Original Article

Biomarker Discovery by Imperialist Competitive Algorithm in Mass Spectrometry Data for Ovarian Cancer Prediction

Shiva Pirhadi[1], Keivan Maghooli[1], Niloofar Yousefi Moteghaed[2], Masoud Garshasbi[3], Seyed Jalaleddin Mousavirad[4]

[1]Department of Biomedical Engineering, Tehran Science and Research Branch, Islamic Azad University, [2]Department of Biomedical Engineering and Medical Physics, Faculty of Medicine, Shahid Beheshti University of Medical Sciences, [3]Department of Medical Genetics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, [4]Department of Engineering, Sabzevar University of New Technologies, Sabzevar, Iran

Abstract

Background: Mass spectrometry is a method for identifying proteins and could be used for distinguishing between proteins in healthy and nonhealthy samples. This study was conducted using mass spectrometry data of ovarian cancer with high resolution. Usually, diagnostic and monitoring tests are done according to sensitivity and specificity rates; thus, the aim of this study is to compare mass spectrometry of healthy and cancerous samples in order to find a set of biomarkers or indicators with a reasonable sensitivity and specificity rates. **Methods:** Therefore, combination methods were used for choosing the optimum feature set as t-test, entropy, Bhattacharya, and an imperialist competitive algorithm with K-nearest neighbors classifier. The resulting feature from each method was feed to the C5 decision tree with 10-fold cross-validation to classify data. **Results:** The most important variables using this method were identified and a set of rules were extracted. Similar to most frequent features, repetitive patterns were not obtained; the generalized rule induction method was used to identify the repetitive patterns. **Conclusion:** Finally, the resulting features were introduced as biomarkers and compared with other studies. It was found that the resulting features were very similar to other studies. In the case of the classifier, higher sensitivity and specificity rates with a lower number of features were achieved when compared with other studies.

Keywords: *Biomarker discovery, imperialist competitive algorithm, mass spectrometry high-throughput proteomics data, ovarian cancer*

Submitted: 22-Mar-2020 **Revised:** 14-May-2020 **Accepted:** 04-Jul-2020 **Published:** 24-May-2021

Introduction

One of the major and unresolved problems in the treatment of cancer disease is the lack of an appropriate method for its timely and early diagnosis. Based on information on the genetic science area, chemical reactions within a living organism might be reflected as protein patterns in fluids such as urine and blood.[1] Recent results of researchers indicate that the pattern of proteins in the blood could be regarded as a fingerprint for the disease.[2] Thus, comparing the proteins in healthy and patient samples might result in the detection of vital biomarkers and the determination of the position of the cell during the disease process. Detecting the reliable biomarkers can help in early diagnosis of disease and its treatment, since in most cases, molecular variations related to them are diagnosable before clinical signs of the diseases emerge. One of the techniques used for extracting the protein patterns is surface-enhanced laser desorption/ionization of time-of-flight mass spectrometry (SELDI-TOF MS).[3] This method, in combination with advanced data mining algorithms, is used to reveal the protein patterns related to diseases.[4,5] In fact, the mass spectrometry yields a signal, which its horizontal axis is the mass-to-load (M/Z) ratio of the specific molecules in Dalton and its vertical axis is severity as a criterion of the frequency of the molecules in the sample.[6] Tang *et al*.[7] used statistical moments for the reduction of the dimension of the characteristics after using a *t*-test on a dataset of ovarian cancer with high resolution. Dataset was categorized using the kernel partial least squares (KPLS) method. The M/Z ratio range is up to about 20,000 Daltons, leading to the generation of a vector of 5000–20,000 numerical values per mass spectrum.[7] While the mass spectrum has high dimensions, the number of healthy and patient samples in the data is relatively low.

Address for correspondence:
Dr. Niloofar Yousefi Moteghaed, Department of Biomedical Engineering and Medical Physics, Faculty of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
E-mail: nilofar.yousefi@gmail.com

How to cite this article: Pirhadi S, Maghooli K, Moteghaed NY, Garshasbi M, Mousavirad SJ. Biomarker discovery by imperialist competitive algorithm in mass spectrometry data for ovarian cancer prediction. J Med Sign Sens 2021;11:108-19.

Access this article online

Website: www.jmssjournal.net

DOI: 10.4103/jmss.JMSS_20_20

Quick Response Code:

Hence, data mining techniques to reduce the dimension, determine the biomarkers, and correct classification of the samples have high importance. Hilario and Kalousis[8] examined various methods for reducing the dimension in the mass spectrometry dataset. The reduction of the dimension in this type of dataset would reduce tens of thousands of variables (M/Z points) to several hundred variables. The reduction of the dimension is generally divided into two groups: feature selection and feature transformation. Feature transformation methods result in linear or nonlinear combinations of the features. These methods can be single variable or multivariable, depending on the fact that one single feature or a subset of features is evaluated. In the concept of classification, feature selection methods are divided into three groups including filter, wrapper, and embedded.[9] Cancer screening and diagnostic tests are evaluated in terms of the rate of sensitivity and specificity. Several studies have been carried out in order to classify or detect biomarkers in the dataset of mass spectrum related to cancer. Using the dimension reduction methods and different classifications has resulted in different sensitivity and specificity rates. Petricoin *et al*.[10] investigated protein patterns generated by SELDI-MS, to distinguish between healthy and ovarian cancer samples. This analysis was conducted in the form of a combination of a genetic algorithm and a self-organizing map. Li *et al*.[11] used both *t*-test and genetic algorithms to select the features, and in both methods, the support vector machine (SVM) method was used as a classifier. Yu *et al*.[12] used a strategy including four stages of binning, a Kolmogorov–Smirnov test, limiting unstable coefficients, wavelet analysis, and SVM classifier. Liu[13] used a multilevel wavelet analysis for the mass spectrum dataset and presented approximation coefficients as input to the SVM classifier. Conrads *et al*.[14] used mass spectrum dataset SELDI with high and low resolutions to diagnose ovarian cancer. Using the binning method and implementing the combination of genetic algorithm and self-organizing systems resulted in the separation of control and cancerous samples with high sensitivity and specificity.

The M/Z vectors for all samples were homogenized using resampling to be able to compare various spectra based on the same resolution and reference. The act of resampling, in fact, implements an anti-alias filter to delete the high-frequency noise available in the mass spectrum.[15] Mass spectrometry with 15,000 M/Z points was used, where values between 710 and 11,900 M/Z were obtained by resampling of the spectra. All the M/Z points were considered as a subset of features. After this stage, mass spectra were normalized to delete the effect of the scale coefficient. In the normalization operation, the mean and variance of each feature for all the samples were obtained. Then, the values of each feature were subtracted from the mean, and the result was divided by the corresponding variance of that feature. Lack of balance between the number of features and samples might result in an increased chance of the wrong classification, due to the use of nonrelevant or additional features. From a medical viewpoint, finding a limited number of markers that play a major role in correct diagnosis has special importance. Thus, reducing the number of features and selecting a number of them seem to be essential. If feature selection is done independently of any learning algorithm and based on a ranking criterion, it would be called a filter method. However, if the evaluation procedure is followed by a classification algorithm, the feature selection method would be called the wrapper method. This method uses the search in the space of the subsets based on the estimation of the accuracy resulting from the selection of special subset under classification algorithm conditions. Researchers have focused on evolutionary search algorithms such as genetic algorithm (GA),[16] simulated annealing (SA),[17-19] particle swarm optimization,[20,21] and cultural algorithm[22] over the past decade.

Wu *et al*.[23] used the Kolmogorov–Smirnov method, logistic regression, and random forest as a feature selection in an ovarian cancer dataset with high resolution. Dataset was classified using one hundred superior features and three classifiers of SVM, regression tree with bagging, and K-nearest neighbors (KNN). The objective of the current research is not just to differentiate the mass spectrum dataset of ovarian cancer and to achieve high sensitivity and specificity rates but also to detect a number of biomarkers for diagnosis of this cancer. The proposed algorithm is a combination of a filter method based on single-variable feature selection and a multivariable wrapper method, which are evaluated in high-resolution ovarian cancer dataset. Using three methods: *t*-test, entropy, and Bhattacharyya, the features were reduced and the output of each of these three methods to the imperialist competitive algorithm (ICA) was provided, to select the optimal subset of the features using the KNN classifier. Then, the results obtained from three combined methods were provided separately to the C5.0 decision tree algorithm to determine the most important features such as biomarkers and extract the rules to distinguish the cancer data from healthy data.

Subjects and Methods

Data preprocessing

In this research, we have used the National Cancer Institute and have gathered two high- and low-resolution datasets for ovarian cancer data.[11]

We have used the combination of different methods for data preprocessing step. In the *t*-test method, M/Z values are ranked by the absolute value of the test statistic, assuming that M/Z value is independent and calculating a two-way *t*-test. In the entropy method, the resolution level of two classes, or ranking of the features, is performed using the entropy criterion. The Bhattacharyya method as the ranking criterion measures the similarity between two probability distributions.

Imperialist competitive algorithm

The ICA in the area of evolutionary computations provides a method to solve optimization problems by mathematical modeling of the sociopolitical evolution process of humans. This algorithm was introduced in 2007,[24] and it has been used so far to solve many problems in the area of optimization.[25-37] Like other evolutionary algorithms, this algorithm is composed of the initial set of possible solutions, which of them is called a country. The ICA gradually improves these initial solutions (countries) and finally provides the desired answer to the optimization problem (the desired country). The main bases of this algorithm are assimilation, imperialist competition, and revolution. As stated, the ICA begins with a number of initial random populations, called a country. Some of the best elements are selected as imperialists, and the rest of the population is considered a colony. Imperialists draw these colonies toward themselves in a special procedure based on their power. The total power of each empire depends on its two parts constituting it, namely the imperialist country (as the core) and its colonies. In the mathematical state, this dependency was modeled by defining the empire power as the sum of the power of the imperialist power and a percentage of mean power of its colonies. By the formation of initial empires, the imperial competition begins among them. Any empire that cannot act successfully in the imperialist competition and cannot increase its power will be eliminated from the imperialist competition arena. Thus, the survival of the empire depends on its power in attracting the empire colonies of the competitor and dominating them. Therefore, during the imperialist competitions, the power of the larger empires will increase gradually and the weaker empires will be eliminated. In order to increase their power, empires will have to develop their own colonies. With the passage of time, colonies will become closer to empires in terms of power and a kind of convergence will be formed. The ultimate limit of the imperialist competition is when there is a single empire in the world. Figure 1 illustrates the flowchart of the ICA.

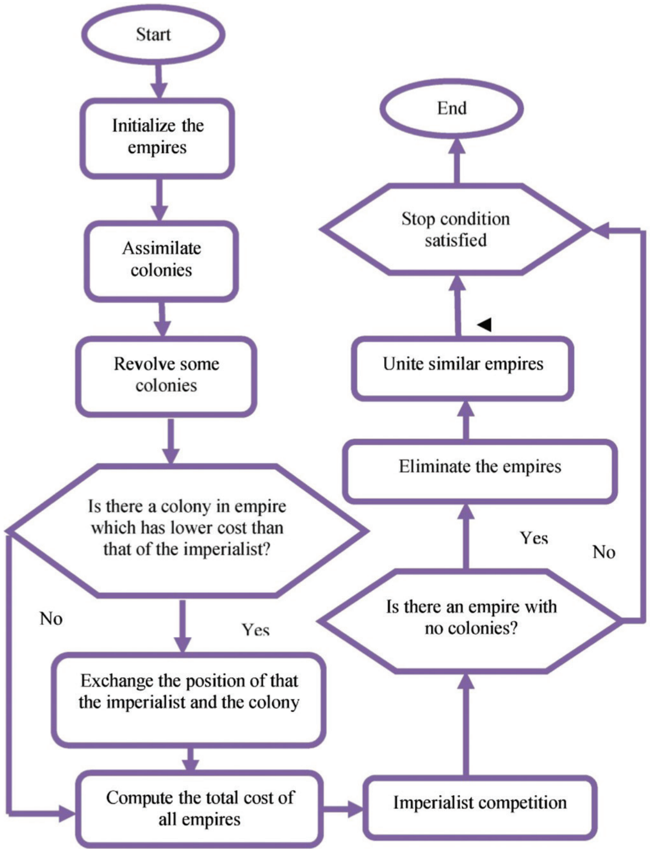
Start
Initialize the empires
Assimilate colonies
Revolve some colonies
Is there a colony in empire which has lower cost than that of the imperialist?
No
Yes
Exchange the position of that the imperialist and the colony
Compute the total cost of all empires
Imperialist competition
Is there an empire with no colonies?
Yes
No
Eliminate the empires
Unite similar empires
Stop condition satisfied
End

Figure 1: Imperialist competitive algorithm flowchart[38]

The objective of using this optimization algorithm is for the process of selecting the features in a way that the most optimal set of features is found. Each country is coded in a binary way. It means that the initial populations are random strings of the numbers zero and one, with the length of the number of the features.. If the feature has been selected from the set of features , it has been shown by number ‘1’ in each string, and when the number is ‘0’, it means that feature has not been selected. In the ICA, each of these strings is called a country. The solution to attraction policy is that the new position includes both previous information and a part of imperialist information. Thus, there is a need to recognize where the position of the colony and imperialist is similar and where it is different. As the problem binary is defined, both of them are strings of zero and one. As a result, if the colony position is reduced from the imperialist position and the absolute value from the resulting string is calculated, the result will be a string of zero or one. Being zero in this string means that both imperialist and colony were similar in that cell (both were zero or one) and being 1 means that one is zero and the other one is one. With this, the difference between the two strings is revealed. By summing up the elements of the string resulting from the deduction of the colony and imperialist, the number of these differences is also revealed. If this number is multiplied by a number between zero and one (for example, 0.2) and the obtained number is rounded up, it can be considered as a criterion to make different cells similar. For example, suppose that the number of variables has been determined to be 200. If colony and imperial are different from each other in 120 out of 200 cells, the 120 in 0.2 is multiplied, and the result will be 24. Hence, the similarity in 24 out of all the cells in which two strings differ should be created. Hence, the colony position in those cells is made exactly like its imperialist position in the corresponding cell (that is, if the imperialist is zero in that cell, we make its colony zero, and if it is one, we make it one). Using this approach, the new colony position would be a position between colony position and its imperialist position. For the revolution operator, an integer random number between

1 and the number of features was generated. This number specifies which cell should be changed in the position of each colony in each empire. Then, it is replaced with zero or one, generated randomly.

Application of the algorithm to ovarian cancer datasets with high resolution

As stated earlier, this dataset has 121 cancerous cases and 95 control samples, and the number of M/Z points reached 15,154 using resampling operations. Eighty-five percent of the samples were considered as the training sample, 7% as the validation samples, and the rest were considered as the test samples. Thus, the numbers of training, validation, and test samples are 184, 16, and 16, respectively. After normalizing the data and applying the filter algorithms (*t*-test, entropy, and Bhattacharyya), in which each of them ranked the features by their own criterion, those features were extracted from the main data. KNN was selected as a classifier. Thus, when the cost function was called up, KNN learning was performed using training samples and tested by the validation samples. In each run, the best cost and mean for different countries were determined.

Finally, this set was tested on the test samples, where the algorithm has not been seen before. Moreover, the value of K in the KNN classifier was selected to be 2, and the distance type was selected as Euclidean. After implementing the ICA several times and calculating the mean of accuracy, sensitivity, and specificity, we provided the answers obtained from each time of implementation to algorithms of the association rule mining and the decision tree, and the rules obtained were presented. Before applying the association rule mining algorithm, there is a need to determine the level of support, the coefficient of confidence, the maximum number of features in the antecedent of each rule, and the maximum number of rules.

The flowchart of the different stages of the proposed algorithm is illustrated in Figure 2. It should be noted that MATLAB 2017a (The Mathworks, Inc., Natick, MA, USA) software and SPSS Clementine 12 (SPSS Inc., Chicago, IL, USA) have been used for the application of algorithms.

Figure 3a shows the range of the M/Z values, several healthy and cancerous spectra, and values of the absolute value obtained from the t-test. Figure 3b illustrates the entropy values, and Figure 3c illustrates the Bhattacharyya criterion together with the spectra of two healthy and cancerous groups.

Results

In this section, the number of features extracted from each filter method and the settings related to ICA [Table 1] were investigated.

When the ICA begins, cost function recalled in each decade and the cost of all countries is calculated using cost function and the best and mean costs are yielded. These values can be reported at each time of implementing ICA, which proceeds based on the number of decades determined. Figure 4 shows these values for the three phases of ICA implementation (these charts refer to the state, where the *t*- and ICA tests were used): Figure 4a related to the first implementation, Figure 4b related to the second implementation, and Figure 4c related to the third implementation of the ICA. The horizontal axis also indicates various decades. As shown, in the three implementations, the best cost in each decade reached 100 and the mean of the costs gradually reached 100. It means that the algorithm could optimize the cost function by a set of selected features and increase the accuracy of the classification of the samples to 100%.

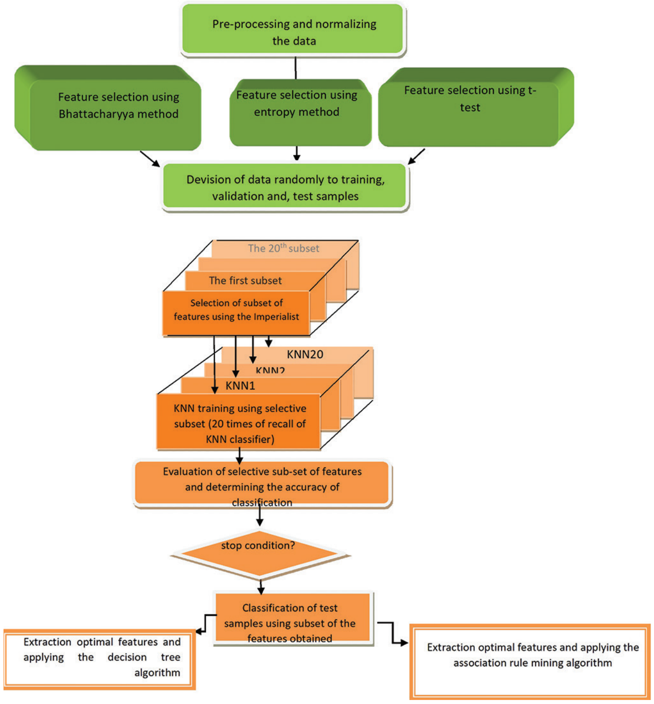
Pre-processing and normalizing the data
Feature selection using Bhattacharyya method
Feature selection using entropy method
Feature selection using t-test
Devision of data randomly to training, validation and, test samples
The 20th subset
The first subset
Selection of subset of features using the Imperialist
KNN20
KNN2
KNN1
KNN training using selective subset (20 times of recall of KNN classifier)
Evaluation of selective sub-set of features and determining the accuracy of classification
stop condition?
Classification of test samples using subset of the features obtained
Extraction optimal features and applying the decision tree algorithm
Extraction optimal features and applying the association rule mining algorithm

Figure 2: Flowchart of the proposed algorithm

The accuracy of the two methods is also shown in Table 2. It should be noted that ICA proceeds with the number of the mentioned decades each time of the implementation, and finally, it selects a number of features among the features selected from the filter methods. Then, this set of features selected by ICA is stored in the last decade and tested on the test samples. Accordingly, the values of accuracy, sensitivity, and specificity were calculated. Then, ICA was run again. Then, it selects another set of features and tests on the test samples. As shown in Table 2, this was performed three times. Finally, the values of accuracy, sensitivity, and specificity obtained from the test samples were averaged in three stages. The result of the average of these values is also shown in Table 3.

With each time of implementation of ICA, the selected features were stored by the algorithm, and those repeated in each of the three times were determined. Table 2 shows information related to the number of features obtained

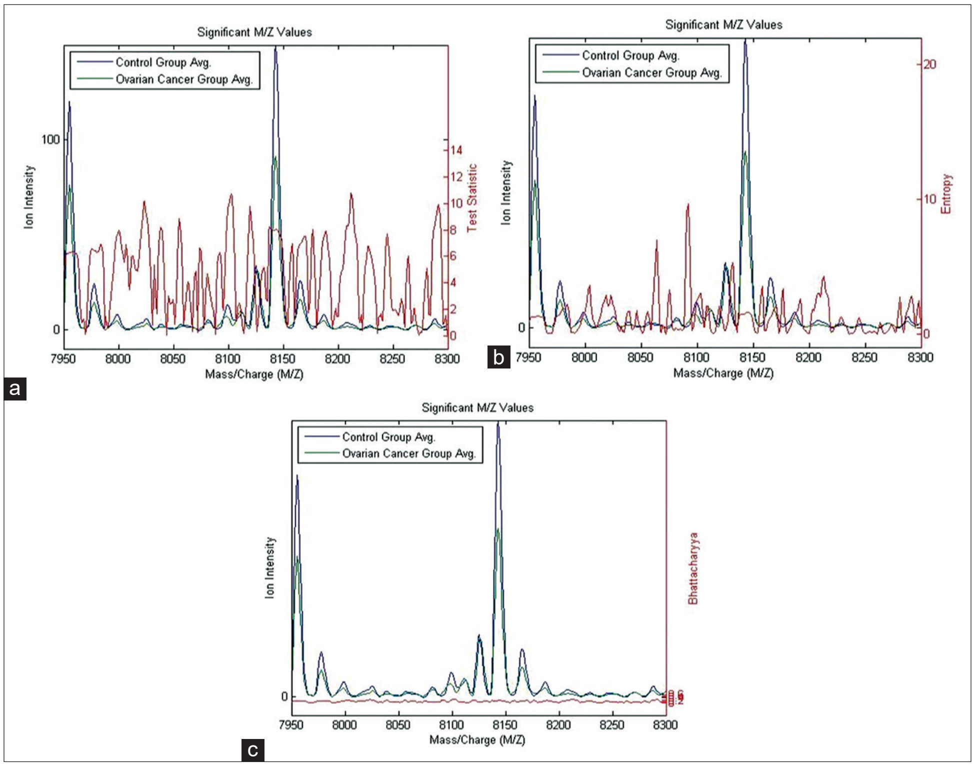

Figure 3: Control and cancerous spectra together with the absolute value (a) *t*-test; (b) entropy values; (c) the values related to Bhattacharyya distance

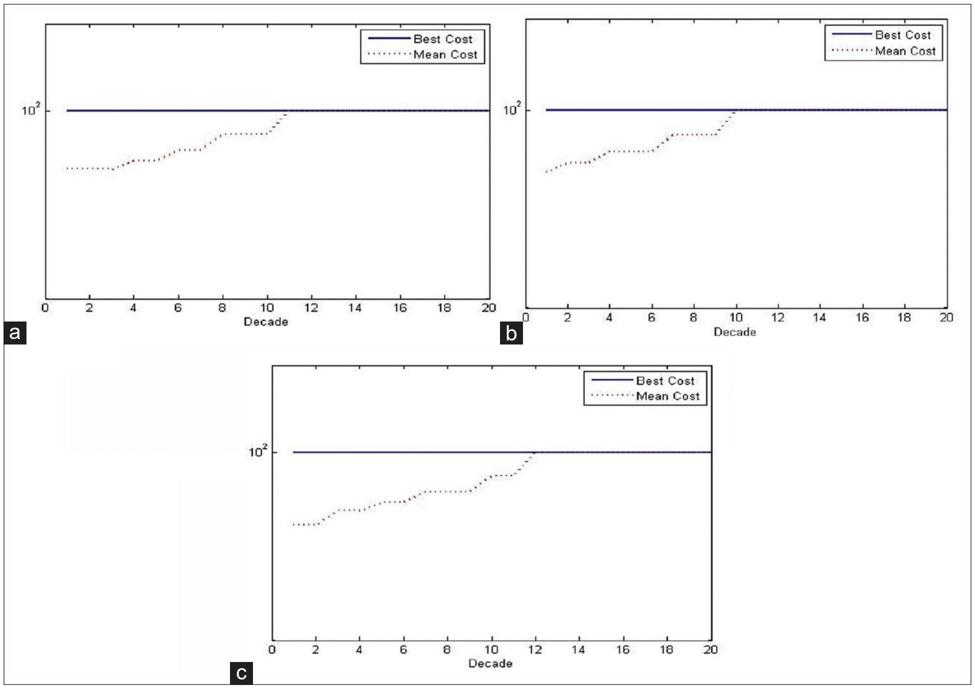

Figure 4: Values of the cost function and meaning of the costs in the implementation of (a) first, (b) second, and (c) third. The horizontal axis indicates the decade, the longest line shows the best cost, and the dotted line shows the mean of costs

from each method in different implementations. Finally, the features which have been repeated in at least two methods (*t*-test and ICA, entropy and ICA, and Bhattacharyya and ICA) as the most frequent biomarkers were reported. Out of the 31, 28, and 44 most frequent features obtained from all three methods, there are five common features. The M/Z values of these 5 features include 861/094, 862/7076, 3428/817, 7053/731, and 7054/654.

Then, the features obtained from the first, second, and third implementation in each method separately were extracted

Table 1: Imperialist competitive algorithm parameters in filter-based methods

Methods	Number of features	Parameters of ICA					
		nPop	nImp	Decades	B	pRevolution	ζ
t-test	200	30	8	20	0.3	0.3	0.1
Entropy	200	30	5	20	0.1	0.1	0.1
Bhattacharyya	400	40	4	15	0.1	0.1	0.1

ICA – Imperialist competitive algorithm

Table 2: Information related to the number of selected characteristics in each time of imperialist competitive algorithm implementation, and finally, the number of common characteristics implemented in each time, three times

Methods	#features from 1st running of ICA	#features from 2nd running of ICA	#features from 3d running of ICA	Common features in tree times of running ICA
t-test-ICA	104	113	111	31
Entropy-ICA	114	97	98	28
Bhattacharyya-ICA	199	194	201	44

ICA – Imperialist competitive algorithm

Table 3: Results obtained from imperialist competitive algorithm implementation on ovarian cancer dataset

Methods	#decade	Best cost	Mean cost	#empires
t-test-ICA	1	100	96.77	8
	20	100	100	4
	1	100	96.77	8
	20	100	100	4
	1	100	96	8
	20	100	100	3
Entropy-ICA	1	100	100	5
	20	100	100	5
	1	100	100	5
	20	100	100	5
	1	100	100	5
	20	100	100	4
Bhattacharyya-ICA	1	100	100	4
	15	100	100	3
	1	100	100	4
	15	100	100	4
	1	100	100	4
	15	100	100	4

ICA – Imperialist competitive algorithm

from the normalized data. Accordingly, there are three matrices with the number of main data samples for each method, and the number of features in each matrix is also equal to the values shown in columns 2–4 of Table 2. The C5 decision tree algorithm was applied separately for each of the matrices of the dataset. By implementing C5, the most important variables were determined and a number of rules were extracted. In Table 4, the accuracy of classification using C5, the most important variable, and the number of rules obtained for the control and cancerous groups were determined. In Table 4, there are features repeated several times among the most important features. M/Z values repeated in at least two methods are 845/042, 8607/152, 7065/738, 1006/425, 7063/890, 8708/407, and 8603/073. Rules extracted by the C5 algorithm, where these common features are present, are listed in Table 5.

In addition to the decision tree, another algorithm known as generalized rule induction (GRI) was used in this study to extract the rules. The algorithm requires determining the support parameter, the confidence coefficient, and the maximum number of features in the antecedent part of a rule, which should be determined by the user. These parameters 30, 100, and 10, respectively, were determined. In addition, these rules can be arranged using the methods with the scoring mechanisms. The most important of these approaches is a method in which the rules are arranged first based on the confidence coefficient in the descending order, then, the rules having the same confidence coefficient are arranged based on the level of support, and if the support of a number of rules is the same, they are arranged based on the number of antecedent features.[39] The association rules obtained from the three methods (*t*-test-ICA, entropy-ICA, and Bhattacharyya-ICA) were compared with the same method. The common rules are listed in Table 6.

Accordingly, the repetitions of each feature were not just considered separately as the criterion to select the biomarker, each occurrence of the features or finding of frequent patterns were examined.

Now, these biomarkers in the control and cancerous samples were investigated and change in their severity in the two mentioned groups was found. In this regard, we can plot the spectra related to both groups and compare the severity of the M/Z values. Another method is the display of the heatmap, which is an effective method for visualizing the complex dataset in the matrices. In the heatmap display, the areas where there is peak were determined by hot colors, and other areas were determined by cold areas.

The common biomarkers provided by the C5 method in the mean of the control and cancerous samples are shown by the red triangle in Figure 5. The heatmap of the cancerous

Table 4: Results obtained from applying the decision tree on ovarian cancer dataset

Important values of M/Z	#rules for cancer group	#rules for control group	Accuracy% C5	#features of original data
1034/163, 8607/152, 7063/890, 8708/408	2	3	98.15	104
845/042, 8711/485, 8607/152, 7065/738, 8713/531	2	4	97.22	113
1036/285, 8022/877, 8100/848, 1072/332, 1006/425, 8604/093, 7065/738, 844/722	2	5	99.54	111
6856/641, 8794/786, 8212/033, 845/042, 1290/083, 8607/152, 4310/126, 8553/187	3	4	98.15	114
8794/786, 8603/073, 6834/813, 8213/03, 4310/126, 1056/195	3	4	99.07	97
8600/015, 8794/786, 8621/435, 845/042, 4310/126, 4003/314	3	4	98.61	98
1006/774, 8025/832, 845/042, 8710/459, 8603/073, 7065/738	3	5	99.07	199
6859/372, 1006/425, 8708/408, 1939/120, 8604/093, 7065/738, 6850/271, 1079/182	3	4	99.07	194
8522/717, 7063/890, 8607/152, 7174/257, 1049/775	3	3	99.54	201

Table 5: Rules obtained from C5 algorithm related to ovarian cancer dataset

1. If 7065/738≤0.073 and 8603/073≤−0.034 then control
2. If 1006/774>0.018 and 8603/073≤−0.034 then control
3. If 1006/425≤0.028 and 7065/738>0.073 then cancer
4. If 845/042>−0.013 and 8607/152≤−0.038 then control
5. If 8603/073≤−0.034 then control
6. If 8708/408>0.078 then control
7. If 8607/152≤−0.038 then control
8. If 7063/890>−0.037 and 8607/152>−0.038 and 8708/408≤0.078 then cancer
9. If 845/042>0.045 then control

group is plotted at the top and that of the control group is plotted at the bottom.

Now, the high frequent biomarkers, which are common in different implementations of the *t*-test-ICA, entropy-ICA, and Bhattacharyya-ICA methods, within the spectra of healthy and cancerous groups were determined. The severity of the spectra of two groups [Figure 6] can be compared by limiting the M/Z values around each of these biomarkers. In this figure, the normal group is shown in red color, and the cancerous group is shown in blue color.

As shown in Figure 7, the severity of 861/094 and 862/7076 values is high in normal samples; in low cancerous samples, the severity of 3428/817, 7053/731, and 7054/654 is high in many cancerous samples, and it is reduced in normal samples. Figure 7 shows the biomarkers (their average is plotted instead of their spectra).

The severity of these biomarkers also varies in the average of healthy and cancerous spectra. However, these M/Z values are shown in Figure 7 alone and as common features among the most important features obtained from C5. Considering the rules in Table 5, they have at least one feature, and if their severity is more or less than value, the normal or cancerous class will be separable using them.

In the antecedent part of the rules obtained from the GRI algorithm, the features were mentioned and can be compared in the average of both healthy and cancerous spectra. Some of these values are observable in the figures for previous biomarkers. Thus, only 1031/516, 4300/749, 4302/912, 4310/126, and 8618/373 are plotted in Figure 8.

Then, we aimed to compare the biomarkers provided by our proposed algorithm and other studies to find if there is a similarity between them or not. For this purpose, studies that provided results on ovarian cancer datasets with high resolution were referred to. Conrads *et al*.[12] collected samples from controls and those who had ovarian cancer and used two mass spectrometers: one with low resolution and another with high resolution. In this study, the biomarkers obtained from the algorithm express the values 845/089, 8602/237, and 8709/548. In 2008, Wu[21] used his algorithm on two samples of ovarian cancer data, in which one of them was the same ovarian cancer dataset with high resolution in relation to the previous study.

We have found three common M/Z values with Wu[21] and Conrads *et al*.[12] as 845/042, 8708/407, and 8603/073. Another important point is that one of these values, 845, was the same in the three studies. This value of M/Z among our three methods, *t*-test-ICA-C5, entropy-ICA-C5, and Bhattacharyya-ICA-C5, was shown to be the same. Thus, it can be regarded as a biomarker with high confidence. Figure 9 shows this M/Z value in the heatmap related to the average of the control and cancer groups. The severity of this biomarker is plotted for all cancerous and control samples [Figure 10]. In terms of data classification, several studies that implemented their algorithm on this data can be compared.

Discussion

The proposed algorithm, which is a combination of filter techniques (*t*-test, entropy, and Bhattacharyya) and ICA, was applied to the ovarian cancer mass spectrum data with high resolution. Each of the filter methods, in fact, ranked the features based on their own criteria. A number of the best features obtained from each method were selected and separately provided to the ICA. ICA, as an intelligent optimization method, was associated with our desired classifier (KNN) during the training process, to achieve the most optimal set of features after implementing several decades by maximizing the classification accuracy. Finally, a set of features was provided by ICA for each of the methods: *t*-test-ICA, entropy-ICA, and Bhattacharyya-ICA. In this regard, there were features repeated in these methods, which were reported as the most frequent biomarkers. In the next stage, the features obtained from the three methods were provided to C5 decision tree, so that the most important features of each method can be extracted using the criterion of this algorithm. In this regard, there were common features, which were also determined. The rules containing these common features were provided for a better understanding. In addition to the C5 tree decision, another algorithm called GRI was used to yield repetitive patterns. Finally, the biomarkers introduced in the main data were investigated to determine the severity of each of them in the control or cancer groups. The severity of some of these biomarkers in the cancer state increased as compared to that of the normal state, while the severity of others reduced.

Conclusion

By comparing the values contained in Table 7, it was realized that the proposed algorithm in all the methods with the lower number of characteristics could achieve an acceptable level of sensitivity and specificity. Another advantage is using M/Z values in the data as characteristic features, in which a number of biomarkers were also presented as a result. These biomarkers showed a high similarity with the biomarkers reported. Moreover, by examining the classifications in the table, it is shown that the KNN, as a simple classifier in our algorithm, could achieve appropriate answers with the lower number of the features, while stronger classifiers such as SVM and KPLS achieved this rate of sensitivity and specificity with the greater number of the features.

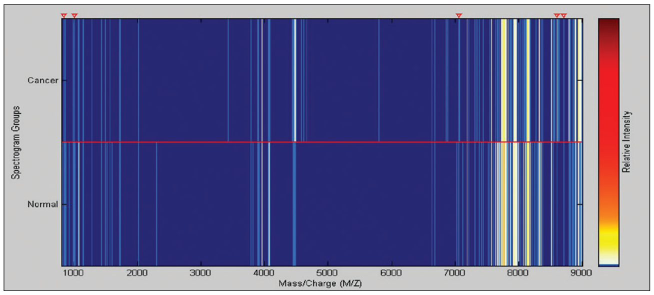

Figure 5: Heatmap display for the average of the cancerous groups (top) and control (bottom). Biomarkers obtained from C5 method are shown with the red triangle

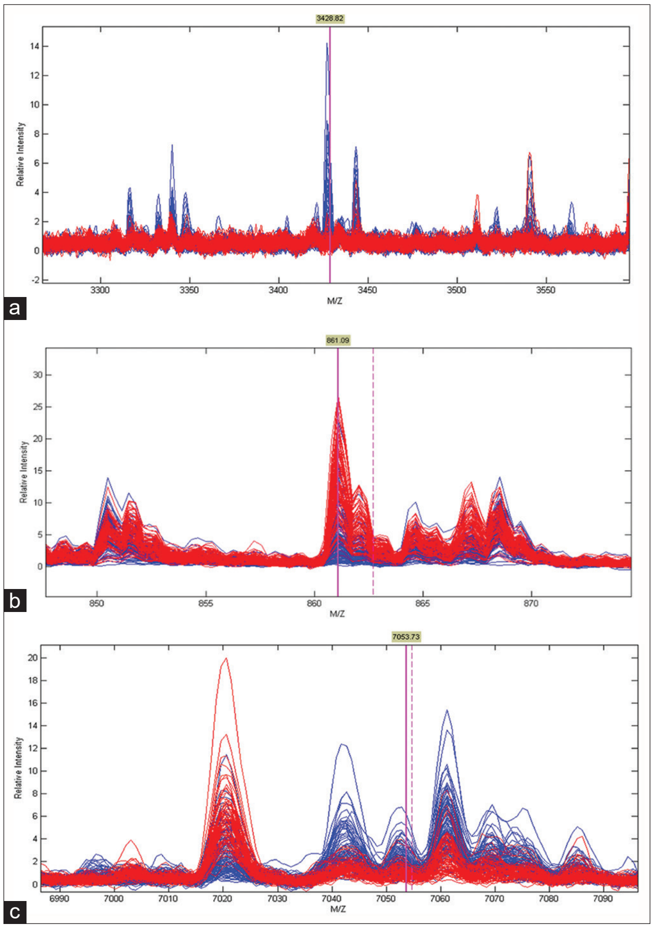

Figure 6: Display of M/Z values for high-frequent biomarkers. (a) 861/094 and 862/076; (b) 3428/817; (c) 7053/731 and 7054/654. The horizontal axis shows the M/Z values and the vertical axis shows their severity

Table 6: Rules obtained from the exploration of generalized rule induction associative rules in the ovarian cancer dataset

Rules	Percentage of confidence coefficient	Percentage of support
If 7065/738> −0.032 and 8601/034> −0.022 then cancer	100	43.06
If 1034/516 < 0.002 and 4310/126> −0.021 then cancer	100	42.59
If 7065/738> −0.032 and 8602/053> −0.021 then cancer	100	42.59
If 7065/738> −0.032 and 1078/821 < 0.021 and 4303/634> −0.025 then cancer	100	40.74
If 8618/373> −0.017 and 4302/912> −0.011 then cancer	100	39.81
If 4300/49 > −0.020 and 4302/912> −0.008 then cancer	100	39.81

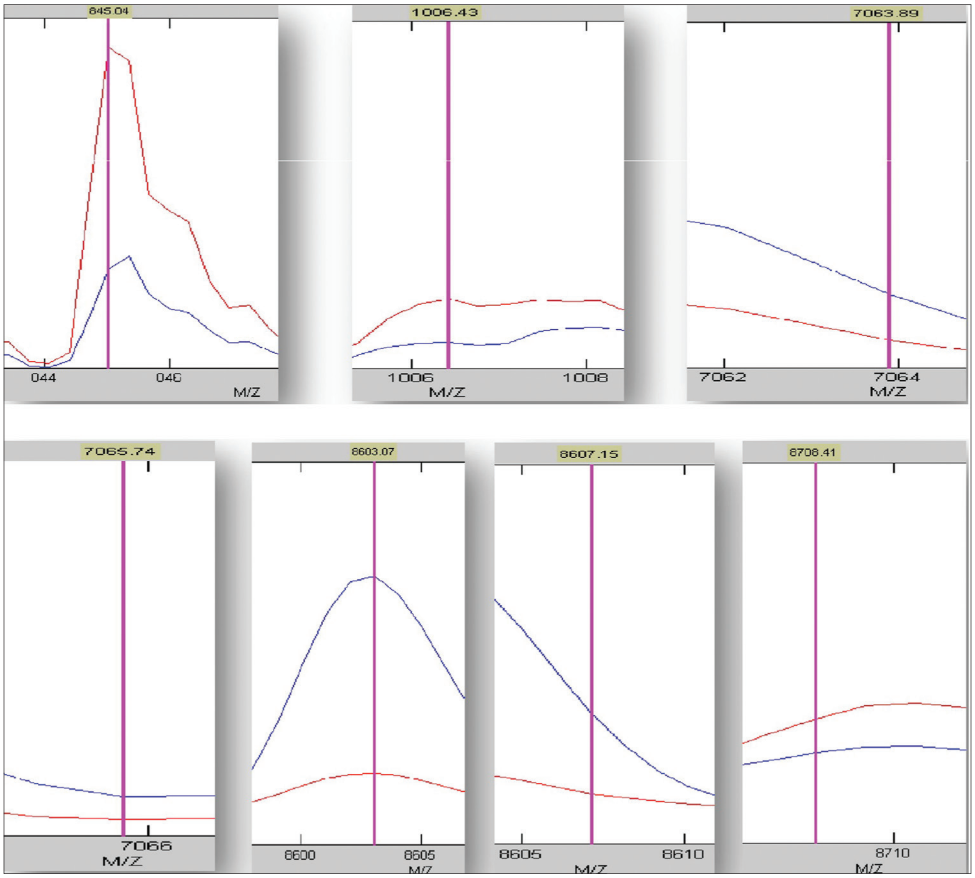

Figure 7: Biomarkers obtained after the implementation of the C5 algorithm. The top row from left to right: 845/04, 1006/43, and 7063/89 and the bottom row from left to right/7065/74, 8603/07, 8607/15, and 8708/41. In this figure, the average control and cancerous samples are plotted with red and blue colors, respectively

Table 7: Comparison of sensitivity and specificity in several studies related to ovarian cancer data with high resolution with our proposed algorithms

Authors/year/ classification type	Cross validation	Using the main features	Number of features	Sensitivity (%)	Specificity (%)
Yu *et al*.[12]	1000 independent k-fold (k=2,…10)	No	3382	97.38	93.30
Wu *et al*.[23]	10-fold	Yes	100	93.9	93.23
Tang *et al*.[7]	5-fold	No	1964	99.50	99.16
Liu[13]	2-fold	No	247-949	98.45-99.55	95.69-97.01
Wu[40]	10-fold	No	215	92.98	88.97
Cui[41]	10-fold	Yes	371	98.16	-
Our proposed algorithm as *t*-test-ICA-C5 KNN	10-fold	Yes	104 113 111	97.52 96.69 100	98.94 97.89 98.94
Our proposed algorithm as Entropy-ICA-C5 KNN	10-fold	Yes	114 97 98	98.34 98.34 99.17	97.89 100% 99.89
Our proposed algorithm as Bhattacharyya-ICA-C5 KNN	10-fold	Yes	199 194 201	99.17 100 99.17	98.94 97.89 100

ICA – Imperialist competitive algorithm

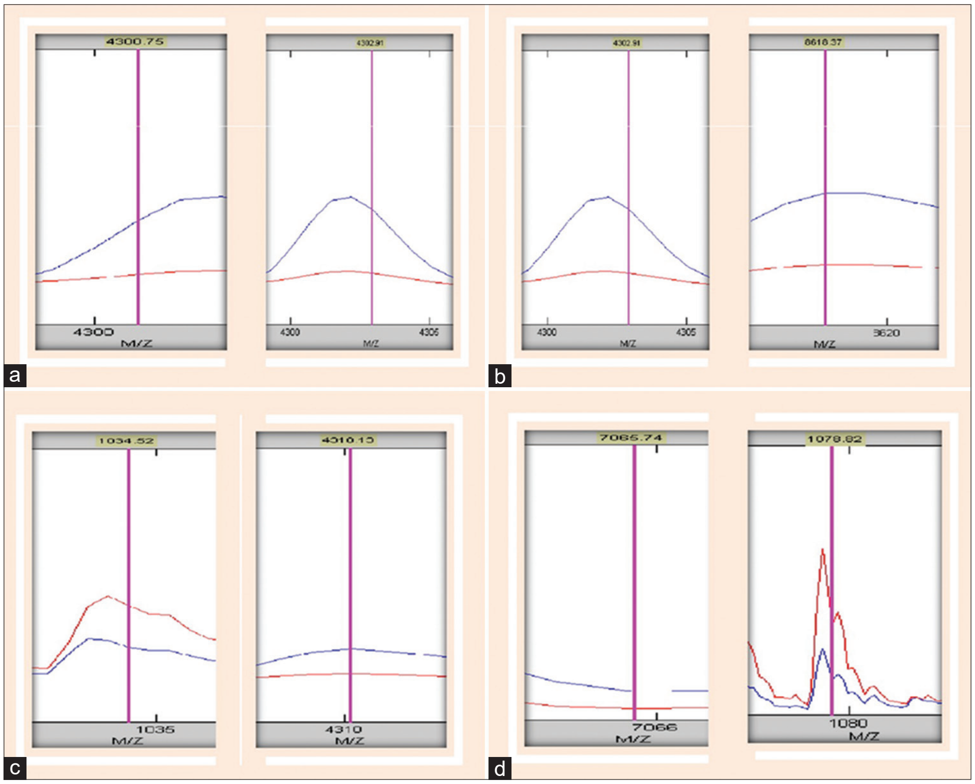

Figure 8: M/Z values in the antecedent part of some of the associative rules (the value for all these rules is the cancer class): (a) 4302/912 and 4300/749; (b) 8618/373 and 4302/912; (c) 4310/126 and 1034/516; (d) 1078/821 and 7065/738

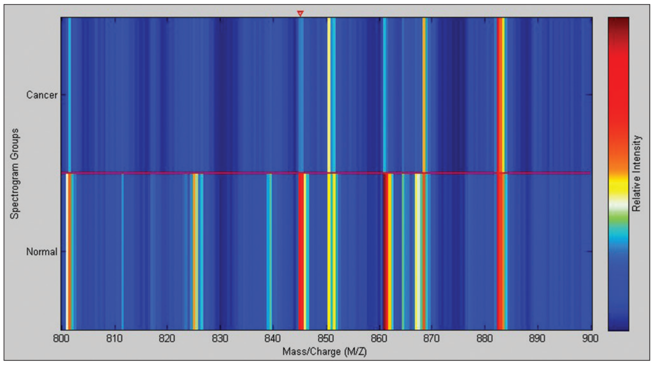

Figure 9: Display of 842/042 value in the average cancerous spectra (high) and control (low). As shown, the severity of this biomarker is greatly different in cancerous and healthy samples

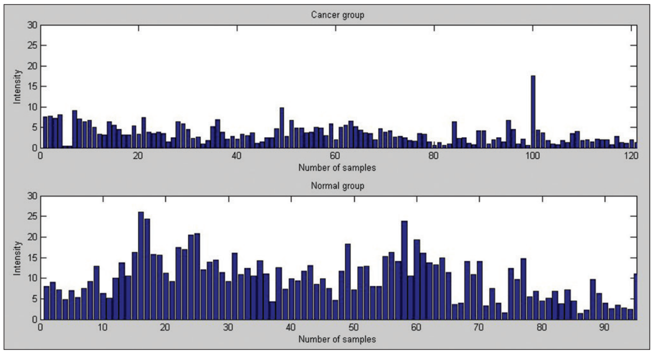

Figure 10: M/Z = 845/042 severity in all cancerous (high) and control (low) samples

Financial support and sponsorship

None.

Conflicts of interest

There are no conflicts of interest.

References

1. Srinivas PR, Srivastava S, Hanash S, Wright GL. Proteomics in early detection of cancer. Clin Chem 2001;47:1901-11.
2. Price ND, Edelman LB, Lee I, Yoo H, Hwang D, Carlson G, *et al*. Systems biology and systems medicine. In: Essentials of Genomic and Personalized Medicine. 2010. p. 131-41.
3. Li Y, Zeng X. Serum SELDI-TOF MS analysis model applied to benign and malignant ovarian tumor identification. Analytical Methods 2016;8:183-8.
4. Srinivas PR, Verma M, Zhao Y, Srivastava S. Proteomics for cancer biomarker discovery. Clin Chem 2002;48:1160-9.
5. Fan Z, Kong F, Zhou Y, Chen Y, Dai Y. Intelligence algorithms for protein classification by mass spectrometry. Biomed Res Int 2018;2018.
6. Assareh A, Moradi MH. Extracting efficient fuzzy if-then rules from mass spectra of blood samples to early diagnosis of ovarian cancer. In: Computational Intelligence and Bioinformatics and Computational Biology, 2007. p. 502-6.
7. Tang KL, Li TH, Xiong WW, Chen K. Ovarian cancer classification based on dimensionality reduction for SELDI-TOF data. BMC Bioinformatics 2010;11:109.
8. Hilario M, Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. Brief Bioinform 2008;9:102-18.
9. Wagner M, Naik D, Pothen A. Protocols for disease classification

from mass spectrometry data. Proteomics 2003;3:1692-8.

10. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, *et al*. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572-7.
11. Li L, Tang H, Wu Z, Gong J, Gruidl M, Zou J, *et al*. Data mining techniques for cancer detection using serum proteomic profiling. Artif Intell Med 2004;32:71-83.
12. Yu JS, Ongarello S, Fiedler R, Chen XW, Toffolo G, Cobelli C, *et al*. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. Bioinformatics 2005;21:2200-9.
13. Liu Y. Dimensionality reduction and main component extraction of mass spectrometry cancer data. Knowledge Based Sys 2012;26:207-15.
14. Conrads TP, Fusaro VA, Ross S, Johann D, Rajapakse V, Hitt BA, *et al*. High-resolution serum proteomic features for ovarian cancer detection. Endocr Relat Cancer 2004;11:163-78.
15. IEEE Acoustics, Speech, Signal Processing Society. Digital Signal Processing Committee. Programs for Digital Signal Processing. IEEE; 1979.
16. Mühlenbein H, Schomisch M, Born J. The parallel genetic algorithm as function optimizer. Parallel Computing 1991;17:619-32.
17. Tokuda I, Aihara K, Nagashima T. Adaptive annealing for chaotic optimization. Phys Rev E 1998;58:5157.
18. Ingber L. Simulated annealing: Practice versus theory. Mathemat Comp Modelling 1993;18:29-57.
19. Cardoso MF, Salcedo RL, de Azevedo SF, Barbosa D. A simulated annealing approach to the solution of MINLP problems. Comp Chem Eng 1997;21:1349-64.
20. Eberhart R, Kennedy J. A new optimizer using particle swarm theory. In: Micro Machine and Human Science; 1995. MHS'95. Proceedings of the Sixth International Symposium on 1995. p. 39-43.
21. Yang X, Yuan J, Yuan J, Mao H. A modified particle swarm optimizer with dynamic adaptation. Appl Mathem Comp 2007;189:1205-13.
22. Franklin B, Bergerman M. Cultural Algorithms: Concepts and Experiments. In: Evolutionary Computation. Vol. 2. Proceedings of the 2000 Congress on 2000. p. 1245-51.
23. Wu LC, Chen TP, Horng JT. Analysis of High-Resolution Protein Mass Spectra Based on Peak Feature Selection. Medical & Biological Engineering and Computing; 2008.
24. Atashpaz-Gargari E, Lucas C. Imperialist Competitive Algorithm: An Algorithm for Optimization Inspired by Imperialistic Competition. In: Evolutionary Computation, IEEE Congress on; 2007. p. 4661-7.
25. Talatahari S, Azar BF, Sheikholeslami R, Gandomi AH. Imperialist competitive algorithm combined with chaos for global optimization. Communicat Nonl Sci Num Simul 2012;17:1312-9.
26. Kaveh A, Talatahari S. Optimum design of skeletal structures using imperialist competitive algorithm. Comput Struct 2010;88:1220-9.
27. Lucas C, Nasiri-Gheidari Z, Tootoonchian F. Application of an imperialist competitive algorithm to the design of a linear induction motor. Energy Conversion Manag 2010;51:1407-11.
28. Khabbazi A, Atashpaz-Gargari E, Lucas C. Imperialist competitive algorithm for minimum bit error rate beamforming. Int J Bio Inspired Comput 2009;1:125-33.
29. Shabani H, Vahidi B, Ebrahimpour M. A robust PID controller based on imperialist competitive algorithm for load-frequency control of power systems. ISA Trans 2013;52:88-95.
30. Coelho LD, Afonso LD, Alotto P. A modified imperialist competitive algorithm for optimization in electromagnetics. IEEE Trans Magnetics 2012;48:579-82.
31. Abdechiri M, Faez K, Bahrami H. Neural network learning based on chaotic imperialist competitive algorithm. In: Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on; 2010. p. 1-5.
32. Mousavi SM, Tavakkoli-Moghaddam R, Vahdani B, Hashemi H, Sanjari MJ. A new support vector model-based imperialist competitive algorithm for time estimation in new product development projects. Rob Comp Int Manufact 2013;29:157-68.
33. Maroufmashat A, Sayedin F, Khavas SS. An imperialist competitive algorithm approach for multi-objective optimization of direct coupling photovoltaic-electrolyzer systems. Int J Hydrog Energy 2014;39:18743-57.
34. Rad SM, Tab FA, Mollazade K. Application of imperialist competitive algorithm for feature selection: A case study on bulk rice classification. Int J Comput Appl 2012;40:41-8.
35. Mojaveriyan M, Ebrahimpour-komleh H, Jalaleddin Mousavirad S. IGICA: A hybrid feature selection approach in text categorization. Int J Intellig Syst Appl 2016;8:42.
36. Wang S, Tang Z, Gao S, Todo Y. Improved Binary Imperialist Competition Algorithm for Feature Selection from Gene Expression Data. In: International Conference on Intelligent Computing. Springer, Cham; 2016. p. 67-78.
37. Wang S, Aorigele , Kong W, Zeng W, Hong X. Hybrid Binary Imperialist Competition Algorithm and Tabu Search Approach for Feature Selection Using Gene Expression Data. Biomed Res Int 2016;2016.
38. Mousavirad SJ, Ebrahimpour-Komleh H. Feature selection using modified imperialist competitive algorithm. In: Computer and Knowledge Engineering (ICCKE), 2013 3rd International eConference on. IEEE; 2013. p. 400-5.
39. Wang YJ, Xin Q, Coenen F. Hybrid rule ordering in classification association rule mining. Transactions on Machine Learning and Data Mining. 2008;1:1-6.
40. Wu J, Ji Y, Zhao L, Ji M, Ye Z, Li S. A mass spectrometric analysis method based on PPCA and SVM for early detection of ovarian cancer. Comput Math Methods Med 2016;2016.
41. Cui L, Ge L, Gan H, Liu X, Zhang Y. Ovarian cancer identification based on feature weighting for high-throughput mass spectrometry data. J Syst Biol 2018;1:1.

BIOGRAPHIES

Shiva Pirhadi received B.Sc. and M.Sc degree in biomedical engineering from the Science and Research branch of Islamic Azad university University, Tehran, Iran in 2010 and 2012 respectively. Now she is Ph.D student in biomedical engineering in Science and Research branch of Azad University since 2012. She works in the area of bioinformatics, data mining and medical image processing.

Email: shv_prhd@gmail.com

Keivan Maghooli has received his B.Sc. in electronic engineering from the Shahid Beheshti University, Tehran, Iran, M.Sc. in biomedical engineering from the Tarbiat Modaress University, Tehran, Iran, and Ph.D. in biomedical engineering from the Research and Science branch, Azad University, Tehran, Iran, majoring in Data Mining, Signal Processing and Artificial Intelligence. He has been with the Biomedical Faculty at Research and Science branch, Azad University, Tehran, Iran, since 2000, where he is currently an Assistance of Professor and Head of Bioelectric department.

Email: k_maghooli@srbiau.ac.ir

Niloofar Yousefi Moteghaed received B.Sc. and M.Sc degree in biomedical engineering from the Science and research branch of Islamic Azad university University, and Ph.D degree in biomedical engineering in Shahid Beheshti University of Medical Sciences and Health ,Tehran in 2010 , 2012,and 2019 respectively. She works in the area of bioinformatics, data science , deep learning, and medical image processing.

Email: nilofar.yosefi@gmail.com

Masoud Garshasbi has received his B.Sc. in Biology from the Ferdowsi University, Mashhad, Iran (2001), and his M.Sc. in Human Genetics from the University of Social Welfare and Rehabilitation (USWR), Tehran, Iran (2003).He obtained his Ph.D. (2009) and Post doctoral (2011) in Human Molecular Genetics from Max Planck Institute, Berlin, Germany by working on the genes involved in Mental retardation. At 2011 he joined as an assistant professor to the Department of Medical Genetics, Faculty of Medical Sciences, Tarbiat Modares, Tehran, Iran. He is also founder and head of Medical Genetic Department at the DNA laboratory, Tehran, Iran.

Email: masoud.garshasbi@gmail.com

Seyed Jalaleddin Mousavirad received the M.Sc. degree (Hons.) in computer engineering from Kurdistan University, Sanandaj, Iran. He is currently independent consultant and researcher on optimization, machine learning, and image processing with an outstanding research record. He has published more than 60 papers in reputable academic journals and conference proceedings. He is an active reviewer for more than 10 international conference and journal papers.

Email: Jalalmoosavirad@gmail.com