Two Enhancement Levels for Male Fertility Rate Categorization Using Whale Optimization and Pegasos Algorithms

Abeer S. Desuky

Al-Azhar University, Mathematics Department, *Faculty of Science, Cairo, Egypt.*

Address For Correspondence:
Abeer S. Desuky, Al-Azhar University, Mathematics Department, Faculty of Science, Cairo, Egypt.

A B S T R A C T

Background: Recently, diseases and health problems which were common only in the elderly, became common also among the youth. Some of these medical problems causes are behavioral, environmental and lifestyle factors. The decrease in fertility rates especially among the male population is one of those problems. Now, machine learning and artificial intelligence algorithms are an emerging methodology as computer aided decision systems in medical diagnosis and health problems. **Objective:** In this paper an incorporation of bio-inspired whale optimization algorithm (WOA) and Pegasos algorithm is used to enhance the male fertility rate categorization in two levels. **Results:** show that implementing WOA as second level of enhancement gives better accuracy than the first level of enhancement in Pegasos algorithm with prediction accuracy value of 90%. **Conclusion:** using two machine learning algorithms to categorize male fertility rate helped in overall improving for the proposed system performance to give results exceeded all recent researches results for fertility data.

INTRODUCTION

There has been a remarkable decrease in fertility over the last two decades and the global health community has recently concerned with the decline in male fertility. Studies showed that thirty percent of infertility is related to male factor. The evaluation of male fertility potential was based mainly on semen analysis which is the core of evaluating men for fertility or infertility, but late studies have proved that the test is not the only available way for predicting the fertility status of men reliably. Infertility has many causes including hormonal imbalances, psychological, environmental and behavioral factors (http://www.who.int/) (Barrat *et al*.,1998). To obtain a quick assessment of the seminal profile of the patients based on these factors, machine learning and artificial intelligence techniques have been applied as computer aided decision systems in medical diagnosis field. Such systems can be used for prior assessment of semen quality before more exhaustive, uncomfortable and costly tests are conducted on patients (Wang *et al*.,2014).

Computer aided decision (CAD) systems which rely on a wide range of classifiers have been investigated and implemented for the diagnosis of various diseases. Constructing a classifier, requires a set of data samples representing previous experience (Mazurowski *et al*.,2008). Generally, the larger and more representative the set of samples, the better the categorization of new query cases can be obtained. However, in medical domain, collecting data faces various challenges and practical limitations. One of these is the time consumed in collecting data from patients. Another challenge, gathering large numbers of patients infected with diseases that has low prevalence (imbalance). Moreover, the clinical presentation of the same disease' patients vary dramatically. According to this variability, computer aided systems are requested to handle large numbers of features, often many of which are correlated and/or have no significant diagnostic value. These described issues (limited sample size, imbalanced data, and large numbers of potentially insignificant/correlated features) can detrimentally affect

the performance and development of CAD classifiers (Barrat *et al.*,1998; Mazurowski *et al.*,2008; Raudys *et al.*,1991).

As is prevailing in medical data diagnosis (Mazurowski *et al.*,2008), most available male fertility rating data also have the class imbalance problem, where, the number of normal (negative) samples far exceeds the number of altered (positive) samples. Owing to this distorted distribution of data, many learning algorithms often perform inaccurately on the minority class (Wang *et al.*,2014; Jose *et al.*,2013). Some researchers have handled classification in the presence of these challenges. David Gil *et al.* (2012), proposed and compared three machine learning classifiers: multilayer perceptron (MLP), decision trees (DT) artificial and support vector machines (SVM) and the highest accuracy was 86% for SVM and MLP. In other work (Macmillan,2015), Artificial Neural Network (ANN) and Naïve Bayes (NB) classifiers presented for the characterization of seminal quality, based on environmental factors and life style habits, comparisons between the two classifiers showed that their accuracy rate was identical (80%). Particle Swarm Optimization (PSO) also proposed in (Puneet *et al.*,2015) and compared to another three learning algorithms. PSO showed the highest performance accuracy (88 %) whereas, SVM and MLP showed less performance accuracy (85%) and DT had the much less accuracy (84%).

The objective of this paper is to enhance the categorization of male fertility rate by considering some environmental and behavioral factors. Due to complexity of the problem itself, the data imbalance and its impact on categorization accuracy, two levels of enhancement will be applied using Pegasos algorithm as the first enhancement level and the bio-inspired whale optimization algorithm as the second enhancement level.

MATERIALS AND METHODS

Fertility Data Set:

The dataset contains 100 instances, which obtained from semen samples taken from 100 volunteers and analyzed according to WHO (World Health Organization) 2010 criteria (http://www.who.int/). The volunteers were asked to fulfill nine survey about health status and life habits. According to the answers of the survey, nine features are determined to form this dataset which are: season in which the analysis was performed, age, childish disease, accident/trauma, high fever in the last one year, surgical intervention, alcohol consumption frequency, smoking habits and number of hours spent sitting per day plus the class which has two values (normal and altered fertility rate). The data were collected and shared in 2013 by the department of Biotechnology of University of Alicante (David *et al.*,2012). The data features and description are represented in **Table 1** and **Table 2** respectively.

Table 1: Fertility data features

No.	Features	Abbreviation
1	Season in which the analysis was performed	SAP
2	Age at the time of analysis	AGE
3	Childish disease	CDS
4	Accident or serious trauma	ACC/TRU
5	Surgical intervention	SUR
6	High fever in the last one year	HF
7	Frequency of alcohol consumption	FAC
8	Smoking habits	SH
9	Number of hours spent sitting per day	HSS
10	Class	DIAGNOSIS

Table 2: Features description

Abbreviation	Feature type	Feature values	Normalized
SAP	Nominal	(1) Winter, (2) spring, (3) summer, (4) fall	(-1, -0.33, 0.33, 1)
AGE	Nominal	18-36	(0, 1)
CDS	Real	(1) Yes, (2) No	(0, 1)
ACC/TRU	Nominal	(1) Yes, (2) No	(0, 1)
SUR	Nominal	(1) Yes, (2) No	(0, 1)
HF	Nominal	(1) less than three months ago, (2) more than three months ago, (3) no	(-1, 0, 1)
FAC	Nominal	(1) several times a day, (2) every day, (3) several times a week, (4) once a week, (5) hardly ever or never	(0, 1)
SH	Nominal	(1) never, (2) occasional, (3) daily	(-1, 0, 1)
HSS	Nominal	1-16	(0, 1)
DIAGNOSIS	Binominal	(1) normal, (2) altered	(0, 1)

The Whale Optimization Algorithm:

The whale optimization algorithm (WOA) is a new addition to meta-heuristic algorithms is which proposed by Mirjalili and Lewis (2016). The WOA algorithm is inspired by the hunting method of humpback whales that is called bubble-net hunting strategy. The humpback whales prefer to hunt small fishes close to the surface. Thus,

they swim within a shrinking circle and along a spiral-shaped path simultaneously around the prey (small fishes) to form distinctive bubbles along a 9-shaped path.

The mathematical simulation of this behavior in WOA, is given as follows:

1. Shrinking encircling preys: In WOA, the prey (target) and the other hunting whales (search agents) try to update their positions towards it is assumed as the currently best candidate solution. This is represented by the following equations:

$$D = |C . W^*(t) - W(t)| \tag{1}$$
$$W(t + 1) = W^*(t) - A . D \tag{2}$$

Where: t is the current iteration, $W^*(t)$ is the historically best position vector of a whale, $W(t)$ is a whale position vector and A, D are coefficient vectors which are calculated as follows:

$$A = 2 . a . r - a \tag{3}$$
$$C = 2 . r \tag{4}$$

where a is a variable linearly decreased from 2 to 0 over the course of iterations and r is a random vector uniformly distributed in [0,1].

2. Spiral bubble-net maneuver: To mimic the helix-shaped movement of whales, a spiral equation between the position of whale and prey is used as follows:

$$W(t + 1) = e^{bk} . cos(2 \pi k) . D^{'} + W^*(t) \tag{5}$$
$$D^{'} = |W^*(t) - W(t)| \tag{6}$$

where b is a constant for defining the shape of the logarithmic spiral, and k is a random number uniformly distributed in [−1,1].

When A is greater than 1 or less than −1, in order to have a global optimizer, the search agent is updated according to a randomly chosen search agent instead of the best search agent avoiding local optima:

$$W(t + 1) = W_{rand} - A . D^{''} \tag{7}$$
$$D^{''} = |C . W_{rand} - W(t)| \tag{8}$$

where W_{rand} is selected randomly from hunting whales in the current iteration (Mirjalili *et al*.; Kaveh *et al*.; Hongping *et al*., 2016).

Pegasos: Primal Estimated sub-GrAdient SOlver for SVM:

Pegasos is a simple and effective stochastic sub-gradient descent algorithm described by Shalev Shwartz *et.al.* (2007) for solving the optimization problem cast by Support Vector Machines (SVM).

Given a training set $(x_i, y_i)(i = 1, ..., n)$, where x is a d-dimensional feature vector and $y_i = \pm1$ is class label, linear support vector machines aim to find a classifier in the form of

$$h_{w,b}(x) = sign(w^T x + b) \tag{9}$$

where w and b are parameters. By introducing a dummy feature of constant value 1s, and ignore the parameter b. Then the parameter w can be learnt from the training set (Charu *et al*.,2015).

Unlike other algorithms which formulate SVM as constrained optimization problem, Pegasos algorithm relies on the following unconstrained formulation of SVM:

$$\min_w = \frac{\lambda}{2} w w^T + \frac{1}{n} \sum_{i=1}^{n} l\big(w, (x_i, y_i)\big) \tag{10}$$

where
$$l\big(w, (x_i, y_i)\big) = \max\{0, 1 - y_i w^T x_i\} \tag{11}$$

In each iteration Pegasos aims to update the weight vector w by using k training examples through the following approximate objective function (Shalev *et al*. ,2007; Charu *et al*.,2015):

$$f(w, i_t) = \frac{\lambda}{2} w w^T + \frac{1}{k} \sum_{i \epsilon A_t} l\big(w, (x_i, y_i)\big) \tag{12}$$
Where $1 \leq k \leq n$ *and* $A_t \subset \{1, ..., n\}$, $||A_t|| = k$.

Proposed Method:

In this paper, we use WOA algorithm to optimize Pegasos algorithm performance. The mechanism of WOA helps us to obtain the optimal *w*-parameter which is the core of Pegasos algorithm. Applying WOA to Pegasos is considered as applying two levels of optimizations to SVM since, Pegasos algorithm described mainly for solving the optimization problem cast by Support Vector Machines. In our proposed algorithm, WOA is used to promote population of candidate solutions which represent the *w*-parameter for attributes. The W* is the best candidate *w*-parameter which is the one getting the highest fitness function (minimum classification error) used in Pegasos classifier. The detailed steps are described as follows:

1) Initialize the population for N hunting whales' positions (the *w*-parameter individuals) randomly. Assume that, the number of attributes for the *w*-parameter individual is d. A single whale position is $W_i = \{x_{i,j}, j = 1,2,...,d\}$, where $x_{i,j}$ is j^{th} *w*-parameter for the ith individual. For each individual, the value of the *w*-parameter is vector of random numbers in the range from -10 to 10 with length d.

2) Initialize the parameters C, A and k.

3) Calculate the corresponding fitness function for each hunting whale and select the best whale (leader) position and store it as W*.

4) The fitness function is calculated from the classification mean squared error which is obtained by Pegasos classifier (first level of optimization) to differentiate the normal and altered fertility rate correctly.

$$Fitness_{min} = MSE \tag{13}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y} - Y)^2 \tag{14}$$

where, $\hat{Y}$ is a vector of n predictions, and Y is the vector of observed values corresponding to the inputs to the function that generated the predictions.

5) Spiral equation (5) between the position of whale and prey and equations (6 -8) are used to update the positions of other whales.

6) Update parameters C, A, and k.

7) Go to step 3 if t is less than maximum number of iterations.

8) Return the position of W* which is the individual vector of best *w*-parameter vector and return the associated performance measures (second level of optimization).

Results:

The experiments were conducted in the Matlab R2015a platform on computer equipped with 2.20GHZ core i7 processor and 6GB RAM. Since classification Accuracy alone is typically not enough information to validate algorithms performance, different performance measures such as Sensitivity, Precision, F Measure are used for validating our method. Measures which are used for the validation of the data are defined as (Lamiaa *and Abeer*,2015; Namita *et al.*, 2012):

Accuracy: It is the number of correct predictions made divided by the total number of predictions.
$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{15}$$

Sensitivity: It is also called the Recall or true positive rate (TPR)
$$Sensitivity = TP/(TP + FN) \tag{16}$$

Precision: It is also called the Positive Predictive Value (PPV).
$$Precision = TP/(TP + FP) \tag{17}$$

F Measure: It is also called F Score and it conveys the balance between the precision and the recall.
$$FMeasure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{18}$$

Table 3, shows the classification performance measures for Pegasos algorithm (first enhancement level) and the optimized Pegasos using WOA (second enhancement level). In our experiment the Pegasos algorithm applied first with parameters (λ=1, k=10 and 5 iterations) on the data and the performance measures are calculated. The second level of enhancement is applied on the data by implementing WOA algorithm on Pegasos (with the same parameters) and the WOA parameters were chosen to be 50 hunting whale agents and a random number in the -10 and 10 range is extracted for each whale in each iteration. If it is greater than 0.5, Eq. (5) is selected; otherwise, Eq. (7) is chosen for updating whale's position. The maximum number of implementing WOA chosen to be 500 iterations.

Table 3: Performance Measures for Pegasos classifier with and without WOA algorithm

Method	Accuracy	Precision	Sensitivity	F Measure
Pegasos	88.00	87.88	98.86	93.05
WOA/Pegasos	90.00	89.80	100	94.63

Fertility data is a new data set, became available for use by researchers in 2013, so only few researches worked on it. **Table 4**, presents a comparison between the performance of our proposed algorithm and some algorithms implemented recently on the fertility data.

Table 4: Performance Comparison

Algorithm	Accuracy	Sensitivity
Proposed	90	100
Pegasos	88	98.9
SVM (David,2012)	86	97.7
MLP (David,2012)	86	94.1
DT (David,2012)	84	96.5
ANN (Macmillan2015)	80	98.9
NB (Macmillan2015)	80	97.7
PSO (Puneet,2015)	88	92.1

From the experiment results shown in **Table 4**, we can see that the classification accuracy of first enhancement level (Pegasos algorithm) is increased by 2 % than classification accuracy obtained by SVM in (David *et al*.,2012) while the increment in the classification accuracy was 4% after applying the second enhancement level (WOA algorithm). Generally, the results showed that the results achieved by the proposed method are higher than other algorithms implemented on fertility data.

Conclusion:

Fertility dataset is highly imbalanced data and biased towards the 'Normal' class so, two enhancement levels have been implemented using two machine learning algorithms in order to categorize male fertility rate. Primal Estimated sub-GrAdient SOlver for SVM (Pegasos) and Whale Optimization Algorithm (WOA) are used for first and second levels of enhancement respectively. experimental results showed that using (WOA) as second level of enhancement for Pegasos algorithm helped in overall improving for the proposed algorithm performance to give results exceeded all recent researches results for fertility data.

REFERENCES

Barrat, C.L., S. Clements, E. Kessopoulou, 1998. Semen Characteristics and fertility tests required for storage of spermatozoa, Human Reproduction (Oxford, England), 13(2): 1-7.

Charu, C., Aggarwal, 2015. Data Classification: Algorithms and Applications, CRC Press, pp: 707.

David Gil, Jose Luis Girela, Joaquin De Juan, M. Jose Gomez-Torres, and Magnus Johnsson, 2012, Predicting seminal quality with artificial intelligence methods, Expert Systems with Applications, 39(16): 12564-12573.

Hongping Hu, Yanping Bai, Ting Xu, 2016, A whale optimization algorithm with inertia weight, WSEAS TRANSACTIONS on COMPUTERS, 15: 319-326.

Jose, L. Girela, David Gil, Magnus Johnsson, María José Gomez-Torres, Joaquín De Juan, 2013. Semen Parameters Can Be Predicted from Environmental Factors and Lifestyle Using Artificial Intelligence Methods, Bio Reprod, 88 (4): 99, 1-8. doi: 10.1095/biolreprod., 112: 104653.

Kaveh, A. and M. Ilchi Ghazaan, 2016. Enhanced Whale Optimization Algorithm for Sizing Optimization of Skeletal Structures, Mechanics Based Design of Structures and Machines, DOI: 10.1080/15397734.2016.1213639

Lamiaa, M. El Bakrawy, Abeer S. Desuky, 2015. A hybrid classification algorithm and its application on four real-world data sets, International Journal of Computer Science and Information Security (IJCSIS), 13: 10.

Macmillan Simfukwe, Douglas Kunda and Christopher Chembe, 2015. Comparing Naive Bayes Method and Artificial Neural Network for Semen Quality Categorization, IJISET - International Journal of Innovative Science, Engineering & Technology, 2(7): 689-695.

Mazurowski, M.A., P.A. Habas, J.M. Zurada, J.Y. Lo, J.A. Baker, G.D. Tourassi, 2008.Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, Neural Networks, 21: 427-436.

Mirjalili, S., and A. Lewis, 2016. The whale optimization algorithm, Advances in Engineering Software, vol. 95, Elsevier Ltd., pp: 51-67.

Namita Aggarwal, R.K. Agrawal, 2012. First and Second Order Statistics Features for Classification of Magnetic Resonance Brain Images, Journal of Signal and Information Processing, 3: 2.

Puneet Singh Duggal, Sanchita Paul and Priyanka Tiwari, 2015. Analytics for the Quality of Fertility Data

using Particle Swarm Optimization, International Journal of Bio-Science and Bio-Technology, 7: 1.

Raudys, S.J., A.K. Jain, 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Transactions on Pattern Analysis and Machine Intelligence, 13: 252-264.

Shalev-Shwartz, Shai, Yoram Singer, and Nathan Srebro, 2007. Pegasos: Primal estimated sub-gradient solver for SVM, Proceedings of the 24th international conference on Machine learning, pp: 807-814, ACM .

Wang, H., Qingsong Xu and Lifeng Zhou, 2014. Seminal Quality Prediction Using Clustering-Based Decision Forests, Algorithms, 7(3): 405-417, doi:10.3390/a7030405

World Health Organization., 2010. WHO laboratory manual for the examination and processing of human semen, http://www.who.int/