ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 652-657
doi: 10.18178/wcse.2019.06.096

Flower Pollination Algorithm and Multilayer Perceptron Artificial Neural Network for Heart Disease Feature Selection and Classification

Nasiru Muhammad Dankolo[1+], Danlami Gabi[2], Nor haizan Mohamed Radzi[3], Noorfa Haszlinna Mustaffa[4] and Roselina Sallehuddin[5]

[1,2] Department of Computer Science, Kebbi State University of Sci. & Tech., Aliero, Nigeria
[3,4,5] Department of Computer Science, Universiti Teknologi Malaysia, Malaysia

Abstract. Heart disease or scientifically known as cardiovascular disease (CVD) is a disease that involves the heart or blood vessels. There are different types of heat diseases and their causes, however the most common one is myocardial infection commonly refers to as heart attack. There are many reasons for heart attack that may be avoidable such as lack of physical fitness and obesity but the unavoidable one is genetic reason. To avoid the serious effect of heart attack and lower the danger of heart failure to patients, early detection of myocardial infection is necessary. Machine learning algorithms such as classification are used in early detection of dieses using historic medical data. Many algorithms are developed for early detection of heart disease, however, because myocardial infection data consists of many features which some of them may not be important to the analysis, there is need to try different alternatives and techniques to come up with the best detection algorithm. In this paper, we proposed flower pollination algorithm and Multilayer perceptron (MLP) Artificial Neural Network (ANN) for feature selection and prediction of myocardial infection. We called this algorithm FPA-ANN. The simulation results of this paper show that FPA-ANN is promising in correct prediction of myocardial infection with 84.2% accuracy.

Keywords: Feature Selection, Classification, Heat Disease

1. Introduction

Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels. Cardiovascular disease includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack) Other CVDs include stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, heart arrhythmia, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, peripheral artery disease, thromboembolic disease, and venous thrombosis.

The underlying mechanisms vary depending on the disease in question. Coronary artery disease, stroke, and peripheral artery disease involve atherosclerosis. This may be caused by high blood pressure, smoking, diabetes, lack of exercise, obesity, high blood cholesterol, poor diet, and excessive alcohol consumption, [1] among others. High blood pressure results in 13% of CVD deaths, while tobacco results in 9%, diabetes 6%), lack of exercise 6% and obesity 5% Rheumatic heart disease may follow untreated strep throat.

It is estimated that 90% of CVD is preventable. Prevention of atherosclerosis involves improving risk factors through: healthy eating, exercise, avoidance of tobacco smoke and limiting alcohol intake. Treating risk factors, such as high blood pressure, blood lipids and diabetes is also beneficial. Treating people who have strep throat with antibiotics can decrease the risk of rheumatic heart disease.

+ Corresponding author. Tel.: + 2348064918672.
E-mail address: nasirdankolo@gmail.com.

However, the estimated 10% of the CVD that are unavoidable (genetic reasons) requires early detection and control. This is where machine learning is of great important. Machine learning algorithms can be train using historic data from known CVD patients and learn from the data to predict the chances of CVD on another patient [2].

Microarray technology is an influential innovation which can be used for disease detection in bioinformatics particularly in cancer detection and diagnosis [3]. The categorisation of the gene expression data is now becoming a central focus to many of researchers in machine learning for bioinformatics data [2]. Using different gene expression forms with normal expression profile, irregularity could be recognized and treated before it develops abnormalities in the patient [4]. The major problem in managing microarray data is the size of its dimension and small sample size [5]. The feature size of microarray data is very vast, which mostly due to the incidence of noisy or unsuitable features that are recorded during the observation, therefore, learning algorithm's performance will significantly be affected if they are to learn on the whole datasets. To address the effect of irrelevant features (genes), feature selection methods are employed. Feature selection works by finding optimal subset of features that can best represent the original features without degradation in performance [4]. Different kind of feature selection algorithms were proposed to scale down the dimension of the features generated by the microarray including the metaheuristic-based search algorithms [6]. In this paper, we are going to develop a heart disease feature selection and classification algorithm based on flower pollination algorithm (FPA) and multilayer perceptron Artificial Neural Network (ANN).

2. Literature Review

The application of machine learning techniques to solve problems relating to knowledge discovery has increased in different fields and areas for many years ago. Since there is a rapid increase in the number of area of applications for some sort of intelligence-based decision process. Generally, the standard structural flow for handling machine learning decision tasks are grouped into three steps: step one is to preprocess the data to be used for the analysis which includes feature selection, step two is to train the learning algorithm to be used for the analysis such as classification algorithms and step three to assess the effectiveness of the applied algorithm [7]. However, the most critical step that determine the effectiveness of the learning algorithms is the preprocessing step, feature selection to be specific. The function of feature selection is to scan over the dataset and return only those features (attributes) that can be used by the learning algorithm to efficiently and effectively solve the given task. In some cases, the accuracy of the learning algorithms does not change even after feature selection, however, the cost of analyzing irrelevant and noisy attributes is relief.

Feature selection has been proven to be effective and efficient in preparing high-dimensional data for data mining and machine learning problem [8]. The objective of this process is to identify and remove irrelevant features from the training dataset. Thus, it increases the performance of prediction algorithms by providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. It attempt to obtain an optimal feature sets from a problem domain while keeping an appropriately high accuracy in representing the original features (Yang et al. 2008). Feature selection process consists of two main components: 1) searching procedure that searches the solution vector, and 2) the evaluation of the searched features. Search methods approach that are widely used in the algorithm include complete, heuristic features [5], and random [9] search. Whilst, the techniques used for feature evaluation are categorized into two either classifiers specific or classifier independent [10]. The classifiers specific requires a learning technique that will be employed to evaluate the quality of feature selection based on the classifier accuracy [11]. However, the classifier independent theories a classifier independent procedure for evaluating the features importance, this kind of measures include mutual information gain, dependence measure and consistency measure [5].

Furthermore, many meta-heuristics algorithms from the evolutionary and swarm intelligent category were applied in the literature to solve feature selection, however, there is still a room of improvement that need to be made. This is due to the complexity and the nature of the high dimensionality of bioinformatic data and the requirement for finding the optimal solution with minimum computational cost [12].

Many metaheuristics algorithms for feature selection such as Particle swam Optimization (PSOA) [13], Ant colony optimization algorithm [14], Artificial Bee colony (ABC) [15].

Flower pollination algorithm (FPA) is inspired from the rules of reproduction process of flowering plants by Yang in 2013 as shown in (fig. 2.1) below. The flower pollination algorithm (FPA) is mainly for optimization, it is applied in feature selection and other optimization tasks to reduce the dimensionality of the search space. In feature selection task, we need a fitness function that will test the fitness of the return features by the FPA. This fitness function may be classifier dependent like K-Nearest Neighbor (KNN), Support Vector Machine (SVM) that use the classifying accuracy of a feature to decides it relevance or classifier independent like entropy that decides the feature important by calculating the amount of information gained before and after splitting the classes using information theory.

Algorithm (or simply Flower Algorithm)

Objective min or max f(x), x = (x1,x2, ...,xd)
Initialize a population of n flowers/pollen gametes with random solutions
Find the best solution g∗ in the initial population
Define a switch probability p ∈ [0, 1]
while (t <Max_Itration)
for i = 1 : n (all n flowers in the population)
if rand < p,
Draw a (d-dimensional) step vector L which obeys a Lévy distribution
Global pollination via $x_i^{t+1} = x_i^t + L\left(x_i^t - g\right)$
else
Draw ϱ from a uniform distribution in [0,1]
Randomly choose j and k among all the solutions
Do local pollination via $L \sim \frac{\lambda\Gamma(\lambda)\sin(\pi\lambda/2)}{\pi}\frac{1}{s^{1+\lambda}}, \quad (s \gg s_0 > 0)$
end if
Evaluate new solutions If new solutions are better, update them in the population
end for
Find the current best solution g∗
end while

Figure 1: Flower Pollination Algorithm (FPA)

where $\mathbf{x}_i^{t+1}$ is the i[th] gamete during the t[th] iteration, *e* is the probability function for switching from global to local pollination drawn from normal probability distribution p[0,1] and x_j and x_k are any random flowers.

After useful features are selected from the original dataset, the actual machine learning task is performed such as classification, clustering, regression and so on. Classification is a machine learning task that categorized objects into predefined classes. Multi-Layer perceptron (MLP) Artificial Neural Network (ANN) for classification is a machine learning algorithm that is effective for classification of bioinformatic dataset [8]. ANN has been experimentally tested for classification in different domain and produce excellent results more than many classification algorithm [16].

In this research, we propose a new method of feature selection and classification of heart disease using high dimensional dataset based on flower pollination algorithm (FPA) and Multilayer Perceptron (MLP) Artificial Neural Network (ANN).

3. Methodology

In this research we proposed a new algorithm for feature selection and classification of Cardiovascular disease (CVD). Using high dimensional data. The dataset was collected from the UCI machine learning repository website which is a well-known public data repository for high dimensional genome dataset. We used the Cleveland heart disease dataset from the UCI website that consist of 76 features, 303 instances and four classes.

We apply flower pollination algorithm (FPA) to perform feature optimization on this high dimensional dataset. The result from the FPA algorithm (number of selected features) is then pass to the Multilayer Perceptron (MLP) artificial neural network for classification of CVD. The performance of the entire algorithm is validated using k-fold cross validation with k=10. The experiment is conducted using mathlab 2017a. The result of the simulation is expressed using confusion matrix.

4. Experiment

In this paper, the proposed FPA-ANN algorithm is developed and tested using heart disease datasets available at UCI machine learning repository. The experiment is conducted using mathlab R2017b. we run the experiment using different parameter setting in order for to achieve the best tune that produce the best result. The experimental setting for this algorithm is given in table 1and 2 below.

Table 1: FPAparameter Settings

S/N	Parameter	Setting
1	Number of population	20
2	Number of iterations	400
3	Probability switch function	0.8

Table 2: ANN parameter Settings

S/N	Parameter	Settings
1	Train Function	trainlm
2	Hidden layer size	30
3	Data partitions	70/30

5. Results

In this section, we present the result obtained after executing our experiment for the proposed FPA-ANN algorithm using heart disease dataset. The flower pollination algorithm (FPA) when applied for feature selection returned 14 best features at the end of the last iteration. These 14 features are used to conduct classification using MLP ANN algorithm. The result is presented in confusion matrix.

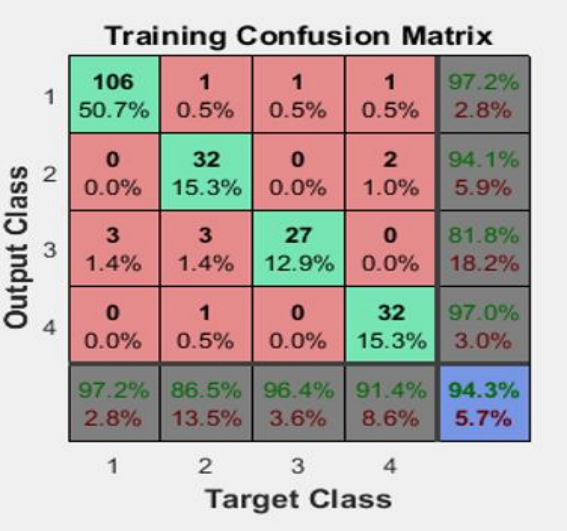

Fig. 2: ANN Training Confusion Matrix

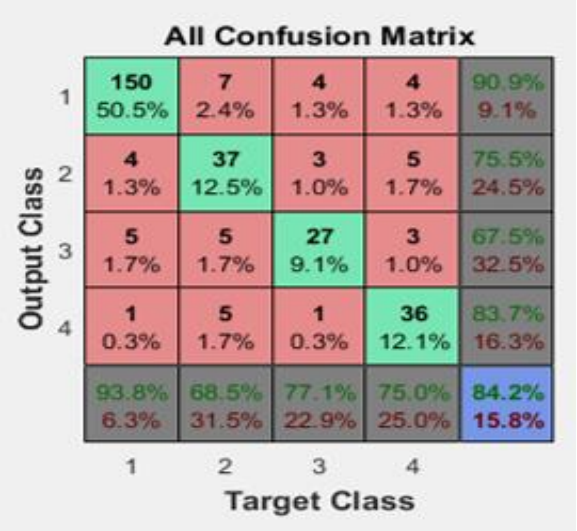

Fig. 3: ANN Testing Confusion Matrix

The result of training ANN for classification of heart disease shows that ANN can achieve an accuracy of 94.3% while the average result of training and testing the ANN for classification of heart disease which is the overall performance of the model shows that ANN can achieve an accuracy of 84.2%.

6. Conclusion

This research work revealed an alternative technique for features selection and classification of heart disease and also reveals the capability of flower pollination algorithm and multilayer perceptron MLP Artificial Neural Network ANN for heart disease feature selection and classification. The results obtained from the experiment proved that FPA-ANN is a good algorithm for early detection of heart disease.

7. References

[1] Alshamlan, Hala M, Ghada H Badr, and Yousef A Alohali. 2016. "ABC-SVM : Artificial Bee Colony and SVM Method for Microarray Gene Selection and Multi Class Cancer Classification." 6(3): 184–90.

[2] Blum, Avrim L., and Pat Langley. 1997. "Selection of Relevant Features and Examples in Machine Learning." *Artificial Intelligence* 97(1–2): 245–71.

[3] Canul-Reich, Juana, Lawrence O. Hall, Dmitry Goldgof, and Steven A. Eschrich. 2008. "Feature Selection for Microarray Data by AUC Analysis." *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*: 768–73.

[4] Dhaenens, Clarisse. 2010. "Metaheuristics for Bioinformatics." : 1–90.

[5] Diao, Ren, and Qiang Shen. 2015. "Nature Inspired Feature Selection Meta-Heuristics." *Artificial Intelligence Review* 44(3): 311–40. http://dx.doi.org/10.1007/s10462-015-9428-8.

[6] Duda, R.O., P.E. Hart, and D.G. Stork. 1997. "Pattern Classification." (April).

[7] Gütlein, Martin, Eibe Frank, Mark Hall, and Andreas Karwath. 2009. "Large-Scale Attribute Selection Using Wrappers." *2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 - Proceedings*: 332–39.

[8] Hira, Zena M., Duncan F. Gillies, Zena M. Hira, and Duncan F. Gillies. 2015. "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data." *Advances in Bioinformatics* 2015(1): 1–13. http://www.hindawi.com/journals/abi/2015/198363/.

[9] Kaur, Parneet, Manpreet Singh, and Gurpreet Singh Josan. 20 "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector." *Procedia Computer Science* 57: 500–508. http://dx.doi.org/10.1016/j.procs.2015.07.372.

[10] Rathasamuth, Wanthanee, and Supakit Nootyaskool. 2016. "Comparison Solving Discrete Space on Flower Pollination Algorithm, PSO and GA." *2016 8th International Conference on Knowledge and Smart Technology, KST 2016*: 18–21.

[11] Shi, Bingbing et al. 2016. "Recent Advances on the Encoding and Selection Methods of DNA-Encoded Chemical Library." *Bioorganic & Medicinal Chemistry Letters* 27(3): 361–69. http://www.sciencedirect.com/science/article/pii/S0960894X16312926?dgcid=raven_sd_aip_email.

[12] Tabakhi, Sina, Ali Najafi, Reza Ranjbar, and Parham Moradi. 2015. "Gene Selection for Microarray Data Classification Using a Novel Ant Colony Optimization." *Neurocomputing* 168: 1024–36.

[13] Thorne, C. J. et al. 2017. "E-Learning in Advanced Life Support—What Factors Influence Assessment Outcome?" *Resuscitation* 114(December 2015): 83–91. http://dx.doi.org/10.1016/j.resuscitation.2017.02.014.

[14] Wang, Lipo, Yaoli Wang, and Qing Chang. 2016. "Feature Selection Methods for Big Data Bioinformatics: A Survey from the Search Perspective." *Methods* 111: 21–31. http://dx.doi.org/10.1016/j.ymeth.2016.08.014.

[15] Wang, Yali, and Brahim Chaib-draa. 2016. "KNN-Based Kalman Filter: An Efficient and Non-Stationary Method for Gaussian Process Regression." *Knowledge-Based Systems* 114: 148–55.

http://dx.doi.org/10.1016/j.knosys.2016.10.002.

[16] Yang, Cheng-San, Li-Yeh Chuang, Chang-Hsuan Ho, and Cheng-Hong Yang. 2008. “Microarray Data Feature Selection Using Hybrid GA-IBPSO.” *Trends in Intelligent Systems and Computer Engineering* 6: 243–53. http://dx.doi.org/10.1007/978-0-387-74935-8_18.