FEATURE SELECTION USING WHALE SWARM ALGORITHM AND A COMPARISON OF CLASSIFIERS FOR PREDICTION OF CARDIOVASCULAR DISEASES

[1]Anuradha.P, [2]Dr.Vasantha Kalyani David

[1]Research Scholar, [2] Professor, Dept of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women,
Deemed to be University, Coimbatore, Tamilnadu, India

Abstract: Selection of relevant features of a dataset is necessary in high dimensional datasets in order to avoid the curse of dimensionality. Feature Selection is performed to reduce overfitting, to improve accuracy and to reduce the training time of the algorithms. In this paper a nature inspired meta heuristic algorithm called Whale Swarm Algorithm devised by Bing Zeng, Liang Gao and Xinyu Li is used for feature subset selection (Bing Zeng, 2017). It focuses on selecting a features subset using Whale Swarm Algorithm (WSA) where Logistic Regression (LR), Random Forest (RF) and k-Nearest Neighbor (KNN) are used as fitness functions. These WSA-LR, WSA-RF and WSA-KNN combinations generate different feature subsets for different number of iterations. Then training and testing is done on the dataset with the subset of selected features using LR, RF, Support Vector Classifier (SVC) and Gaussian Naive Bayes (GNB) and prediction accuracies generated are analyzed.

IndexTerms **– Feature Selection, Whale Swarm Algorithm, Logistic Regression (LR), Random Forest (RF) and k-Nearest Neighbor (KNN)**

1. Introduction

According to WHO, 31% of all global deaths is due to Cardio Vascular Diseases (CVDs) [18]. Machine Learning algorithms can be applied on the health care data to predict heart diseases. These algorithms would support the medical practitioners to gain insight into the high dimensional data thereby assisting them to predict and prevent heart diseases.

1.1 Machine Learning

Machine Learning is the ability of systems to learn from data (by training) without being explicitly programmed. Machine learning algorithms can be classified into two groups: supervised learning and unsupervised learning.

In supervised learning, a set of independent variables/ features with the corresponding output variables is input to train the model (Mathworks, 2012). This model would then predict the output to the new input variables. Supervised learning uses classification and regression techniques to develop predictive models (Mathworks, 2012). Classification techniques predict discrete responses and Regression techniques predict continuous responses (Mathworks, 2012).

Common algorithms for performing classification include Support Vector Machine (SVM), Decision Trees, Random Forest, K-Nearest Neighbors, Naive Bayes, Discriminant Analysis, Logistic Regression, and Neural Networks (Mathworks, 2012).

Common algorithms for performing Regression include Linear model, Non Linear model, Regularization, Stepwise Regression, Boosted and Bagged Decision Trees, Neural Networks, and Adaptive Neuro-Fuzzy learning (Mathworks, 2012).

Unsupervised Learning finds hidden patterns or intrinsic structures in data (Mathworks, 2012). It draws inferences from datasets consisting of input data without labelled outputs [10]. Clustering is the most common unsupervised learning technique (Mathworks, 2012).

Common algorithms for performing clustering include K-Means, Hierarchical Clustering, Gaussian Mixture models, Hidden Markov models, Self-Organizing Maps, Fuzzy C-Means Clustering, and Subtractive Clustering (Mathworks, 2012).

1.2 Feature Selection

In High dimensional datasets, Feature Selection aims at reducing the redundant and irrelevant features. The refined dataset with only the relevant features would improve the learning accuracy and reduce the learning time (Jie Cai, 2018). The features which are used to train the machine learning model highly influence the performance of the model. Irrelevant features can bring down the performance of the model.

Feature Selection can be broadly classified into filter method, wrapper method and Embedded method. In Filter method, various statistical tests are used to select the features based on their correlation with the outcome or dependent variable. In wrapper method, a subset of features is used to train a model. Based on the performance of the model features will be added or removed to / from the subset. In embedded method, both the advantages of filter and wrapper methods are combined. The embedded method algorithms perform subset selection, train a model and also execute a penalization function to reduce overfitting.

1.3 Nature-inspired Metaheuristic algorithms

If there are 'm' features in a dataset then there would be 2^m possible subsets of features, where a complete search to get an ideal solution would be almost impossible, especially when 'm' is large. The wrapper methods can be used to extensively search, but, these methods could get trapped in local optima (Ah. E. Hegazy, 2018). Moreover, exploring the entire problem space and evaluating all subsets is costly in term of the computational complexity and response time (Hoda Zamani,2016). Recently, many nature-inspired metaheuristic techniques that result in global optimization are used to solve NP-Hard problems (Ah. E. Hegazy, 2018).

Generic meta-heuristic algorithms include Bat Algorithm (BA), Bacterial Foraging Optimization (BFO), Cuckoo Search (CS), Genetic Algorithm (GA), Particle Swarm optimization (PSO), Whale Optimization Algorithm (WSO), Multi-Verse Optimizer (MVO), Grey Wolf optimization Algorithm (GWO) and Whale Swarm Algorithm (WSA) (Yuefeng Zheng ,2019). Swarm intelligence (SI) techniques has been used to solve NP-hard (Non-deterministic Polynomial time) computational problems. It has been used successfully for Feature Selection in some applications (Lucija Brezo cnik, 2018).

1.4 Whale Hunting Behavior

Whales are social animals and live in groups in the oceans. They make different sounds to indicate their migration, feeding and mating patterns. They determine food azimuth and keep in touch with each other from large distances by ultrasound. When a whale has found food source, it will make sounds to notify other whales nearby of the quality and quantity of food (Bing Zeng, 2017). So each whale will receive lots of notifications from the neighbors, and then move to the proper place to find food based on these notifications (Bing Zeng, 2017). The behavior of whales communicating with each other by sound for hunting inspired Bing Zeng et al., (Bing Zeng, 2017), to develop a metaheuristic algorithm for function optimization problems.

In (Bing Zeng, 2017), Whale Swarm Algorithm, the following four idealized rules are employed:

1) all the whales communicate with each other by ultrasound in the search area

2) each whale calculates the distance of it from other whales

3) the quality and quantity of food found by each whale are associated to its fitness

4) the movement of a whale is guided by the nearest one among the whales that are better (judged by fitness) than it, such nearest whale is called the "better and nearest whale" (Bing Zeng, 2017).

Figure 1: gathering of sperm whales

The random movement of a whale X guided by its better and nearest whale Y can be formulated as follows:

$$X^{t+1}_{i+1} = X^{t}_{i} + rand(0, \rho_0 . e^{-\eta . dxy}) . (y_i - x_i) \quad (1)$$

where, x_i and x_{i+1} are the i^{th} elements of X's position at t and t+1 iterations respectively, similarly, y_i denotes the i^{th} element of Y's position at t iteration.
ρ_0 is the intensity of ultrasound at the origin of source, η is the probability of message distortion at large distances, $d_{x,y}$ represents the Euclidean distance between X and Y. And rand $(0, \rho_0 . e^{-\eta . dxy})$ means a random number between 0 and $\rho_0) . e^{-\eta . dxy}$. Based on a large number of experiments, ρ_0 can be set to 2 for almost all the cases (Bing Zeng, 2017).

1.5 Whale Swarm Algorithm (Bing Zeng, 2017)**:**

Input: An objective function, the whale swarm n.
Output: The global best.

1. Start
2. Initialize whale positions
3. Evaluate all the whales (calculate their fitness).
4. Find current best
5. global best=current best
6. for i=1 to n do
7. Find the better and nearest whale Y of whale i
8. if Y exists then
9. Whale i moves under the guidance of whale Y according to equation 1;
10. Evaluate whale i; find current best
11. if current best < global best then
12. global best= current best
13. end if
14. end if
15. end for
16. return global best
17. stop

The pseudo code of finding a whale's better and nearest whale (Bing Zeng, 2017):

Input: The whale swarm n, a whale u .
Output: The better and nearest whale u .

```
1: begin
2: Define an integer variable v initialized with 0;
3: Define a float variable temp initialized with infinity;
4: for i=1 to n do
5:     if i≠u then
6:         if f(whale i )<f(whale u ) then
7:             if dist(whale i , whale u )<temp then
8:                 v=i;
9:                 temp=dist(whale i , whale u );
10:            end if
11:        end if
12:    end if

13: end for
14: return whale v ;
15: end
```

2. Related Work

Researchers had worked on various nature-inspired meta heuristic algorithms in the recent years to get global optimal solution.

Majdi M. Mafarja and Seyedali Mirjalili (2017), proposed variants of Whale Optimization Algorithm (WOA) namely WOA-T (Whale Optimization Algorithm with tournament selection), WOA-R (Whale Optimization Algorithm with roulette wheel selection) and WOA-CM (Whale Optimization Algorithm with Crossover and Mutation), which were applied on the feature selection domain. From the UCI repository datasets, eighteen well-known datasets were used to assess the performance of these three approaches. These approaches were compared to Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Ant Lion Optimizer (ALO), and five standard filter feature selection methods and the proposed approaches yielded better results than the others.

Ah. E. Hegazy et al., (2018), had proposed an Improved Whale Optimization Algorithm (IWOA), where a binary form of the feature subsets is used and the result obtained was tested against five optimizers. Further IWOA was compared against WOA, PSO, GA, ALO and GWO by applying on 27 datasets. IWOA was found to produce better classification accuracy when compared to others.

Hoda Zamani et al., (2016), had done Feature Selection based on Whale Optimization Algorithm (FSWOA) with K-Nearest Neighbors algorithm as the fitness function on four different medical datasets. Their work had reduced the dimension of the dataset considerably and produced acceptable accuracy.

Bing Zeng et al., (2017), had proposed Whale Swarm Algorithm (WSA) for function optimization. WSA was compared with several popular metaheuristic algorithms like PSO, GA, locally informed PSO (LIPS), Speciation-based DE (SDE), The original crowding DE (CDE) and Speciation-based PSO (SPSO) on four performance metrics (i.e., SR, ANOF, MPR and

Convergence speed). Their work proved that WSA had a quite competitive performance when compared with other algorithms, in terms of efficiency and in locating multiple global optima.

B Subanya and R R Rajalaxmi (2014), had used a metaheuristic algorithm Binary Artificial Bee Colony Algorithm (BABC -kNN) to select the best features in heart disease diagnosis. The experiment proved that the results converge to the optimal solution quickly.

Tad Gonsalves [13], introduced a novel approach in fine-tuning the Constructive Cost Model.

Model. He applied the Fireworks Algorithm (FWA) to determine the optimal subset of features.

The experimental results imply that accurate cost estimates for a project can be made with fewer cost drivers. This implies that with fewer features, complexity of the model decreases and thereby indirectly reduces the cost involved in collecting data.

P. Mohapatra et al., (2014), proposed modified cat swarm optimization (MCSO) to select the most relevant features of microarray gene expression based medical data. The selected features had been classified applying two variations of kernel ridge regression (KRR), namely wavelet kernel ridge regression (WKRR) and radial basis kernel ridge regression (RKRR). The results showed that KRR outperforms ridge regression (RR), online sequential ridge regression (OSRR), support vector machine radial basis function (SVMRBF) and Random Forest irrespective of the datasets used.

Ismail Babaoglu et al., (2014), used binary particle swarm optimization (BPSO) and genetic algorithm (GA) techniques as feature selection models on coronary artery disease (CAD) existence based upon exercise stress test (EST) data. The classification process implemented by utilizing BPSO had better classification accuracy and minimal process time compared to GA.

Experimental Data and setup

Statlog heart disease dataset from UC Irvine Machine Learning Repository is used. This dataset contains 13 attributes (which have been extracted from a larger set of 75). There are no missing values and there are 270 observations.

Attribute Information:
---
1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Variable to be predicted: Absence (1) or presence (2) of heart disease

The experiment was done using Python on Jupyter notebook on Ubuntu OS run on a system with i5 processor and 4 GB RAM. The Statlog dataset is divided into 230 rows of training set and 40 rows of test set. Three variants of Whale Swarm Algorithm (WSA) where Logistic Regression (LR), Random Forest (RF) and K-Nearest Neighbors (KNN) are used as fitness functions are experimented for obtaining the feature subset selection. The resulting subset of features yielded by each combination is then chosen from the training set and classification is performed using Random Forest, Logistic Regression, Gaussian Naive Bayes and SVC classification algorithms. The accuracy on the testing set is obtained and a comparison is made.

Results and Discussion

The WSA algorithm with Logistic Regression(LR) as fitness function(WSA-LR) gives a subset with seven features on an average. The WSA algorithm with Random Forest (RF) as fitness function(WSA-LR) gives a subset of six features on an average. The WSA algorithm with K-Nearest Neighbors (KNN) as fitness function(WSA-KNN) gives a subset of seven features on an average. Table 1 gives the details of the number of selected features output during the different iterations of 5, 20, 40, 100 for the three variants of WSA.

Table 1:: Number of selected features output for each variant of WSA

	ITERATIONS				
	5	**20**	**40**	**100**	**Average**
WSA-LR	10	6	9	5	8
WSA-RF	7	4	9	7	7
WSA-KNN	6	5	4	5	5
Average no.of features	8	5	7	6	7

The Statlog Heart Dataset which has 270 observations is split into training set with 200 observations and test set with 70 observations.

Based on the subset of features generated by the WSA-LR, WSA-RF, WSA-KNN during the different number of iterations, the training dataset is fed to the classifiers namely Logistic Regression (LR), Random Forest (RF), Support Vector Classifier (SVC) and Gaussian Naive Bayes (GNB) and the Accuracy is tested against test dataset. The accuracies obtained is recorded as follows in table 2, table 3 and table 4.

Table 2: WSA-LR and Average accuracies after classification

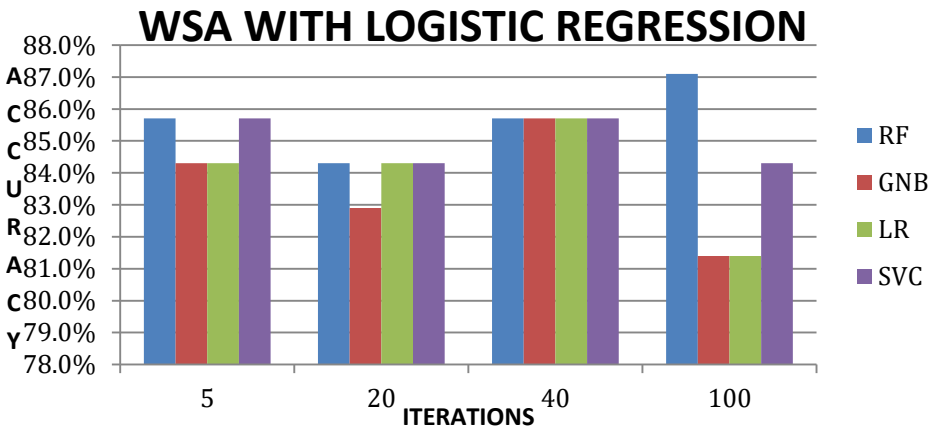
	ITERATIONS				
MODEL	5	20	40	100	Average
RF	85.7%	84.3%	85.7%	87.1%	85.70%
GNB	84.3%	82.9%	85.7%	81.4%	83.58%
LR	84.3%	84.3%	85.7%	81.4%	83.93%
SVC	85.7%	84.3%	85.7%	84.3%	85.00%
Average Accuracy	85.00%	83.95%	85.70%	83.55%	84.55%

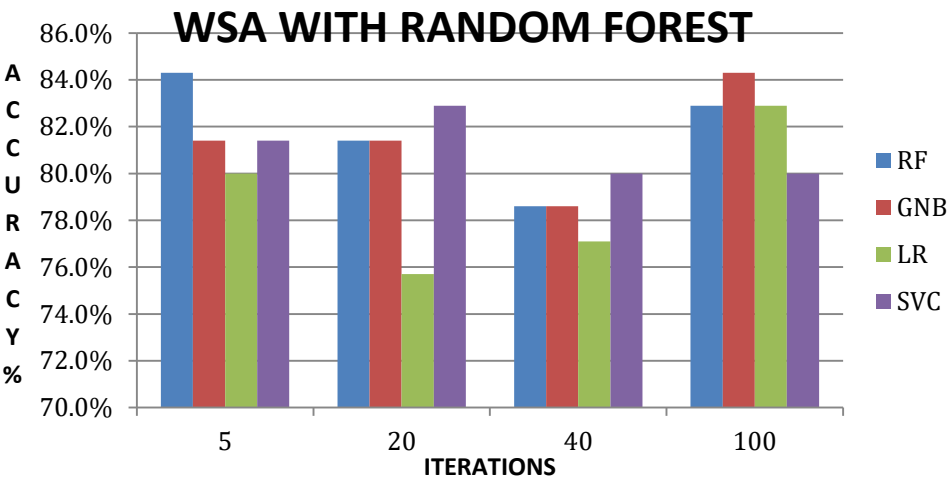
Table 3:WSA-RF and Average accuracies after classification

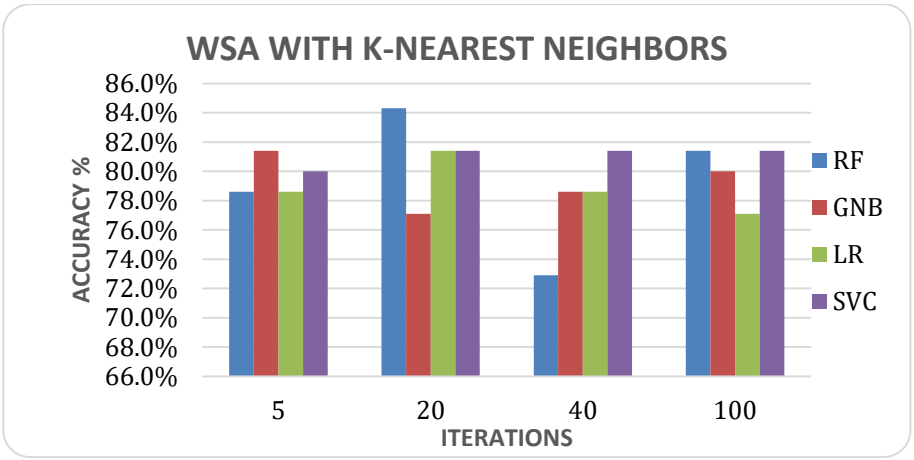
	ITERATIONS				
MODEL	**5**	**20**	**40**	**100**	**Average**
RF	84.3%	81.4%	78.6%	82.9%	81.80%
GNB	81.4%	81.4%	78.6%	84.3%	81.43%
LR	80.0%	75.7%	77.1%	82.9%	78.93%
SVC	81.4%	82.9%	80.0%	80.0%	81.08%
Average Accuracy	81.78%	80.35%	78.58%	82.53%	80.81%

Table 4:WSA-KNN and Average Accuracies after Classification

	ITERATIONS				
MODEL	5	20	40	100	Average
RF	78.6%	84.3%	72.9%	81.4%	79.30%
GNB	81.4%	77.1%	78.6%	80.0%	79.28%
LR	78.6%	81.4%	78.6%	77.1%	78.93%
SVC	80.0%	81.4%	81.4%	81.4%	81.05%
Average Accuracy	79.65%	81.05%	77.88%	79.98%	79.64%







According to the algorithm execution and classification done, WSA-LR with the different classifiers gives an average of 84.6%, WSA-RF with the different classifiers gives an average of 80.8% and WSA-KNN with the different classifiers gives an average of 79.6%.

Among these WSA with Logistic Regression as the fitness function gives a subset of eight features on an average and the accuracy of Random Forest Classifier is found to be 85.7% which is better than the other classifiers.

Conclusion

Feature selection is crucial in high dimensional datasets. Selecting the right set of features has no one particular rule. Recently, swarm intelligence is being used in the search for the best feature subset. In this paper Whale Swarm Algorithm (WSA) is used for the feature selection task. For the fitness function three different classifiers namely, Logistic Regression, Random Forest and K-Nearest Neighbors are used and the subsets are obtained for different number of iterations. The dataset based on the selected subsets are then used for classification. The classification accuracy of four different classifiers namely, Random forest, Logistic Regression, Support Vector Classification and Gaussian Naive Bayes are compared. Among these WSA with Logistic Regression as the fitness function (WSA-LR) gives a subset of eight features on an average and the accuracy of Random Forest Classifier is found to be 85.7% which is better than the other classifiers. In future, other classifiers can be tried as fitness function in the WSA, and the prediction accuracy can be compared among other classifiers.

References

[1] Jie Cai. Jiawei Luo. Shulin Wang. Sheng Yang. 2018. Feature selection in machine learning: A new perspective. Neurocomputing, 300:70–79.

[2] Ah. E. Hegazy. M. A. Makhlouf. Gh. S. El-Tawel. 2018. Dimensionality Reduction Using an Improved Whale Optimization Algorithm for Data Classification. I.J. Modern Education and Computer Science, 7: 37-49.

[3] Hoda Zamani. Mohammad-Hossein. Nadimi-Shahraki. 2016. Feature Selection Based on Whale Optimization Algorithm for Diseases Diagnosis. International Journal of Computer Science and Information Security

(IJCSIS), 14(9).

[4] Z, Zhao. F, Morstatter. S, Sharma. S, Alelyani. A, Anand. H, Liu. 2010. Advancing Feature Selection Research. ASU Feature Selection Repository, 1–28.

[5] P, Langley. 1996. Selection of relevant features in machine learning. Proceedings of the AAAI, Fall.

[6] P. Langley. Elements of Machine Learning. Morgan Kaufmann Series.

[7] Bing Zeng. Liang Gao. Xinyu Li. 2017. Whale swarm algorithm for function optimization. LNCS, 10361: 624-639

[8] Yuefeng Zheng. Ying Li. Gang Wang. Yupeng Chen. Qian Xu. Jiahao Fan. Xueting Cui. 2019. A Novel Hybrid Algorithm for Feature Selection Based on Whale Optimization Algorithm. IEEE Access, 7: 14908-4923.

[9] Lucija Brezo cnik. Iztok Fister Jr. Vili Podgorelec. 2018. Swarm Intelligence Algorithms for Feature Selection -A Review. Applied Sciences, 8:1521.

[10] Machine Learning with MATLAB. https://in.mathworks.com/solutions/machine-learning.

[11] Majdi M, Mafarja. Seyedali Mirjalili. 2017.Whale Optimization Approaches for Wrapper Feature Selection. Applied Soft Computing, 62:441 – 453.

[12] B Subanya. R R Rajalaxmi. 2014. A Novel Feature Selection Algorithm for Heart Disease Classification. International Journal of Computational Intelligence and Informatics, 4(2).

[13] Tad Gonsalves. Feature Subset Optimization through the Fireworks Algorithm. International Journal of Electronics and Computer Science Engineering. ISSN : 2277-1956.

[14] P,Mohapatra. S, Chakravarty. P K, Dash. 2014. Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. Elsevier.

[15] Ismail Babaoglu. Oguz Findik. Erkan Ulker. 2014. A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine. Expert Systems with Applications, Elsevier, 37: 3177–3183.

[16] Dua.D. Graff, C. UCI Machine Learning Repository, Irvine, University of California, School of Information and Computer Science.

[17] https://towardsdatascience.com/getting-data-ready-for-modelling-feature-engineering-feature-selection-dimension-reduction

[18] https://www.who.int/cardiovascular_diseases/en/

[19] Fabian Pedregosa. Gael Varoquaux. Alexandre Gramfort. Vincent Michel. Bertrand Thirion. Olivier Grisel, Mathieu Blondel. Peter Prettenhofer. Ron Weiss. Vincent Dubourg. Jake Vanderplas. Alexandre Passos. David Cournapeau. Matthieu Brucher. Matthieu Perrot. Édouard Duchesnay. 2011. Sci kit-learn: Machine Learning in Python. JMLR, 12, 2825-2830.