SVM with LOOCV classifier, t_g and s_g is the number of total and selected genes, respectively. w_1 and w_2 are the coefficients of each part of fitness.

Algorithm 1 Genes selection pseudo-code

```
1: function Gene_select(i, NbrGenes, ϑ)        ▷ Where i: number of the individual
2:     SG[NbrGenes] array of Boolean
3:     for (j ← 1 to NbrGenes) do
4:         D ← sqrt((m^x_{i,j})^2 + (m^y_{i,j})^2)
5:         if sigmoid(D) > ϑ then
6:             SG[j] ← true
7:         else
8:             SG[j] ← false
9:         end if
10:    end for
11:    return SG
12: end function
```

Step 3: *Update moth position*
The next moth coordinates are calculated by the following Eq.7:

$$\begin{cases} X = \rho\cos(\theta+\alpha) + X_F \\ Y = \rho\sin(\theta+\alpha) + Y_F \end{cases} \tag{7}$$

where ρ is the distance between the next position of i^{th} moth and the j^{th} flame given by the Archimedes spiral function $\rho = a.\theta$ where θ is a random angle in $[0, 2k\pi]$ (see Fig.1). α is the angle between the line (FM) and the x-axis in the landmarks R.

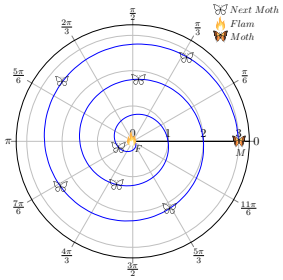

Fig. 1. Logarithmic Archimedes spiral, space around a flame, and some next possible moth positions (when k=3).

Step 4: *Update number of flames*
In our work, we used the following formula proposed by [9].

$$Nbr\_F = round(Max_{Nbr\_F} - C_{it} * \frac{Max_{Nbr\_F} - 1}{Max_{it}}). \tag{8}$$

Where C_{it} and Max_{it} represent the current and maximum number of iterations, respectively, Max_{Nbr_F} is the maximum number of flames.

Step 5: *Stopping Criterion is satisfied*
If a maximum number of iterations is reached, the iterative process stops for extract and evaluate the best subset of genes. Otherwise, return to Step 2.

The pseudo-code of mMFOA is given by Algorithm 2.

Algorithm 2 Pseudo-code of the mMFOA

for ($i \leftarrow 1$ **to** Pop_Size) **do**
  $Random\_initialization(P[i])$
end for
$Iteration \leftarrow 1$
repeat
  Nbr_F is calculated using (Eq. 8)
  for ($i \leftarrow 1$ **to** Pop_Size) **do**
    $SG_M[i] \leftarrow selection - genes(P[i])$
    $FM[i] \leftarrow$ Evaluate $PM[i]$ to fitness function using (Eq. 6)
  end for
  if $Iteration = 1$ **then**
    $PF, FF, Subset\_F \leftarrow sort(PM, FM, Subset\_M)$;
  else
    $PF, FF, Subset\_F \leftarrow sort([PF, PM], [FF, FM], [Subset\_F, Subset\_M])$;
  end if
  for ($i \leftarrow 1$ **to** Pop_Size) **do**
    for ($j \leftarrow 1$ **to** Nbr_F) **do**
      Calculate (ρ) using $\rho = a.\theta$.
      Update the moth position using (Eq. 7)
    end for
  end for
  $Subset_{best} \leftarrow PF[0], FF[0], Subset\_F[0]$
  $Iteration \leftarrow Iteration + 1$
until ($Iteration > Max\_iteration$)
return $Subset_{best}$

3 Results and discussion

In our work, we used two comparison parameters : classification accuracy and the number of predicted genes.

3.1 Dataset and Parameters Settings

In this study, Table 1 detailing seven binary microarray datasets. The MIM-mMFOA parameters used in our experiments are shown in Table 2.

Table 1. Summary of gene expression datasets.

Dataset Name	Samples	Features	Classes	Source
CNS	60	7129	2 (Binary class)	[16]
Colon	62	2000	2 (Binary class)	[2]
Leukemia1	72	7129	2 (Binary class)	[5]
Breast	97	24481	2 (Binary class)	[16]
Ovarian	253	15154	2 (Binary class)	[12]
DLBCL	77	5469	2 (Binary class)	[13]
Prostate_Tumor	102	10509	2 (Binary class)	[14]

Table 2. MIM-mMFOA parameters.

Parameters	Setting value
Population size	50
Normalization interval	[-1, 1]
Top-ranked genes	100
Random angle	$[0, 6\pi]$
w_1	0.70
w_2	0.30
Number of generation	30

3.2 Experimental Results and Analysis

In this section, we present and analyze the results obtained by MIM-mMFOA. In order to prove the high-performance of MIM-mMFOA, it should be compared with various gene selection methods.

Table 3 shows the accuracy and the number of genes selected by MIM-mMFOA in seven binary Microarray datasets.

From the Table 3, we can see, that MIM-mMFOA can obtain 100% (Best, worst and average) accuracy with zero standard deviation (S.D) for the leukemia1, DLBCL, Colon, and Prostate_Tumor. Also, we can see, that leukemia1 the best number of selected genes is (7) seven. For the datasets DLBCL, CNS, Colon, Breast and Prostate_Tumor, the MIM-mMFOA can provide the best number of selected genes between 20 and 10.

Table 4 compared MIM-mMFOA with well-known gene selection algorithms published in the literature, applied to the binary class.

In Table 4, we have highlighted the best results in classification accuracy and number number of genes selected. As we can see, the MIM-mMFOA obtains the highest results in five out of seven cancer microarray datasets in terms of classification accuracy, and in three out of seven datasets in terms of number of genes.

Based on the above analysis, we can conclude that MIM-mMFOA can perform better classification accuracy than other algorithms.

Table 3. Experimental results by MIM-mMFOA on all datasets.

Dataset	Accuracy				# Genes			
	Best	**Worst**	**Avg.**	**S.D.**	**Best**	**Worst**	**Avg.**	**S.D.**
Leukemia	100,00	100,00	100,00	0,00	6,00	9,00	7,50	0,97
DLBCL	100,00	100,00	100,00	0,00	11,00	18,00	14,70	2,00
CNS	100,00	98,33	99,83	0,53	13,00	31,00	24,70	6,09
Colon	100,00	100,00	100,00	0,00	20,00	31,00	26,30	3,56
Ovarian	98,42	98,02	98,18	0,20	26,00	40,00	35,90	4,70
Breast	91,75	83,51	86,80	3,10	11,00	45,00	25,90	9,60
Prostate_Tumor	100,00	100,00	100,00	0,00	14,00	23,00	18,60	2,63

Table 4. Comparison of experimental results obtained by MIM-mMFOA with other methods for binary class datasets.

Algorithms		**Dataset**	**Leukemia1**	**DLBCL**	**Prostate Tumor**	**CNS**	**Colon**	**Breast**	**Ovarian**
MIM-mMFOA	Accuracy	Best	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	91,75	98,42
		Worst	**100,00**	**100,00**	**100,00**	**98,33**	**100,00**	83,51	98,02
		Avg.	**100,00**	**100,00**	**100,00**	**99,83**	**100,00**	86,80	98,18
		S.D.	**0,00**	**0,00**	**0,00**	**0,53**	**0,00**	3,10	0,20
	# Genes	Best	6,00	11,00	14,00	**13,00**	**20,00**	**11,00**	26,00
		Worst	9,00	18,00	23,00	**31,00**	**31,00**	**45,00**	40,00
		Avg.	7,50	14,70	18,60	**24,70**	**26,30**	**25,90**	35,90
		S.D.	0,97	2,00	2,63	**6,09**	**3,56**	**9,60**	4,70
PCC-BPSO	Accuracy	**Best**	**100,00**	-	97,06	98,33	91,94	90,72	**100,00**
	# Genes	Best	18,00	-	33,00	39,00	25,00	41,00	**17,00**
PCC-GA	Accuracy	**Best**	**100,00**	-	96,08	98,33	91,94	88,66	**100,00**
	# Genes	Best	35,00	-	26,00	48,00	29,00	38,00	22,00
GBC	Accuracy	Best	**100,00**	-	-	-	98,38	-	-
		worst	93,05	-	-	-	91,93	-	-
		Avg	96,43	-	-	-	94,62	-	-
	# Genes	Best	5,00	-	-	-	20,00	-	-
mABC	Accuracy	Best	**100,00**	**100,00**	**100,00**	-	-	-	-
		Avg.	**100,00**	**100,00**	**100,00**	-	-	-	-
		S.D.	0,00	0,00	0,00	-	-	-	-
	# Genes	Best	**4,00**	**3,00**	**5,00**	-	-	-	-
		Avg.	5,67	**4,05**	10,73	-	-	-	-
		S.D.	0,73	**0,78**	3,15	-	-	-	-

4 Conclusions

In this paper, we presented a new bio-inspired algorithm for gene selection problem. Our approach consists of two stages : pre-processing in order to reduce the initial size of the input dataset and mMFOA to select the best gene subset.

The overall goal of this paper is to select a smaller number of genes and achieve similar or better classification accuracy than using all genes. The tests of MIM-mMFOA on seven binary classes of datasets show that our algorithm is better than all other compared algorithms to the classification accuracy and provides competitive results with the number of genes.

We plan, in future work, to extend our algorithm to address other issues not yet tackled.

Bibliography

[1] Alanni R, Hou J, Azzawi H, Xiang Y (2019) A novel gene selection algorithm for cancer classification using microarray datasets. BMC medical genomics 12(1):10

[2] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences 96(12):6745–6750

[3] Alshamlan HM, Badr GH, Alohali YA (2015) Genetic bee colony (gbc) algorithm: A new gene selection method for microarray cancer classification. Computational biology and chemistry 56:49–60

[4] Du D, Li K, Li X, Fei M (2014) A novel forward gene selection algorithm for microarray data. Neurocomputing 133:446–458

[5] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science 286(5439):531–537

[6] Hameed SS, Muhammad FF, Hassan R, Saeed F (2018) Gene selection and classification in microarray datasets using a hybrid approach of pcc-bpso/ga with multi classifiers. JCS 14(6):868–880

[7] Huerta EB, Duval B, Hao JK (2006) A hybrid ga/svm approach for gene selection and classification of microarray data. In: Workshops on Applications of Evolutionary Computation, Springer, pp 34–44

[8] Jeffrey SS, Lønning PE, Hillner BE (2005) Genomics-based prognosis and therapeutic prediction in breast cancer. Journal of the National Comprehensive Cancer Network 3(3):291–300

[9] Mirjalili S (2015) Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. Knowledge-Based Systems 89:228–249

[10] Moosa JM, Shakur R, Kaykobad M, Rahman MS (2016) Gene selection for cancer classification with the help of bees. BMC medical genomics 9(2):47

[11] Mundra PA, Rajapakse JC (2010) Gene and sample selection for cancer classification with support vectors based t-statistic. Neurocomputing 73(13-15):2353–2362

[12] Petricoin III EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, et al (2002) Use of proteomic patterns in serum to identify ovarian cancer. The lancet 359(9306):572–577

[13] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, et al (2002) Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature medicine 8(1):68

[14] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, et al (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer cell 1(2):203–209

[15] Venkateswara H, Lade P, Lin B, Ye J, Panchanathan S (2015) Efficient approximate solutions to mutual information based global feature selection. In: 2015 IEEE International Conference on Data Mining, IEEE, pp 1009–1014

[16] Zhu Z, Ong YS, Dash M (2007) Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognition 40(11):3236–3248