Report

on

Feature Selection Method Based on Grey Wolf Optimization for Coronary Artery Disease Classification

Submitted By

Rishu Raj (MIT2018050)

Shubham Patre (MIT2018062)

Kaushal Sahu (MIT2018072)

Submitted to

Dr. Vrijendra Singh

Abstract

Cardiovascular disease also called heart disease are heart conditions that include diseased vessels, structural problems and blood clots. One of the most common cardiovascular disease is coronary heart disease also known as Coronary Artery Disease (CAD). CAD is damage or disease in heart's major blood vessels. Various Computational Intelligence (CI) techniques have been quite effective in providing insights about such deadly diseases by analyzing huge datasets. The datasets involved in the process may contain irrelevant features and redundant data which affects the performance and accuracy of the CI techniques involved. Therefore we need to apply feature selection techniques to eliminate such redundancies and irrelevance form datasets. In this report we propose a feature selection method using Grey Wolf Optimization (GWO) to determine the optimal feature subset for diagnosing coronary artery disease. Our proposed method consists of two stages feature selection and classification of data. In first stage we find the best features for disease using Grey Wolf Optimization and in second stage the fitness function of GWO is evaluated with help of classification technique. For classification purpose we use Support Vector Machine (SVM). Cleveland Heart disease dataset is use for performance evaluation of the proposed method.

Introduction

In recent years, advancements in electronical medical records have been remarkable. Using this electronic medical records with help of Computational Intelligence (CI) techniques doctors can be provided with insights that can assist them in better decision making process which is relatively faster and accurate than manual process. Heart disorders are also called Cardiovascular Disease (CVD) are diseases that describe narrowed or blocked blood vessel that can lead to heart attack or failure [1]. One of the most common Cardiovascular Disease is Coronary Heart Disease (CAH) or Coronary Artery Disease (CAD) which is generally result of cholesterol or fatty deposition on internal walls of arteries.

Coronary Artery Disease (CAD) prediction using Computation techniques is generally carried out in two stages. First stage is feature selection stage and second is classification stage. In feature selection stage we select the most useful features using CI techniques eliminating the redundant and irrelevant data which generally leads to inaccuracy and error in predictions. Feature selection also helps in reduction of computational complexity [2].

Currently various feature selection process has been proposed but for the purpose of this paper we will be using Grey Wolf Optimization (GWO) technique to extract the most optimal features in disease identification. For classification purpose also several methods have been proposed like Naive Bayes, Linear Regression, Neural Networks (NN), SVM and Fuzzy classifier [4, 5]. Form all the classification methods stated above SVMs have been reported to produce results with highest accuracy. Therefore in this paper we have used Grey Wolf Optimization for feature selection and Support Vector Machine classifier for classification purpose namely GWO-SVM.

Gray Wolf Optimization (GWO)

GWO is inspired by social hierarchy and the hunting approach of grey wolves proposed by [5]. Grey wolves typically prefer to live in a pack of 5–12 individuals and have a strict social hierarchy. As shown in Fig. 1, GWO Consist of four levels as follows:

(1) Alpha (a): male and female are the leaders of a pack of wolves that are responsible for making decisions such as wake-up time, hunting, and sleep place.

(2) Beta (b): either male or female wolves, beta probably the best candidate of replacement for alpha. Assisting a in decisions making and suggesting feedbacks are the main roles of b.

(3) Delta (d): The wolves at this level obey a and b wolves and control x wolves.

Delta acts as sentinels, scouts, elders, sentinels, caretakers in the pack, and hunters.

(4) Omega (x): the wolves at this level are the weakest. Omega (x) plays a role of scapegoat. Omega (x) should obey other individuals' orders.

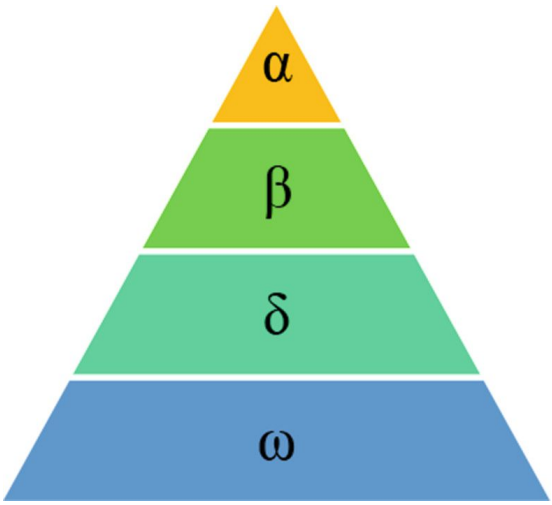

Fig. 1. Grey wolves' social hierarchy represented by [5].

In order to mathematically simulate the encircling behavior of grey wolves, the following equations are proposed:

$$\vec{D} = \vec{C} \cdot \vec{X}p(t) - \vec{X}(t),$$

$$\vec{X}(t+1) = \vec{X}p(t) - \vec{A} \cdot \vec{D},$$

where t indicates the current iteration, $\vec{A} = 2\vec{a} \cdot \vec{r}1 - \vec{a}$, $\vec{C} = 2\vec{r}2$, $\vec{X}p$ is the position vector of the prey, $\vec{X}$ is the position vector of a grey wolf, $\vec{a}$ is linearly decreased from 2 to 0, and $\vec{r}1$ and $\vec{r}2$ are random vectors in [0, 1].

In order to mathematically simulate the hunting behavior of grey wolves, the following equations are proposed:

$$\vec{D}\alpha = \vec{C}1 \cdot \vec{X}\alpha - \vec{X},$$

$$\vec{D}\beta = \vec{C}2 \cdot \vec{X}\beta - \vec{X},$$

$$\vec{D}\delta = \vec{C}3 \cdot \vec{X}\delta - \vec{X},$$

$$\vec{X}1 = \vec{X}\alpha - \vec{A}1 \cdot \vec{D}\alpha,$$

$$\vec{X}2 = \vec{X}\beta - \vec{A}2 \cdot \vec{D}\beta,$$

$$\vec{X}3 = \vec{X}\delta - \vec{A}3 \cdot \vec{D}\delta,$$

$$\vec{X}(t+1) = (\vec{X}1 + \vec{X}2 + \vec{X}3)/3.$$

Finally, the trade-off between exploration and exploitation is controlled by the updating of the ~a parameter. In each iteration ~a parameter is updated linearly to range from 2 to 0 as according to the equation below:

a=2-t*(2/maxiter)

Where maxiter indicates the total number of iterations allowed for the optimization and t is the number of iteration.

Proposed Methodology

In this paper we used Grey Wolf Optimization [5] along with Support Vector Machine (GWO-SVM) classifier for effective feature selection to help with diagnosis of Coronary Artery Disease (CAD). GWO-SVM works in two stages where in the first stage we use GWO for effective feature selection from the Cleveland heart dataset. Initially GWO produce the initial positions of the population and then current positions with each iteration until a stopping criteria is satisfied. In second stage then SVM is used for classification on the optimal feature subset obtained from the first stage.

Table 1 shows the parameter setting used for the proposed method. Where the iterations, wolves, dimension numbers and search domain are identified. a and b parameters for the fitness function are declared.

Table 1. Parameter setting for the proposed method

Parameter	Numbers
Iterations no.	100
Wolves no.	5
Dimensions no.	14
Search domain	[0 1]
α in fitness function	0.99
β in fitness function	0.01

Table 2 shows the details of the features of dataset. It has thirteen features and one target variable. GWO gives best features out of 13 to make SVM more accurate. Each time the GWO operates the features changes and accordingly the accuracy also varies.

Table 2. Attributes of Cleveland dataset

No	Attributes	Description
1	Age	Age in year
2	Sex	0 for female and 1 for male
3	Cp	Chest pain type Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
4	Trestbps	Resting blood sugar in mm Hg on admission to the hospital
5	Chol	Serum cholesterol in mg/dl
6	Fbd	(Fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	Restecg	Resting ECG result
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope or peak exercise ST segment
12	Ca	Number of major vessels colored by fluoroscopy
13	Thal	Defect type
14	num	The predicted attribute

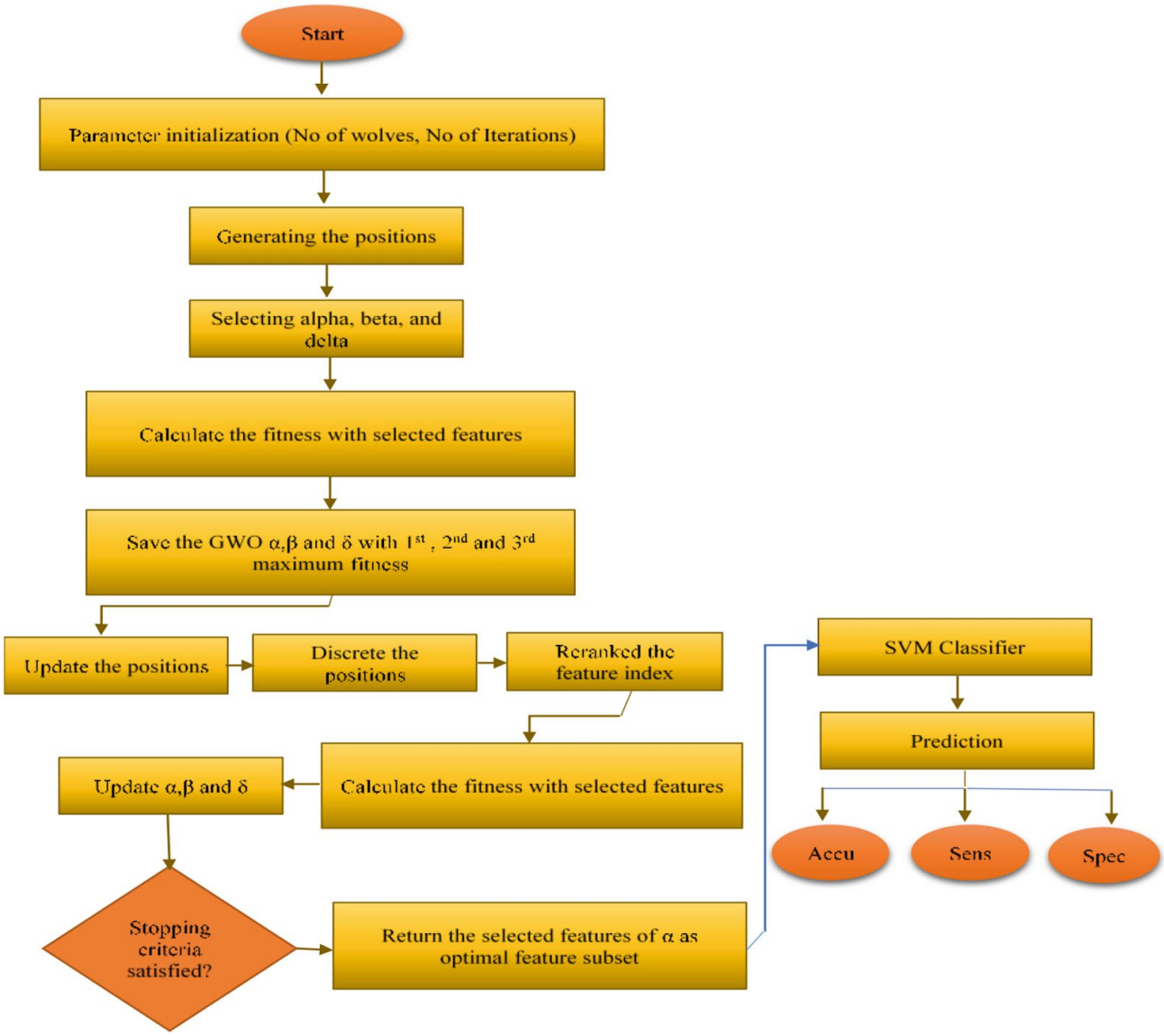

Fig:- Proposed Feature Selection Method

Dataset Used

Cleveland dataset freely available and can be downloaded from UCI repository.

Results

After using the GWO-SVM on the given dataset 10 times, different features were obtained and the accuracy obtained on the algorithm using GWO-SVM was 61.4%. While SVM has accuracy 59.8%, which was obtained when applied on the feature subset. Several parameters like number of wolves, benchmark function matters for the accuracy achieved. The open source code for the GWO has been used for writing the entire code of the algorithm GWO-SVM.

References

1. Krishnaiah, V., Narsimha, G., Chandra, N.S.: Heart disease prediction system using data mining technique by fuzzy K-NN approach. In: Emerging ICT for Bridging the Future Proceedings of the 49th Annual Convention of the Computer Society of India (CSI), vol. 1, pp. 371–384 (2015)
2. Shilaskar, S., Ghatol, A.: Feature selection for medical diagnosis: evaluation for cardiovascular diseases. Expert Syst. Appl. 40(10), 4146–4153 (2013)
3. Srinivas, K., Rao, G.R., Govardhan, A.: Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In: 2010 5th International Conference on Computer Science and Education (ICCSE), pp. 1344–1349 (2010)
4. 21. Das, R., Turkoglu, I., Sengur, A.: Effective diagnosis of heart disease through neural networks ensembles. Expert Syst. Appl. 36(4), 7675–7680 (2009)
5. Mirjalili, S., et al.: Grey Wolf Optimizer. Adv. Eng. Softw. 69, 46–61 (2014)
6. Cleveland dataset. http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/ processed.cleveland.data. Accessed 27 May 2018
7. Al-Tashi, Qasem & Rais, Helmi & Jadid Abdulkadir, Said. (2018). “Feature Selection Method Based on Grey Wolf Optimization for Coronary Artery Disease Classification.” 257-266. 10.1007/978-3-319-99007-1_25.