Genetic Algorithm as a Key Parameter of SVM parameter optimization and feature selection for acute Leukemia diagnosis

Najaat Ahmed Abdullah[1], Mohammed Abduljalil Ibrahim[2] and Adel Sallam Haider[1]
[1] Department of Information Technology, Faculty of Engineering, Aden University, Yemen
[2] Department of Information Technology, Faculty of Engineering, Taiz University, Yemen
[1] najaat200020@gmail.com, [2] Sabri1966@yahoo.com, [3] Haider.Adel@gmail.com
DOI: https://doi.org/10.47372/uajnas.2020.n2.a07

Abstract

The selection process of the kernel parameters and the relevant features are very crucial to enhance the classification tasks. Thus, in this work, a genetic algorithm that mimics the biological evaluation is used to optimize the support vector machine kernel parameters in order to achieve a high classification accuracy of an acute leukemia diagnosis. The results proved that the combination of genetic algorithm with support vector machine increased the classification accuracy of acute leukemia diagnosis to 99.19%, compared with the value of 89.43% obtained under default support vector machine kernel parameters. This can be directly attributed to the elimination of the irrelevant features and the suitable selection of the kernel parameters. This implies that the genetic algorithm model can be adequately used to solve the optimization problem and features subset selection that gives the optimal accuracy.

Keywords: acute leukemia, support vector machine, genetic algorithm, optimization, feature selection

Introduction

Increasing a training time and the associated overfitting risk are the major problems which affect the performance of the models in the image recognition systems due to the very high dimensionality of the data (feature set) [2]. In addition, the presence of informative features reduces the classification accuracy. In this connection, feature selection methods in machine learning process have been selected for removing unnecessary, redundant or irrelevant features in the dataset. Generally, feature subset selection methods are classified into two approaches; a wrapper approach and a filter approach [16]. The first is related to the use of the classification algorithm to evaluate the goodness of the features during the feature selection process, while the second is independent of any classification algorithm.

Support Vector Machine (SVM), as a classifier in the wrapper approach, is a supervised learning model that tries to find the optimal hyperplane which separates the data points according to their class labels [3]. This can be done by maximizing the margin between separating hyperplane and the closest data points of each class. To do this proposed or model the SVM used different kernel functions such as linear, polynomial, Radial Basis Function (RBF) and sigmoid. These mathematical functions transform the input data into the required form in order to separate the non-linearly separable classes [5]. However, the selection of vital subset features and the best parameters of the kernel functions limits the accuracy of the SVM classification. To solve these problems, Genetic Algorithm (GA) and Practical Swarm Optimization (PSO) have been cited in the literature as appropriate methods for feature subset selection and parameter optimization. Thus, the objective of this work is to study the impact of the combination process between the GA and the SVM on the classification efficiency of the acute leukemia diagnosis process.

This paper is organized as follows; after a brief reviewing of the related works, a brief description of the SVM and the GA techniques are presented in the following sections. Then, the

experimental approach and the results of the study are discussed to make up the paper's body. Finally, the ability of GA as an efficient tool for acute leukemia diagnosis is concluded.

Related works

Over the last several years, the related works of our proposed method have been divided into three different research areas. In the first area, most of the researchers focused on the optimization of the SVM parameters using different techniques. For example, Bamakan et al. [3] used a hybrid approach based on Practical Swarm Optimization (PSO) technique to determine the optimal value for Non Parallel SVM (NPSVM) parameters in order to overcome the drawbacks of Twin-SVM. Although the PSO-NPSVM achieved better classification accuracy, it can be trapped into local optimum. Under the same area, Kharrat et al. [6] built the GA-SVM model by five selection features for optimization of the SVM parameters. The combine GA with SVM can have benefits in terms of accuracy and computational efficiency in spite of its long processing time, compared to a statistical approach (Grid search). However, Syarif et al. [15] demonstrated that the optimization of the SVM parameters, using GA, was more stable and almost 16 times faster than the grid search which can be attributed to high dimensional datasets with a suitable range of parameters.

Due to the negative impact of high dimensional features on the performance of the classifiers, the second area of studies involves the researchers who focused on the feature subset selection. Babatunde et al. [2] developed a GA-based feature selector using a novel fitness function (K Nearest Neighbors KNN-based classification error). The ability of GA based feature selector to change the fitness function of all of the selected features produces better classification accuracy, compared to the Waikato Environment for Knowledge Analysis (WEKA) ranker. Similarly, Singh et al. [14] showed that the good performance of the GA-based feature selection to remove irrelevant features from the medical dataset can be evaluated by using Naïve Bayes, J48 and KNN classifiers.

In the final area, researchers tried to combine the above two areas in order to decrease the training time and associated overfitting risk to enhance the classification accuracy. In this manner, Huang and Wang [5] and Chen et al. [4] suggested the GA-based approach and its coarse-grained parallel (CPGGA), respectively, for optimizing the SVM parameters and subset feature selection in order to overcome the degrading in SVM classification accuracy from UIC database. This combination process helped in finding the optimal feature subset and efficient kernel parameters which significantly decrease the training time and increase the quality obtained solution.

Support vector machine brief description

In actual practice, separating data into training and testing sets is the first stage in a classification process. To do this, many useful techniques have been used in the literature such as SVM, KNN, Naive Bayes... etc. Conceptually, in the classification process of binary problems, each sample (instance) of the training set contains one label and several features. Thus, the SVM uses this training data to construct a model that works as an indicator to predict the label of the test data [4, 5, 8].

Suppose that the sample vector of the training set $x_k, x_k \in R^n$ and the corresponding labels $y_k \in \{+1, -1\}$, where R^n is the sample space, the hyperplane can be described as,

$$w^T x + b = 0 \quad (1)$$

, where w is a bias (scalar). The distance D(k) from a point x_k in the feature space to the hyperplane can be defined as

$$D(k) = \frac{w^T x_k + b}{\|w\|}, k = 1, \ldots, m \quad (2)$$

1- Linear SVM

The SVM can find the optimal separating hyperplane which maximizes the minimum value of the distance by solving the optimization problem described in the equation when the training samples are linearly separable.

$$\min_{w,b} \frac{1}{2}\|w\|^2 \qquad \text{Subjected to: } y_k(w^T x_k + b) \geq 1, k = 1, \dots, m \tag{3}$$

However, the optimal hyperplane cannot be used to classify the data points correctly without errors in the case of linearly non-separable. Thus, the slack variable ξ_k will be introduced into the optimization problem, as defined in equation

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{k=1}^{m} \xi_k \quad \text{Subjected to: } y_k(w^T x_k + b) \geq 1 - \xi_k \, , k = 1, \dots, m \tag{4}$$

, where C is the penalty parameter that is use to balance the training error and margin. After that, the Lagrange multipliers α_k is used to solve the optimization problem in equation (4) by transforming it into a dual form as

$$\max_{\alpha} L = \sum_{k=1}^{m} \alpha_k - \frac{1}{2}\sum_{k,l=1}^{m} \alpha_k \alpha_l y_k y_l (x_k . x_l) \qquad \text{Subjected to: } \alpha_k \geq 0 \ \& \ \sum_{k=1}^{m} \alpha_k y_k = 0 \tag{5}$$

Finally, the label is assigned to a sample in the feature space according to the equation

$$y = sign(\sum \alpha_k y_k x_k x + b) \tag{6}$$

2- Non-Linear SVM

When the SVM cannot draw a straight line to classify the data points, the data are converted to linearly separable by mapping into higher dimensional space. In this case, the inner product of the two vectors $(x_k . x_l)$ in equation (5) will be replaced by the kernel function, i.e.

$$(x_k, x_l) = \phi(x_k).\phi(x_l) \tag{7}$$

Therefore, the decision function in the equation becomes as

$$y = sign(\sum \alpha_k y_k k(x_k, x) + b) \tag{8}$$

, where the popular kernel functions are $x_k . x_l$, $(1 + x_k . x_l)^d$, $\exp(-\gamma\|x_k - x_l\|^2)$, $\tanh(kx_k . x_l - \delta)$ for Linear kernel, Polynomial kernel, Radial Bases Function (RBF) kernel and Sigmoid kernel, respectively.

Genetic algorithm brief description

Genetic algorithms GAs are a heuristic search and optimization techniques built based on a natural selection process to mimic the biological evaluation and to find the optimal solutions of the difficult problems. For any given problem in machine learning, GAs manipulated the population of candidate solutions (chromosomes or individuals) that can solve the problem. Subsequently, each candidate solution is evaluated by assigning a fitness value to reproduce and to mate in order to form a new population for the next generation [4]. After a number of generations, GA can be able to get acceptable results that satisfy the termination criteria, as shown in Fig. 1 [8].

SVM parameter optimization and feature selection based on GA for diagnosis acute leukemia

In general, GA uses bit string in order to design a suitable chromosome, which is use to produce a fitness value that is evaluated by the system architecture. The general steps used in the chromosome design, fitness function building and system architecture in our proposed system are described as follows:

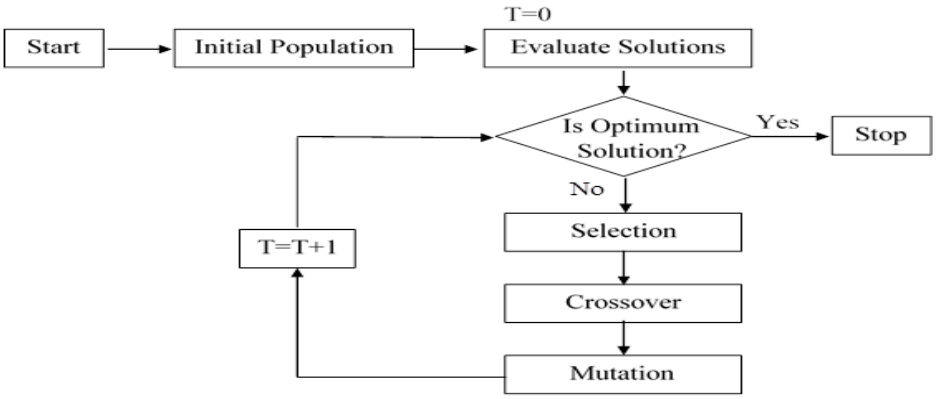

Fig. 1 Simple Flowchart for genetic algorithm

1- Chromosome Design

In this step, feature subset of the input data and different kernel function parameters are chosen to represent the binary code of each chromosome, which consists of three parts; parameter $\mathrm{C}$and γ and feature subset, as shown in Fig. 2.

Penalty Cost C	Gamma γ	Features mask
$\mathrm{C}_1 \ldots . \mathrm{C}_{n\mathrm{C}}$	$\gamma_1 \ldots . \gamma_{n\gamma}$	$f_1 \ldots . f_{nf}$

Fig. 2 Chromosome Coding

where, $\mathrm{C}_1 \ldots . \mathrm{C}_{n\mathrm{C}}$ is the binary code of, $\gamma_1 \ldots . \gamma_{n\gamma}$ is the binary code of, $f_1 \ldots . f_{nf}$ is the binary code of feature mask and $n\mathrm{C}, n\gamma$ and nf are the number of bits of the parameters C, γ and f, respectively.

The number of bits $n\mathrm{C}$ and $n\gamma$ are selected according to the computational precision. Since the decimal natural of SVM, the genotype of the parameters (C, γ) should be transformed into a phenotype, using the equation

$$p = min_p + \frac{max_p - min_p}{2^m - 1} * dv \quad (9)$$

, where p is the bit string, min_p and max_p are the minimum value and the maximum value of the parameter respectively, dv is the decimal value of the bit string and m is the length of the bit string. In the feature subset, coding 1indicates that the feature is chosen and coding 0 the feature is not chosen. In chromosome design, the bit with value 1 indicates that the feature is selected. The bit with value 0, on the other hand, indicates that the feature is not selected.

2- Fitness Function

To assess the performance of individual chromosome fitness function that produces high classification accuracy is selected in our proposed approach, as defined in the equation.

$$fitness = w_{acc} * accuracy \quad (10)$$

The predefined weight w_{acc} can be adjusted from 75% to 100%, according to study requirements [5].

3- System Architecture

To establish GA-SVM system architecture, Fig. 3 presents a details that should be proceeded. Firstly, the input data extracted from acute leukemia images are split into two sets; training set and

testing set. The dominating of the attributes in greater numeric range is avoided by data scaling in the range of [0, 1] by the equation.

$$sv = \frac{ov - min_f}{max_f - min_f} \tag{11}$$

, where sv is the scaled value, ov is the original value min_f and max_f are the lower and upper bound of the feature value respectively.

Secondly, the genotype parameters C and γ of the randomly generated population will be converted to the phenotype. each chromosome phenotype and selected features with 70% of the input.

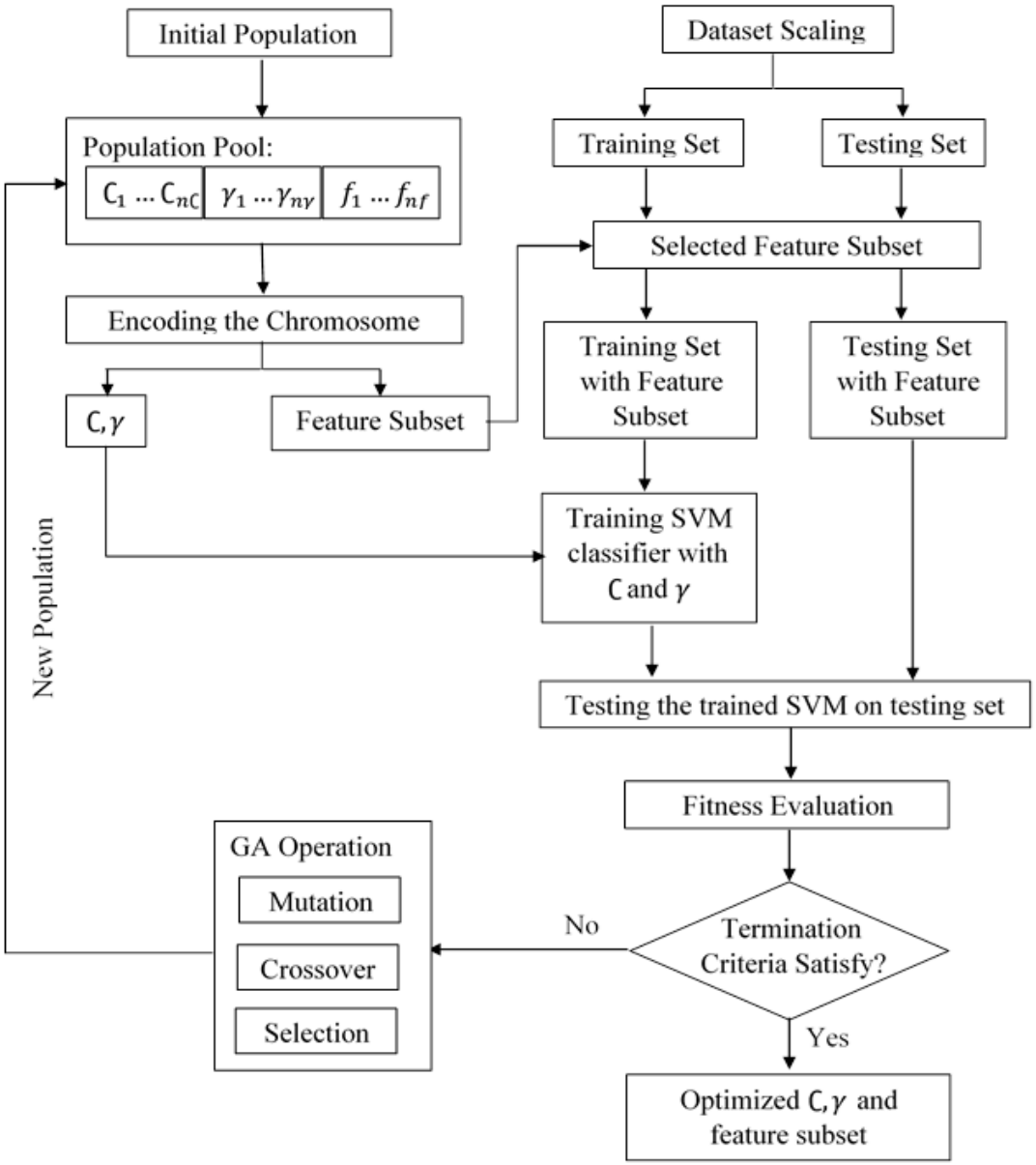

Fig. 3 Flowchart for the proposed SVM parameter optimization and feature subset selection based on GA

data, i.e. training data, are chosen to construct the SVM model. The remaining 30% of the input data will be use as testing data to calculate the fitness value through the classification accuracy. Finally, if the generation numbers satisfy the termination criteria, the process will stop [6]. That is the fitness value does not increase during the last number of generations or the maximum number of generations that have been reached; otherwise, the system searches for the best solutions using genetic operations, including selection, crossover and mutation.

Experiments Description

The GA of the initial parameters used are listed in Table 1. The parameter values chosen is based on the highest accuracy of training set experiments.

Table 1. Initial Parameters of GA used

GA Parameter	Setting
C	$0 - 2^{20}$
γ	$2^{-20} - 0$
Generations Number	300
Population Size	100
Selection type	Tournament
Crossover type	Single-point crossover
Crossover Probability	0.8
Mutation Probability	0.1

The empirical evaluation was performed in Intel® Core™ i7-2670 QM CPU@ (2.20 GHz, 8 cores), 8 GRAM. Run under Windows 7 operating system. The development environment is MATLAB 2017 and the SVM software is Libsvm [9].

The extracted data were collected from Acute Lymphoblastic Leukemia Image Database for Image Processing IDB-ALL [7], American Society of Hematology (ASH) image bank [1], Pathpedia [11] and sutterstock image bank of medical images [13]. The selected data consist of 132 images: round 50 normal images and 82 belong to Acute Lymphoblastic Leukemia (ALL) and Acute Myeloblastic Leukemia (AML). According to the equation, the w_{acc} was set to 100% under the termination criteria of 300 generations, or the fixed fitness value during the last 150 generation.

Therefore, classification measurements used to evaluate the SVM performance are given in Table 2 [15].

Table 2. classification measurements used to evaluate the SVM performance

Measure	Formula
$Precision$	$Precision = \frac{TP}{TP + FP}$
$Sensitivity$	$Sensitivity = \frac{TP}{TP + FN}$
$Accuracy$	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
F-Measure	$F - Measure = \frac{2}{(1/Precision) + (1/Sensitivity)}$

, where TP is the True Positive, FN is the False Negative, FP is the False Positive and TN is the True Negative.

Result and Discussion

Table 3 gives the details parameters for the GA-SVM model used for initial and optimized kernel parameters and feature subset selection.

Table 3. Performance of the proposed approach

	Initial parameters			**With optimization and feature selection**
Kernel Function	Linear	Polynomial	RBF	RBF
Number of features	283	283	283	132
Parameter C	1	1	1	1028771
Parameter γ	1/k	1/k	1/k	$8.67*10^{-7}$
Accuracy %	87.8049	87.8049	**89.4309**	**99.1870**
Precision %	85.9524	83.6859	94.0367	99.4845
Sensitivity %	76.2153	78.8773	75.9259	98.1481
F-Measure %	80.7915	81.2105	**84.0164**	**98.8118**
where; k is the number of the features				

The highest accuracy of the initial values of the RBF kernel function was 89.43%. Thus, the GA-SVM module used RBF kernel function for optimization and feature subset selection. However, the classification accuracy value has improved to 99.19% when the genetic algorithm was implemented to optimize the RBF kernel parameters and feature subset selection. The number of features used was 132 on the basis of the maximum value of the fitness function. This accuracy is approximately the same with the value obtained by Rawat et al. [12] and Negm et al. [10] which is 99.5%. However, our accuracy is still better than the value obtained from other literatures.

Figure 4 depicts the evolving process of the fitness value during the genetic algorithm implementation. The process is approximately characterized by four phases, in the first phase ,The fitness value at the 5 generation has slowly increased from 95.12% to 96.75% ,during 5 generations, in the second phase, the fitness value has stepped significantly to 97.56% during 27 generations. in the third phase, the fitness value has increased to 98.37% during 31 generations and, in the final phase, the value has risen up to 99.19% and remains at the same value until the termination criteria satisfy at the 217 generation. These results imply that the proposed GA-SVM model significantly improved the acute leukemia diagnostic process.

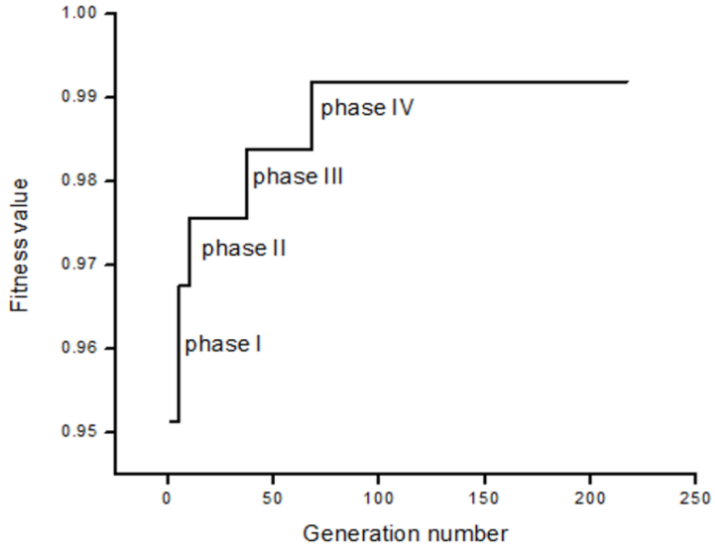

Fig. 4 The evolution process of the fitness value during the GA implementation

Conclusion

The genetic Algorithm, with the support vector machine has been successfully used for computing optimization of the kernel parameters and feature subset selection. This model has higher accuracy of 99.19% with fewer number of features of 132 obtained with the RBF kernel function. These results point up the significance of searching for optimal parameter values and the feature subset selection that achieve the highest accuracy performance. Hence, this model can well used for the medical application of acute leukemia and it might be possible to extend to other types of cancer.

References

1. ASH Image Bank. (2018). "Acute myeloid leukemia," *American Society of Hematology*,
2. Babatunde, O. H., Armstrong, L., Leng, J. and Diepeveen, D. (2014), "A genetic algorithm-based feature selection," *International Journal of Electronics Communication and Computer Engineering*, (Vol. 5, pp. 899-905).
3. Bamakan, S. M. H., Wang, H. and Ravasan, A. Z. (2016), "Parameters optimization for nonparallel support vector machine by particle swarm optimization," *Procedia Computer*
4. Chen, Z., Lin, T., Tang, N. and Xia, X. (2016), "A parallel genetic algorithm based feature selection and parameter optimization for support vector machine," *Scientific Programming*, (Vol. 2016).
5. Huang, C.-L. and Wang, C.-J. (2006), "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Systems with applications*, (Vol. 31, pp. 231-240).
6. Kharrat, A., Benamrane, N., Messaoud, M. B. and Abid, M. (2011), "Evolutionary Support Vector Machine for Parameters Optimization Applied to Medical Diagnostic," in VISAPP, (pp. 201-204).
7. Labati, R. D., Piuri, V. and Scotti, F. (2011), "All-IDB: The acute lymphoblastic leukemia image database for image processing," in Image processing (ICIP), 2011 18th IEEE international conference on, (pp. 2045-2048): IEEE.
8. Li, J., Zhang, B. and Shi, J. (2017), "Combining a genetic algorithm and support vector machine to study the factors influencing CO2 emissions in Beijing with scenario analysis," Energies, (Vol. 10, pp. 1520).
9. LIBSVM. (2018, July 15). Libsvm -- a library for support vector machines. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/libsvm/
10. Negm, A. S., Hassan, O. A. and Kandil, A. H. (2017), "A decision support system for Acute Leukaemia classification based on digital microscopic images," Alexandria Engineering Journal, (Vol. 57, pp. 2319-2332).
11. PathPedia. (2019). "Histopathology of blood cells," PATHPEDIA, LLC (c), USA, Retrieved from https://www.pathpedia.com/education/eatlas/histopathology/blood_cells.aspx
12. Rawat, J., Singh, A., Bhadauria, H., Virmani, J. and Devgun, J. S. (2017), "Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia," *Biocybernetics and Biomedical Engineering*, (Vol. 37, pp. 637-654).
13. Shutterstock. (2019). "Acute myelogenous leukemia images," *Shutterstock, Inc.*, New York, Retrieved from https://www.shutterstock.com/search/acute+myelogenous+leukemia
14. Singh, D. A. A. G., Leavline, E. J., Priyanka, R. and Priya, P. P. (2016), "Dimensionality reduction using genetic algorithm for improving accuracy in medical diagnosis," *International Journal of Intelligent Systems and Applications*, (Vol. 8, pp. 67).
15. Syarif, I., Prugel-Bennett, A. and Wills, G. (2016), "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," *Telkomnika*, (Vol. 14, pp. 1502).
16. Zhuo, L., Zheng, J., Li, X., Wang, F., Ai, B. and Qian, J. (2008), "A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine," in *Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images*, (Vol. 7147, pp. 71471J): International Society for Optics and Photonics.

الخوارزمية الوراثية كحجر أساس لاختيار القيم المثلى لمعاملات آلة المتجهات الداعمة والسمات المتباينة في عملية تشخيص سرطان الدم الحاد

نجاة أحمد عبدالله[1]، محمد عبدالجليل إبراهيم[2] و عادل سلام حيدر[1]

[1] قسم ت تكنولوجيا المعلومات، كلية الهندسة، جامعة عدن، اليمن.

[2] قسم تكنولوجيا المعلومات، كلية الهندسة، جامعة تعز، اليمن.

[1] najaat200020@gmail.com, [2]Sabri1966@yahoo.com, [3]Haider.Adel@gmail.com

DOI: https://doi.org/10.47372/uajnas.2020.n2.a07

الملخص

في أنظمة التعلم الآلي تُحَّديد السمات المتباينة واختيار معاملات دالة kernel التي تستخدم لربط بيانات العنصر المدروس بأبعاد أعلى حتى نتمكن من تشكيل مستوى فائق hyperplane لفصلها وتحديد السمات هدة يُعد من المعايير الهامه لتطوير عملية التصنيف. لذا في هذا العمل، الخوارزمية الوراثية Genetic Algorithm التي تحاكي التطور البيولوجي استخدمت لتحديد القيم المثلى لمعاملات دالة kernel في آلة المتجهات الداعمة Support Vector Machine لغرض إنجاز عملية التصنيف لسرطان الدم الحاد بدقة عالية. نتائج هذه الدراسة اثبتت أن عملية الدمج بين الخوارزمية الوراثية وآلة المتجهات الداعمة رفعت من دقة عملية التصنيف لسرطان الدم الحاد إلى 99.19% مقارنة بـ 89.43% التي تم الحصول عليها باستخدام معاملات دالة kernel الافتراضية. هذه الدقة العالية في التصنيف يمكن إرجاعها إلى قدرة الخوارزمية الوراثية على استبعاد السمات الاقل تبايناً وإيجاد القيم المثلى لمعاملات دالة kernel، مشيرةً إلى أن النظام المقترح يُعد حلاً مناسبًا لعملية الأمثلة optimization وانتقاء السمات الفرعية feature subset selection في عملية تصنيف سرطان الدم الحاد.

الكلمات المفتاحية: سرطان الدم الحاد، آلة المتجهات الداعمة، الخوارزمية الوراثية، الأمثلة، انتقاء السمات.