A New Approach Based Symbiotic Organism Search using Data Mining for Medical Decision

Samia Nouredine, Abida Toumi, Baarir Zineeddine
Electrical Engineering Department, University of Biskra
{Samia.nouredine@univ-biskra.dz} {zineeddine.baarir@univ-biskra.dz}
(a.toumi@univ-biskra.dz)

ABSTRACT: In the medical field the decision represents an extremely important asset because the risk must be zero. This is why decision approaches which are based on a predictive vision are a must solution. The decision is generally based on the exploitation of a large volume of medical data. The processing and analysis of mass data is only possible through an extraction of knowledge allowing the medical experts to make the best decision. Thus to meet this need, data mining has become the most promising approach. There are several techniques of datamining, and although they are quite developed they still remain even less efficient notably the classical meta-heuristics. In this paper, we are exploiting a new meta-heuristic called symbiotic organisms search (SOS) that is based on a biological process. In this paper, we develop the formal model of the SOS based data mining process in the medical field with a comparative study with other metaheuristics to show its performance and credibility of treatment.

Subject Categories and Descriptors:
[G.1.6 Optimization]; [H.2.8 Database Applications]; Data mining; [J.3Medical information systems]

General Terms: Decision Making, Medical Information System, Data Mining

Keywords: Symbiotic, Symbiosis, Datamining, Metaheuristic, Optimization, Classification

Received: 16 February 2020, Revised 12 July 2020, Accepted 27 July 2020

Review Metrics: Review Scale: 0/6, Review Score: 5.05, Inter-reviewer consistency: 84%

DOI: 10.6025/jdim/2020/18/5-6/195-209

1. Introduction

Nowadays, scientists and engineers find a refuge of research solutions in combinatorial optimization, the interest in this form of problem representation lies in the difficulties encountered in solving these problems in an exact manner in addition to its massive use in several fields of engineering. These areas concern robotics, telecommunications, industrial safety, prediction of industrial machine behavior, medical decision support etc. For such systems the expectations and results are of great importance in terms of saving time, money, energy, or satisfaction [1]. In these areas of application, the exact solution is sometimes difficult or impossible to achieve. For this purpose meta-heuristics have become an alternative helping researchers to have an acceptable and sufficient solution [2]. Metaheuristics are generally of social or biological inspiration based on the collective behavior of organisms living together and generating a phenomenon that can be modeled by a mathematical system [1]. The main limitations of metaheuristic statistics concern above all their non-determinism and do not guarantee optimality. Another limitation also concerns the inability to precisely define the form of the solutions and also how to generate these solutions (admissible or ineligible) [2].

The main goal of this paper consists on one hand to evaluate the performance of SOS by using test functions, and to see its efficiency and its behavior through its use for data mining for classification or supervised clustering applied on public medical data bases, on the other hand, the process of data mining is a phase of knowledge extraction for decision support. It consists of representing in a very expressive way the data extracted from a raw database ensuring analysis and a decision-making rather thoughtful by the decision-maker of a company. In the
literature, several techniques have been worked on for datamining [3].

SOS is applied in different fields for many data treatments, and we are interested in the medical field and classification. In 2016, SOS was used for the training of the FNNs for the classification of different databases [4]. Prinolan and Absalom used SOS to solve the problem of blood allocation [5] and the problem of blood assignment [6].

The paper is organized as follow: in the first section we present the state of the art about combinatory optimization, section two is dedicated to four used metaheuristics in this research work. In section three, we develop SOS specification and its experimentations using six fitness functions followed by last section containing classification based on datamining for medical area.

2. Combinatorial Optimization

Automatic problem solving has invaded many areas of engineering and our social life. Certain types of problems are difficult to solve by an algorithmic approach because solution modeling does not follow a mathematical principle that can be modeled by formal methods. Therefore, researchers have turned to this type of problem by necessity and for the sake of finding a solution to combinatorial optimization problems [1]. This type of problems are sometimes easy to define, but they are usually difficult to solve.

Combinatorial optimization is a topical field at the cutting edge of combinatorics and theoretical computing that aims to use combinatorial techniques to solve discrete optimization problems. A discrete optimization problem seeks to determine the best possible solution from a finite set of possibilities [1]. Combinatorial optimization problems occur in a multitude of real-world applications, such as routing, assignment, scheduling, splitting and packaging, network design, protein alignment, and more areas of economic, industrial and scientific importance. The techniques available for these problems can be broadly classified into two main categories: the exact and metaheuristic methods (approximate methods) [1].

The exact resolution methods aim at obtaining a solution whose optimality is guaranteed. In some cases, it is more interesting to find better quality solutions with no guarantee of optimality in favor of a shorter calculation time [1] [2]. For this purpose, it is advisable to use methods called metaheuristic, adapted to each problem treated, with the disadvantage of not having in return any information on the quality of the solutions obtained [1].

Heuristics or meta-heuristics generally exploit random processes in the exploration of the research space to cope with the combinatorial explosion engendered by the use of exact methods.

Metaheuristics are based on neighborhood methods such as simulated annealing and taboo research, and evolutionary algorithms such as genetic algorithms and evolution strategies. These metaheuristics have the advantage of generating approximate solutions for large conventional optimization problems. During these last years a great application focuses on the metaheuristics more especially in the field of operational research but especially in artificial intelligence [2] [1].

The Artificial Beeh Colon (ABC) algorithm was proposed by Karaboga in 2005 to solve constrained optimization problems [7].

The cultural algorithm (CA) has two basic components: the space of the population and the space of belief, it is a system with double heredity. Its principle is based on the idea that the experiences of selected individuals are then added to the content of the belief space using an update function [8].

The Gray Wolf Optimization (GWO) is an intelligent approach to the swarm proposed by Mirjalili et al in 2014. This approach is inspired by hunting behavior with a social hierarchy of gray wolves [9].

The Chaotic Dynamic Weight Particle Swarm Optimization algorithm (CDWPSO): In the objective of improving the efficiency of the PSO and to make its two mechanisms of exploration and exploitation, a possible solution is in the form of a chaotic optimization of the particle swarm in using the idea of dynamic weight [10].

Genetic algorithms (GA) are optimization algorithms based on the mechanisms of natural selection and genetics. They were adapted to the optimization by John [11], another researcher in this case David Goldberg who worked hard to improve the formal and functional aspect [12]. The inspiration for this technique is based on the theory of evolution and genetics. This is why, we use the concept of individual (potential solution), population (set of solutions), genotype (a representation of the solution), gene (part of the genotype), parent, child, reproduction, crossover, mutation, generation, etc.

The differential evolution algorithm (DE) is a branch of scalable programming developed by Rainer Storn and Kenneth Price for optimization problems on continuous domains. In DE, the value of each variable is represented by a real number. The advantages of DE are its simple structure, ease of use, speed and robustness [13].

Particle Swarm Optimization (PSO) is a stochastic optimization method for nonlinear functions based on the reproduction of social behavior. The origin of this method comes from observations made during computer simulations of group flights of birds and schools of fish. These simulations highlighted the ability of individuals in a moving group to maintain an optimal distance from each other and to follow a global movement in relation to local
movements in their neighborhood [14]. Particles are individuals and they move in the hyperspace of research based on limited information.

3. Symbiotic Organisms Search (SOS) Algorithm

SOS is a new metaheuristic algorithm developed by [15] in 2014. It is bio-inspired by the dependence based on the lasting association between two or more ecosystem organisms, and profitable to each of them, each of them deriving a benefit from this association.

SOS is like most metaheuristic population-based algorithms, it has the following characteristics [15] [16]:

- This algorithm proposes candidate solutions that constitute a population of organisms dispersed in the search space;
- Candidate solutions are the key elements of the operators for the orientation of research;
- The algorithm exploits a selective strategy to achieve the best solutions;
- This selectivity requires a mechanism that validates intrinsically the intrinsic characteristics, namely the size of the population and the maximum number of evaluations.

However, unlike most metaheuristic algorithms that have control parameters (for example, GA has the crossover and mutation rate; DE has crossover and mutation vectors as well; the PSO is characterized by specific aspects same as inertia weight, cognitive factors, and social factors), [15]. SOS does not require any specific parameters to the algorithm. This is considered an advantage over competing algorithms because SOS does not need to perform parameter adjustment [15]. The inaccurate setting of algorithm-specific parameters can increase the compute time and result in local optimal solutions [16]. SOS process contains three phases: mutualism, commensalism and parasitism.

3.1 Mutualism Phase

In this step, the organisms help each other to ensure an existence in the ecosystem. The survival process is based on a change of state of each of the two candidate organisms $X_i$ and $X_j$ expressed by the following expressions [15]:

$$X_i_{\text{new}} = X_i + \lambda (X_{\text{best}} - \text{Mutual\_Vector} \times BF1) \quad (1)$$

$$X_j_{\text{new}} = X_j + \beta (X_{\text{best}} - \text{Mutual\_Vector} \times BF2) \quad (2)$$

$$\text{Mutual\_Vector} = (X_i + X_j)/2 \quad (3)$$

The new value of the gain factor of both organisms are $BF1$ and $BF2$, they are calculated randomly as 1 or 2. Each factor is the partial or full gain level of each organism.

The best level of adaptation of the organism is represented by $X_{\text{best}}$, $\lambda$ and $\beta$ are random values in $[0, 1]$. The symbiotic relationship, also called characteristic, between the $X_i$ and $X_j$ organisms, is represented by a vector called “Mutual\_Vector”.

3.2 Commensalism Phase

In this phase, one of the two organisms obtains a gain where the other go without gaining or losing something. The organism $X_j$ represents the one that does not gain or lose something of the relation and the new organism $X_i$ is calculated according to the symbiosis of commensalism between the organisms $X_i$ and $X_j$. Its expression is given as follows [15]:

$$X_i_{\text{new}} = X_i - \alpha (X_{\text{best}} - X_j) \quad (4)$$

Where $\alpha$ represents a random number in $[-1, 1]$ and $X_{\text{best}}$ is the highest degree of adaptation organism.

3.3 Parasitism Phase

Parasitism is a behavior where one organism obtains gain from one relationship but actively harms the other organism. An artificial parasite called “Parasite\_vector” is created in the search space using duplication of the organism $X_i$ and then modifying the dimensions concerned randomly by a random number. This Parasite\_vector tries to replace another organism $X_j$ in the ecosystem. For this goal, if Parasite\_vector is better, he will kill the organism $X_j$ and assume its position; otherwise, $X_j$ will be immune to the parasite and can no longer exist in this ecosystem [15].

The three phases are formalized by the following algorithm:

Initialization (ecosystem, fitness function and maximum iteration)

For counter=1: maximum iteration

For counter=1: size (ecosystem)

Mutualism Phase

Commensalism Phase

Parasitism Phase

Identify better organisms

End For

End For

Figure 1. SOS algorithm pseudocode

4. Experimentations and Discussion

In this section, the SOS algorithm is executed using six
we known benchmark functions [17], then we experiment SOS by four real and public medical databases for the classifications. Test functions, were chosen according following three categories:

• **Category 1:** Unimodal functions with high dimension: function Schwelef 2.22 (F1), function Schwelef 2.22 (F2), with 30 dimensions.

• **Category 2:** Multimodal functions with high dimension: function Rastrigin (F3), function Ackley (F4) with 30 dimensions.

• **Category 3:** Multimodal functions with low dimension: functions on Six hum (F5) with 4 dimensions, function Six hum camel back (F6) with 2 dimensions.

The classification datasets used in this experimentation are: breast cancer, heartDisease, liver disorders (LD) and Indian Diabetes Pima (PID) obtained from the machine learning repository of the University of California at Irvine (UCI) [18]. For an excellent estimation and appreciation performance of SOS, we compared it with GA, PSO, and DE. The other comparative measures presented in the results section are: matrix confusion, classification accuracy rates, and ROC curves to evaluate our results.

### 4.1 Test Functions

In table.1 we list the different functions used for the test process. This function list allowed us to verify the robustness and performance of the proposed algorithmic solution concerning symbiotic metaheuristic. We have chosen these functions because they are complex and difficult to obtain their global optimum because of the existence of several local optima.

Three dimensions functions curves are presented in figure 1 this type of curves gives us clearly the number of local optima for each one.

<table>
<thead>
<tr>
<th>Category</th>
<th>Functions of Test</th>
<th>Dimension</th>
<th>Interval</th>
<th>Maximum Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unimodal functions with high dimensions</td>
<td>$F_{01} = \sum_{i=1}^{n}</td>
<td>x_i</td>
<td>+ \prod_{i=1}^{n}</td>
<td>x_i</td>
</tr>
<tr>
<td></td>
<td>$F_{02} = \max_{1 \leq i \leq D}</td>
<td>x_i</td>
<td>$</td>
<td>30</td>
</tr>
<tr>
<td>Multimodal functions with high dimensions</td>
<td>$F_{03} = \sum_{i=1}^{D} x_i^2 - 10 \cos(2\pi x_i) + 10$</td>
<td>30</td>
<td>[-5.12,5.12]</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>$F_{04} = -20 \exp \left(-0.2 \sqrt{\frac{1}{D} \sum_{i=1}^{D} x_i^2} \right) - \exp \left(\frac{1}{D} \sum_{i=1}^{D} \cos(2\pi x_i) \right) + 20 + e$</td>
<td>30</td>
<td>[-32,32]</td>
<td>60</td>
</tr>
<tr>
<td>Small multimodal functions</td>
<td>$F_{05} = \sum_{i=1}^{4} \left[ a_i - \frac{x_i (b_i^2 + b_i x_i)}{b_i^2 + b_1 x_4 + x_4} \right]^2$</td>
<td>4</td>
<td>[-5,5]</td>
<td>400</td>
</tr>
<tr>
<td>Small multimodal functions</td>
<td>$F_{06} = [1 + (x_1 + x_2 + 1)^2 \left( 19 - 14x_1 - 3x_1^2 - 14x_2^2 + 6x_1x_2 + 3x_2^2 \right) \times \left[ 30 + (2x_1 - 3x_2)^2 \right] \times \left[ 18 - 32x_1^2 + 12x_1^3 + 48x_2^2 - 36x_1x_2^2 + 27x_2^4 \right] ]$</td>
<td>2</td>
<td>[-5,5]</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 1. List of test functions used [11] [13]
4.2 Comparative Study
To have a good appreciation of proposed approach, we compared the symbiotic method (SOS) with other methods: PSO, ABC, CA, GWO and CDWPSO. These results are obtained after 25 iterations and shown in Table 2. For the evaluation of each function, we used three variables which are the mean, the standard deviation and the ranking of each method.

This ranking is obtained on the basis of the best average of the values of the objective functions and the best standard deviations with respect to the optimal value of the three categories, the SOS method give F1 to F6 effective results compared to the other methods.

4.3 Fitness Curve
To show more performance of SOS metaheuristic, we present the curves of the fitness functions. Figure 2 shows the convergence curves obtained by four algorithms on six functions.

<table>
<thead>
<tr>
<th>Functions</th>
<th>Optimum</th>
<th>PSO</th>
<th>ABC</th>
<th>CA</th>
<th>GWO</th>
<th>CD-PSO</th>
<th>SOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>0</td>
<td>0.0352</td>
<td>0.1752</td>
<td>59.2268</td>
<td>8.0399e-47</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.1624</td>
<td>0.0990</td>
<td>20.9776</td>
<td>8.3453e-47</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>0</td>
<td>2.1999</td>
<td>12.3781</td>
<td>3.1395</td>
<td>1.3831e-21</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1.7124</td>
<td>1.6148</td>
<td>3.2499</td>
<td>2.8568e-21</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td>0</td>
<td>275.2287</td>
<td>308.3371</td>
<td>316.9494</td>
<td>47.6089</td>
<td>2.3177</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>17.2402</td>
<td>20.9358</td>
<td>18.6367</td>
<td>15.1017</td>
<td>8.2990</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>F4</td>
<td>8.8818e-16</td>
<td>13.4428</td>
<td>16.4168</td>
<td>11.1855</td>
<td>0.0844</td>
<td>0.9060</td>
<td>8.8818e-16</td>
</tr>
<tr>
<td></td>
<td>0.7387</td>
<td>0.7822</td>
<td>0.8800</td>
<td>0.0292</td>
<td>1.6656</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>F5</td>
<td>0.0003075</td>
<td>0.0036</td>
<td>1.0010</td>
<td>0.0044</td>
<td>0.0019</td>
<td>5.96e-04</td>
<td>3.4411e-04</td>
</tr>
<tr>
<td></td>
<td>0.0075</td>
<td>3.3967e-05</td>
<td>0.0122</td>
<td>0.0056</td>
<td>6.14e-04</td>
<td>1.8314e-04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>F6</td>
<td>3</td>
<td>3.0000</td>
<td>3.001</td>
<td>3.0000</td>
<td>3.0000</td>
<td>8.1731</td>
<td>3.0000</td>
</tr>
<tr>
<td></td>
<td>1.8965e-012</td>
<td>1.85137e-04</td>
<td>1.3628 e-15</td>
<td>2.1140e-05</td>
<td>5.3154</td>
<td>0.5368e-16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Middle Class</td>
<td>4.16</td>
<td>5</td>
<td>4.83</td>
<td>3.16</td>
<td>2.33</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Overall ranking</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Methods evaluation results
The values shown in these figures represent a comparison of the fitness results of the six functions according to SOS, PSO, ABC, CA, GWO and CDW-PSO. From these figures, it is clear that our method has a better search accuracy than the other three methods; this figure also shows that the SOS algorithm has a faster convergence speed. This proves that SOS has a strong robustness and a good stability ensuring a clear convergence compared to the other methods.

4.4 ANOVA Variation Analysis

To explain our evaluation of the proposed method compared to other comparison algorithms, we used in Figure 3, the “ANOVA test” which is a schematic representation of the distribution of a variable. The Anova test invented in 1977 by John Tukey, also called Tukey box or box-and-whisker plot, is used to represent the essential scheme of a quantitative statistical series [19]. The Anova test which is performed on the six test functions gives the result of 25 independent execution of our method. These results show that SOS approach is more suitable for large-scale problems.

The shapes in the figures obtained by this test, which are boxes, represent the variations of the optimal solutions between the minimum and maximum values, centered on the average, which is the optimal solution of the test subject function.

We notice that the SOS, for the functions from F1 to F6, is very stable that proves that it generates a good performance with a clear precision.
4.5 SOS Evaluation for Datamining

There are many pieces of information circulating in the information industry in general and in any medical, economic or administrative sector. These data have no cognitive or economic value until they are converted into useful information. It is interesting to see the need to analyze this large volume of data and to extract useful and usable information [20].

Datamining is the process of retrieving information from huge data stores. In other words, datamining is also defined as data mining to extract knowledge [21].

After checking the robustness of SOS, in what follows we will make an evaluation of the symbiotic for the automatic clustering mechanism in datamining.

The experiments carried out with the SOS algorithm aim to classify in a supervised manner. For a qualitative evaluation, the same experiments are performed on the GA,
DE and PSO algorithms on the same dataset.

4.6 The Fitness Function
The quality of the results obtained depends directly on the formal specification of the objective function. The study of the existing functions gave us the inspiration of a fitness function having as fundamental parameter the minimal distance between the instances and the centers of classes proposed randomly. The formula of this fitness function is as follows [22] [23]:

\[ F = \sum_{i=1}^{n} \sum_{j=1}^{k} (d_{\text{min}}(X_i, c_k)) \]  

\( d_{\text{min}} \): minimum distance between each instance and the class centers.

\( X \): data matrix.

\( c_k \): the class centers.

\( k \): number of classes.

\( n \): number of instances of the database.

4.7 The Databases used for Datamining
To show that SOS is excellent technic for datamining, we have uses the same four database cited previously and detailed as follow:

**Breast Cancer (BC) Data Base**
This database was prepared by William H. Wolberby of the University of Wisconsin Hospitals, Madison. It contains 683 instances and 11 attributes such as: clump thickness, uniformity of cell size, uniformity of cell shape marginal adhesion.

**Heart Disease (HD) Data Base**
This database was created for the processing and decision analysis of single proton computer tomography (SPECT) images. This database contains 267 images, 80 instances and 22 entities. All functionalities are in binary (0 or 1), after the treatment the diagnosis declares if the patient is normal or not.

**Liver Disorder (LD) Data Base**
The database on liver disorders has been provided by BUPA Medical Research Ltd for liver disorders in the form of a binary label. It contains values of 7 attributes measured for 345 male patients. The first five represent blood test data considered sensitive to liver disorders due to excessive alcohol consumption.

**Pima Indian Diabetes (PID) Data Base**
This database concerns pregnant women aged at least 21 years old and of Indian origin. Pima contains a number of instances equal to 768 with a number of attributes equal to 8.

Table 3 summarizes the database used for calcification. We note that the data belongs to several levels of complexity. We note that only the breast cancer database is relatively simple compared to others.

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of instances (n)</th>
<th>Number of attributes</th>
<th>Number of Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer (BC)</td>
<td>683</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Heart Disease (HD)</td>
<td>80</td>
<td>23</td>
<td>2</td>
</tr>
<tr>
<td>Liver Disorder (LD)</td>
<td>345</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Diabetes (PID)</td>
<td>768</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3. The databases used [18]

![Figure 4. Classification process of Diabetes (PID) Data base](image-url)
obtained, the classification is performed. Because SOS has three phases, so it has automatically tree solutions in each iteration, it produces a matrix of solutions (6), the fitness function choose the best solution between them [23]. Figure 4 explains the classification process.

\[ X = \{x_1, x_2, \ldots, x_{it}\} \]  

such that \( x_i \) is a matrix solution and \( it \) is a number of maximum iteration.

4.9 Classification Results

The results were obtained after 200 iterations of the four algorithms. To show more detail results, we use the fitness curves and classification histograms. For more estimation of the results, the validation is performed for the four methods by the confusion matrix and the ROC curves.

4.9.1 Curves of the Fitness function

In this section, we evaluated the fitness function for the four algorithms on the four data bases previously described. The goal of this evaluation is to show SOS convergence comparatively with other methods.

Figures 5a, 5b, 5c and 5d represent the variations of the fitness function. It is clear, according to these curves, that the SOS algorithm is more efficient and performs faster converges than the other methods. This proves the good quality of the SOS solution evaluation compared to the other used metaheuristics.

4.9.2 Histogram Classification

In this section, we applied the four algorithms for classification task by using the four databases to discover if each algorithm add each instance to the appropriate class. To show the summary with data analysis we use the following table. 4.
The number of instances in each class

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Class1</th>
<th>Class2</th>
<th>Class1</th>
<th>Class2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>n = 444 benign</td>
<td>n = 239 malignant</td>
<td>n = 450</td>
<td>n = 450</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>n = 23 abnormal</td>
<td>n = 54 normal</td>
<td>n = 24</td>
<td>n = 27</td>
</tr>
<tr>
<td>Liver Disorder</td>
<td>n = 145 not</td>
<td>n = 200 disorder</td>
<td>n = 272</td>
<td>n = 268</td>
</tr>
<tr>
<td>Diabetes (PID)</td>
<td>n = 500 not infected</td>
<td>n = 273</td>
<td>n = 546</td>
<td>n = 273</td>
</tr>
</tbody>
</table>

Table 4. Distribution of the numbers of instances of each class by the four algorithms

In the following schema, we represent the classified data in the form of histograms to give a significant visibility in order to make a good comparison between the methods.

Discussion
From the results presented on the various classification histograms above, it is clear that (see Table 4), the SOS histogram classification is almost identical to the real result, which means that SOS has given a very good classification because just few elements are misclassified, whereas the other methods generate more misclassified numbers.
4.9.3 Validation of Classification using Confusion Matrix

To evaluate the quality of SOS classification and the results of other approaches, the confusion matrices (CM) is used to clarify obtained results. The diagonal of each CM gives the distribution of instances in the corresponding class [24].

**The Accuracy Rate (AR) or accuracy**: deduced from the CM, it represent the set of observations of good classification, it is defined by the following equation [24]:

$$AR = \frac{\sum_{j=1}^{n} \text{diag}(CM_j)}{\sum CM_i} \quad (7)$$

Where :

- $\Sigma \text{diag}(CM_i)$: represents the sum of the diagonals of the confusion matrix of well ranked instances
- $\Sigma CM_i$: it is the sum of all the coefficients of the confusion matrix

$n$: number of instances of the database.

We applied the confusion matrix to evaluate each method used. CM gives us the exact classification of each instance shown in the table 5.

<table>
<thead>
<tr>
<th>The rate of accuracy or precision in %</th>
<th>GA</th>
<th>DE</th>
<th>PSO</th>
<th>SOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer Database (BC)</td>
<td>96.48</td>
<td>95.31</td>
<td>95.75</td>
<td>96.48</td>
</tr>
<tr>
<td>Heart Disease Database (HD)</td>
<td>68.25</td>
<td>63.25</td>
<td>66.25</td>
<td>63.75</td>
</tr>
<tr>
<td>Liver Disorder Database (LD)</td>
<td>52.75</td>
<td>54.65</td>
<td>49.96</td>
<td>61.57</td>
</tr>
<tr>
<td>Diabetes Database (PID)</td>
<td>39.32</td>
<td>39.18</td>
<td>39.19</td>
<td>60.10</td>
</tr>
</tbody>
</table>

Table 5. Results from the confusion matrix for the four databases used by the four metaheuristics
Breast Cancer Database
For this basis, and from Table 5 we note that the four algorithms gave good results but SOS and GA have a gave the same and better results which is 96.48%. We note that SOS always gives the best result from the beginning and a quick convergence of its fitness function compared to the others.

Heart Disease Database
We note from Table 5 that the SOS algorithm gave the best results that are of the order of 63.75%. This particularity comes from the fact that SOS gave at the first execution 65.75% and the same results in all the executions, unlike the other algorithms where the results vary between 28% and 63 or 68.25%. We conclude that SOS is credible and others have not a stable behavior, the SOS fitness feature converges quickly.

Liver Disorder Database
For this difficult data base, SOS is clearly better than the other metaheuristics; two of them have middle results but PSO is the worst because it gave us less than 50%.

Diabetes Database
By using this very complex data base, only SOS gives us good results but the three other methods gives us less than 50% as classification results.

4.9.4 Validation of Classification using ROC Curves
To clearly visualize the performances of the models used, we have plotted the ROC curves that use the confusion matrix for the comparison of the models with the real classification by varying the threshold from 1 to 0 for each case [25].

If the results curves are upper than the diagonal (50%), then the algorithm used gave good classification, else, which means under the diagonal, the method used gave an incoherent result.

The results of ROC curve to show the classification of each instance in corresponding real class using the same data bases are in the following figures.

Breast Cancer Database
Our algorithm SOS and GA give a very good classification but the other algorithms DE and PSO are less significant is shown in figure 10.

Heart Disease Database
Although SOS is the third method for the classification of this base, its ROC curve keeps the same stability during several executions with the same conditions. Contrariwise, the other three methods are unstable as shown in Figure 11 which indicates that the classification is less than the diagonal. The corresponding histogram of this behavior is shown in figures 12 and 13.
Figure 12. ROC curves of GA, DE and PSO executed on Heart Disease Database gives inconsistent result

Figure 13. Histogram of GA, DE and PSO executed on Heart Disease Database gave inconsistent result

Figure 14. ROC curves of Liver Disorder Database for 4 algorithms

Figure 15. ROC curves of Diabetes (PID) Database for 4 algorithms
Liver Disorder Database
For this complex data base SOS gave us good results that other methods clarified in the figure 14.

Diabetes Database
By using this very complex and difficult data base, we found that SOS gave very good results corresponding to the complexity of this data base, but the other methods gave us bad classification under the diagonal as shown in figure 15.

5. Conclusion and Future Perspectives
In this paper, we have presented the Symbiotic Organisms Search (SOS) algorithm inspired by biological interactions between organisms in an ecosystem. To highlight the robustness of the SOS algorithm, the domain of valorization and experimentation of the latter is data mining. The process of datamining is a type of combinatorial optimization problem. For the enhancement of our theoretical choices, we have applied the SOS algorithm on medical field. Through the experiments used, it is clear that the symbiotic method has compared to the other methods a very large aspect of exploitation allowing him to have the exact value in the majority of the practical cases realized. From the results of the confusion matrix we find that the SOS algorithm converges quickly to the optimal solution and remains stable in this value unlike the other algorithms. Although SOS is taking its first steps and only its basic version has been used, contrary to the other AG, PSO and DE algorithms with their latest versions, the results obtained show that SOS is the first algorithm in convergence. Statistically, SOS gave us better results in all experiments performed on complex databases. The classification accuracy rates achieved by SOS are higher than those of the other algorithm in three out of four datasets. According to the convergence curves, SOS generally has the fastest convergence behavior on all databases. We point out that the new SOS algorithm is robust, it is easy to implement. We also note its ability to solve various numerical optimization problems despite the use of fewer control parameters than competing algorithms.

We also see, at the end of this work, which perspectives we assume are very promising and which are as follows: improvement of SOS for the treatment of its operating phase; in data mining we will be able to exploit the unsupervised approach; application of this research work in other areas like textmining, email analysis and classification, data analysis in social networks. Finally, we find that we can do the hybridization of SOS with other metaheuristics.

References


[22] Kalami Heris, Moustapha., S. Project Title: Evolutionary Clustering and Automatic Clustering, Khaje Nasir Toosi University Tehran, Iran.

