Feature Selection Based on Whale Optimization Algorithm for Diseases Diagnosis

Hoda Zamani, Mohammad-Hossein Nadimi-Shahraki*
Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad, Iran
hoda_zamani@sco.iaun.ac.ir
nadimi@iaun.ac.ir

***Abstract*—Medical datasets are mainly composed of countless irrelevant and redundant features in a series of patient records. All these features are not required to obtain a medical decision-making process. On the other hand, the huge size of data is caused to increase the dimensionality and to reduce the performance of classifier. Recently, there have been many methods proposed to solve this problem and their results show that the feature selection can be an effective solution. The feature selection methods are mostly aim to reduce the size of data and enhance the efficiency of learning algorithms by eliminating the unrelated and redundant features. In this paper, a meta-heuristic algorithm is proposed named FSWOA for feature selection. This algorithm is based on the hunting methods of Humpback Whales consisting of three main steps: encircling prey, spiral bubble-net attacking and search for prey. The performance of proposed algorithm is evaluated conducted by four standard medical datasets: Pima Indians Diabetes, Original Wisconsin Breast Cancer, Statlog and Hepatitis. The results show that the proposed algorithm can reduce the dimensionality of medical datasets with acceptable accuracy for diseases diagnosis.**

***Index Terms*— Feature selection, Whale optimization algorithm (WOA), Dimensionality reduction, Diseases diagnosis.**

I. INTRODUCTION

Medical data mining is one of the most important issues in recent years, which is relying on analysis and statistical reasoning, machine learning techniques and pattern recognition to discover relations and hidden patterns in the datasets of the patients. Generally, all activities in medicine can be divided into six areas: screening, diagnosis, treatment, prognosis, monitoring and management [1]. The accuracy and sensitivity have particular importance in the diagnosis and prediction of diseases. Its positive feedback can predispose until the doctor by its analysis to speed up the process of diagnosis and prognosis. Therefore, the costs of treatment can be reduced and the rate of health in society can be increased. In the real world, medical databases are usually filled by irrelevant and redundant features which increase the dimension of database or lead them to curse of dimensionality. Then, the accuracy, computational cost and speed of the learning process are affected. Dimensionality reduction methods have been proposed to solve this problem. One of the most famous dimensionality reduction techniques is Feature selection. The feature subset selection problem consists of identifying and selecting a useful subset of features from a larger set of often mutually redundant, possibly irrelevant, features with different associated importance [2, 3].

Features selection techniques can be divided based on their dependence on the classification algorithms in two main categories: model-free and model-based [4]. In the model-free methods, feature extraction is done based on statistical functions and independent of specific data model. Some common model-free methods include: F-Score criterion, information gain, correlation function and maximum relevance minimum redundancy technique. In the model-based approach, the process of Selection and feature extraction are dependent on the performance of the predictor. The model-free is better than model-based because of the independence on specific data model in features extraction in term of speed, scalability and computational cost. But this independence compared with the model-based techniques causes to reduce the performance.

In general, the space of problem must be completely explored-all search-to extract the effective subset of features. However, using all search for the most of real-world problems is impossible because of having the high dimensionality especially in NP-hard problems. Obviously, exploring the whole problem space and evaluating all states are very costly in term of the computational complexity and response time. Therefore, many meta-heuristic algorithms have been proposed to find the optimal solutions inspiring of the fauna's foraging behavior in the nature. They consider trade-off between computational complexity and time mainly based on swarm intelligence which is shared by cooperation and competition between agents. Consequently, some efficient meta-heuristic algorithms have been proposed for feature selection such as Ant Colony Optimization algorithm (ACO) [5], Particle Swarm Optimization (PSO) [6], Artificial Bee Colony (ABC) [7]. Recently, they apply to a large number of applications in the medical sciences [8, 9].

In this paper, a meta-heuristic algorithm is proposed named Feature Selection based on Whale Optimization Algorithm or FSWOA in short. FSWOA is mainly aim to reduce the dimensionality of medical data. In fact, this algorithm is based on the hunting methods of humpback whales including three main operators: encircling prey, spiral bubble-net attacking and search for prey. The rest of paper is organized as follows.

* Corresponding Author

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Section 2 is to review some related works on feature selection based on meta-heuristic. The proposed algorithm is described in section 3. In section 4, the performance of the proposed algorithm is evaluated conducted by well-known medical datasets. Then, the results are shown in section 5 and the conclusions are finally discussed in section 6.

II. Related WORKS

Feature selection methods have been applied to classification problems in order to select a reduced feature set that makes the classifier more accurate and faster [10]. For a large number of features, evaluating all states is computationally non-feasible requiring meta-heuristic search methods [11]. These methods tries to solve the challenges that related to real world problems with competition and cooperation strategy between agents. Many studies have been done in the feature selection, which are intersection with swarm intelligence. Advanced binary ACO (ABACO) was proposed for feature selection and dimension reduction [11]. In 2005, Şahan et al. applied the Attribute Weighted Artificial Immune System (AWAIS) to diagnose Heart and Diabetic diseases. In this study has shown the negative effects of irrelative features in diseases diagnosis process [12]. In other study, Huang proposed a hybrid method of ant colony optimization algorithm and support vector machine for feature selection [13]. Inbarani et al. offered a model based on PSO and rough sets strategy for selected features [14]. Nahar et al. used the computational intelligence to diagnose Heart disease [15]. Another swarm-based meta-heuristic optimization algorithm inspired by the hunting behavior of humpback whales. This algorithm was proposed by Mirjalili and Lewis [16] and called whale optimization algorithm (WOA).

III. Proposed Algorithm

In this section, the proposed algorithm is described which is a meta-heuristic algorithm named Feature Selection based on Whale Optimization Algorithm or FSWOA in short. FSWOA is a new algorithm for feature selection based on the hunting methods of Humpback Whales including three main steps: encircling prey, spiral bubble-net attacking and search for prey. Fig.1. shows the flowchart and main steps of FSWOA. In the first step, it generates k humpback whales and randomly scatter them in the search space. Then, the position of each humpback whale is evaluated and the best whales are selected. The other whales will try to update their positions towards the best whale. In the second step, humpback whales start to attack with a bubble-net strategy. There are two strategies: shrinking encircling and spiral updating position for bubble-net attacking. In fact, this step is similar the exploitation phase in which each whale suggests a subset of features. Then, these subsets of feature are evaluated based on the accuracy of classifier on the testing set. In the third step or the exploration phase, the humpback whales search prey randomly according to the position of each other.

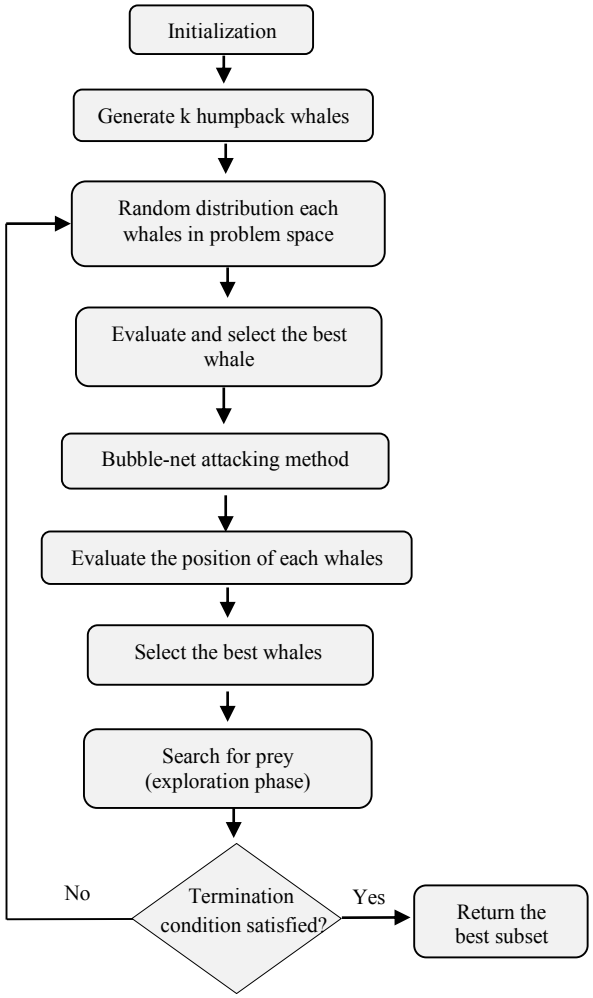

Fig.1.The flowchart of proposed FSWOA

IV. Experimental Evaluation

In this section, the performance of proposed algorithm is evaluated conducted by four benchmark medical datasets downloaded from UCI machine learning repository [17]. These datasets include Pima Indians Diabetes, Original Wisconsin Breast Cancer, Statlog and Hepatitis which are popular for feature selection problems. Table I shows the statistical information of the datasets.

TABLE I
Statistical Information of Datasets

Dataset	Features	Sample	Classes	Missing data
Pima Indians Diabetes	8	768	2	Yes
Original Wisconsin Breast Cancer	10	699	2	yes
Statlog	13	270	2	no
Hepatitis	19	155	2	yes

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The Pima Indians Diabetes dataset (PID) is to diagnose a person has Diabetes or not based on clinical and laboratory data. The purpose of Original Wisconsin Breast Cancer dataset is Breast cancer diagnosis. The heart disease is predicted by Statlog dataset. The objective of Hepatitis dataset is to predict whether a person will be live or die with Hepatitis disease. Since in real world the medical datasets are noisy and incomplete, therefor these datasets are firstly normalized and then the proposed algorithm is evaluated by each dataset.

A. Evaluation Functions

The subsets of features selected by FSWOA algorithm are evaluated by well-known evaluation functions such as: sensitivity, specificity, precision, negative predictive value (NPV), area under the curve (AUC) and accuracy. The sensitivity and specificity shown in Eq. (1) and (2) indicate respectively the samples in the positive and negative classes which are correctly classified. Precision or positive predictive value (PPV) and negative predictive value (NPV) are computed by Eq. (3) and (4). AUC is the true positive rate vs the false positive rate, its value is between 0.0 and 1.0. The cost function of the proposed algorithm is defined by the accuracy of classifier shown by Eq. (5) where the sum TP and FP is total number of subjects with positive test and sum FN and TN is total number of subjects with negative test.

$$\text{Sensitivity (True positive rate)} = \frac{TP}{TP+FN}\,(\%) \quad (1)$$

$$\text{Specificity (True negative rate)} = \frac{TN}{TN+FP}\,(\%) \quad (2)$$

$$\text{Precision (positive predictive value)} = \frac{TP}{TP+FP}\,(\%) \quad (3)$$

$$\text{Negative predictive value} = \frac{TN}{TN+FN}\,(\%) \quad (4)$$

$$\text{Accuracy (ACC)} = \frac{TP+TN}{TP+TN+FP+FN}\,(\%) \quad (5)$$

B. Experimental Setup

The proposed algorithm is implemented using MATLAB on an Intel Core-i5 CPU with 6GB of RAM. To find the best subset of features, our algorithm is tested 15 times by using the evaluation functions described in Section IV-A. During each time, firstly, the datasets were randomly split into two sets of 70% and 30% as a training set and a test set respectively. Then, the proposed algorithm uses K-Nearest Neighbors algorithm with K=3 for evaluating the subset of selected features. In addition, the maximum iteration is set to 60, the initial population size is set to 30 and the lower and upper bound are set to 0 and 1 respectively.

C. Experimental Results

This section shows the result of experimental evaluation of the proposed algorithm in Table II and III where max and mean indicate the maximum and average value respectively. The accuracy of the proposed algorithm observed on these medical datasets are 87.10 % for Hepatitis, 97.86 % for Breast Cancer, 78.57 % for Pima Indians Diabetes and 77.05 % for Statlog Disease. Moreover, the proposed algorithm selects 6 features for diagnosis of heart disease, 7 for Pima Indians Diabetes, 8 for Hepatitis and 4 for Breast cancer.

Finally, the feature reduction rate is computed for all datasets. As Table 3 shown, its results for Hepatitis, Pima Indians Diabetes, Breast Cancer and Heart Disease are 57.89 %, 60 %, 12.5 % and 53.85%, respectively. In addition, Fig. 2 and 3 show the dimensionality reduction rate and the accuracy of classifier of Diabetes and Heart diseases.

TABLE II
Evaluation functions (in %) of FSWOA algorithm

Datasets	# NF	Accuracy	sensitivity	specificity
Hepatitis	8			
Max		87.10	100.00	94.12
Mean		73.33	70.51	79.97
Breast Cancer	4			
Max		97.86	98.90	100.00
Mean		96.57	96.60	96.47
Pima Indians Diabetes	7			
Max		78.57	90.65	63.46
Mean		70.87	82.72	48.28
Statlog	6			
Max		77.05	100.00	82.76
Mean		62.84	94.71	64.36

TABLE III
Evaluation functions (in %) of FSWOA algorithm

Datasets	# NF	AUC	PPV	NPV
Hepatitis	8			
Max		0.971	88.89	100.00
Mean		0.752	75.99	76.56
Breast Cancer	4			
Max		0.994	100.00	97.92
Mean		0.965	98.17	93.64
Pima Indians Diabetes	7			
Max		0.771	80.83	70.59
Mean		0.655	75.37	59.67
Statlog	6			
Max		0.913	90.24	100.00
Mean		0.795	76.74	91.53

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE IV
Features reduction rate and subset of effective features

Dataset	Features Reduction rate (%)	Effective features
Hepatitis	57.89	[27,5,28,11,15]
Breast Cancer	60.00	[7,2,3,10]
Statlog	53.85	[2,3,6,9,10,11]
PID	12.5%	[5,6,1,3,8,4,7]

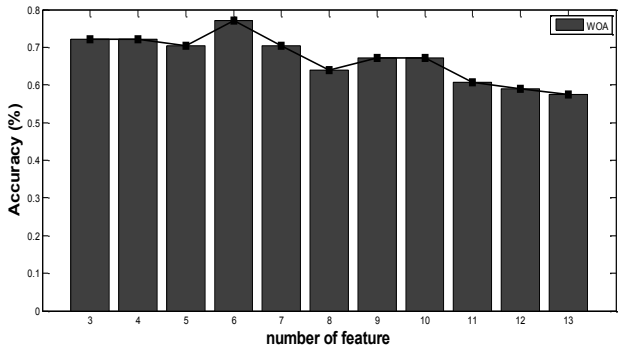

Fig. 2. The accuracy of classifier for different number of features in Heart disease dataset

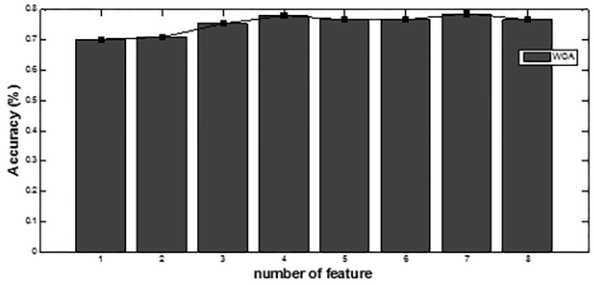

Fig. 3. The accuracy of classifier for different number of features in Pima Indian Diabetes dataset.

V. CONCLUSION

Feature selection is the process for choosing a subset of features that maximizes the performance of learning algorithm and reduces the dimensionality of the problem space. In this study, an efficient feature selection algorithm was proposed based on whale optimization algorithm named FSWOA. The proposed algorithm searches the problem space and extract a subset of optimal features for medical decision-making. The performance of this algorithm was experimentally evaluated conducted by four different medical datasets. The experimental results show that our method can reduce the dimension of medical datasets in diseases diagnosis with an acceptable accuracy.

REFERENCES

[1] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend," *Expert Systems with Applications,* vol. 41, pp. 4434-4463, 7// 2014.

[2] H. Liu and H. Motoda, "Feature Selection Methods," in *Feature Selection for Knowledge Discovery and Data Mining*, ed Boston, MA: Springer US, 1998, pp. 73-95.

[3] S. M. Vieira, J. M. C. Sousa, and T. A. Runkler, "Two cooperative ant colonies for feature selection using fuzzy models," *Expert Systems with Applications*, vol. 37, pp. 2714-2723, 4// 2010.

[4] I. Guyon, Andr, #233, and Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157-1182, 2003.

[5] M. Dorigo and T. Stützle, "The Ant Colony Optimization Metaheuristic: Algorithms, Applications, and Advances," in Handbook of Metaheuristics, F. Glover and G. A. Kochenberger, Eds., ed Boston, MA: Springer US, 2003, pp. 250-285.

[6] B. Xue, M. Zhang, and W. N. Browne, "Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach," *IEEE Transactions on Cybernetics,* vol. 43, pp. 1656-1671, 2013.

[7] B. Akay and D. Karaboga, "Artificial bee colony algorithm for large-scale problems and engineering design optimization," Journal of Intelligent Manufacturing, vol. 23, pp. 1001-1014, 2012.

[8] S. Al-Muhaideb and M. El Bachir Menai, "Hybrid Metaheuristics for Medical Data Classification," in *Hybrid Metaheuristics*, E.-G. Talbi, Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 187-217.

[9] S. M. Vieira, L. F. Mendonça, G. J. Farinha, and J. M. C. Sousa, "Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients," *Applied Soft Computing,* vol. 13, pp. 3494-3504, 8// 2013.

[10] H.-H. Hsu, C.-W. Hsieh, and M.-D. Lu, "Hybrid feature selection by combining filters and wrappers," *Expert Systems with Applications,* vol. 38, pp. 8144-8150, 7// 2011.

[11] S. Kashef and H. Nezamabadi-pour, "An advanced ACO algorithm for feature subset selection," *Neurocomputing,* vol. 147, pp. 271-279, 1/5/ 2015.

[12] S. Şahan, K. Polat, H. Kodaz, and S. Güneş, "The Medical Applications of Attribute Weighted Artificial Immune System (AWAIS): Diagnosis of Heart and Diabetes Diseases," in Artificial Immune Systems: 4th International Conference, ICARIS 2005, Banff, Alberta, Canada, August 14-17, 2005. Proceedings, C. Jacob, M. L. Pilat, P. J. Bentley, and J. I. Timmis, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 456-468.

[13] C.-L. Huang, "ACO-based hybrid classification system with feature subset selection and model parameters optimization," *Neurocomputing*, vol. 73, pp. 438-448, 12// 2009.

[14] H. H. Inbarani, A. T. Azar, and G. Jothi, "Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis," *Computer Methods and Programs in Biomedicine*, vol. 113, pp. 175-185, 1// 2014.

[15] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," *Expert Systems with Applications,* vol. 40, pp. 96-104, 1// 2013.

[16] S. Mirjalili and A. Lewis, "The Whale Optimization Algorithm," *Advances in Engineering Software,* vol. 95, pp. 51-67, 5// 2016.

[17] K. Bache, M. Lichman (2016) UCI machine learning repository. School of Information and Computer Science, University of California, Irvine. Available: http://archive.ics.uci.edu/ml.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Hoda Zamani was born in Iran. She received both B.S. and M.S. degrees in software engineering in 2012 and 2015 respectively from the Faculty of Computer Engineering, Islamic Azad University of Najafabad (IAUN) in Iran. Her research interests include data mining, medical data mining and meta-heuristic algorithm.

Mohammad-Hossein Nadimi-Shahraki was born in Iran. He received his PhD in computer science with major of artificial intelligence and data mining from University Putra of Malaysia (UPM) in 2010. His research interests include data mining, medical data mining, social network mining and big data mining. He was director general of research in IAUN from 2012 to 2014 and currently he is dean of faculty of computer engineering of Islamic Azad University of Najafabad (IAUN) in Iran. Dr. Nadimi is a member of professional societies such as IEEE and IAENG. He was awarded in International Research and Technology Expo, Malaysia Invention & Innovation Awards 2010 (MTE2010). He was also awarded as top researcher in 2012 and 2014 in IAUN and his data mining book was awarded as top book in 2016 in Iran.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500