AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

ISSN:1991-8178 EISSN: 2309-8414
Journal home page: www.ajbasweb.com

Grey Wolf Optimization and Naive Bayes classifier Incorporation for Heart Disease Diagnosis

Lamiaa M. El Bakrawy

Al-Azhar University Faculty of Science, Cairo, Egypt.

Address For Correspondence:
Lamiaa M. El Bakrawy. Al-Azhar University Faculty of Science, Cairo, Egypt.

ARTICLE INFO

Article history:
Received 18 February 2017
Accepted 5 May 2017
Available online 10 May 2017

Keywords:
Heart Disease, Diagnosis, Naive Bayes classifier, Grey Wolf Optimization.

ABSTRACT

Background: Heart disease is a significant health problem and many people are influenced by it. It can be defined as any type of disorder that affects the heart so diagnosis of heart disease is an important task which must be executed in an efficient way. The Heart disease diagnosis is used to detect the presence or absence of heart disease in the patient. Objective: A new Heart disease diagnosis algorithm is proposed by integrating Grey Wolf Optimization (GWO) and Naive Bayes classifier (NB), named by (GWO-NB). The attributes of heart disease data is discretized using CAIM (Class-Attribute Interdependence Maximization) method in order to increase the accuracy of independent classifiers. The grey wolf optimization is used to determine the weights of attributes of Naive Bayes classifier automatically and then these weights are used to maximize NB's classification accuracy. The proposed algorithm has been implemented by applying 5-fold cross-validation on the Cleveland Heart Database obtained from the UCI Machine Learning. Conclusion: Experimental results demonstrate that the proposed (GWO-NB) algorithm achieved better classification accuracy than the traditional Naive Bayes classifier algorithm. Also, it shows that the proposed algorithm outperforms the performance of other techniques for the diagnosis of heart disease.

INTRODUCTION

The heart is one of the strongest muscles in the body which pumps blood, giving nutrients and oxygen to all parts of the body. The heart is a part of body which playing an important role in Human life. Human life is completely dependent on the proper operation of the heart which will affect the vital organs of human body such as brain and kidney. There are different forms of heart disease such as inflammatory heart disease, rheumatic heart disease and coronary heart disease Vijayashree and SrimanNarayanaIyengar (2016) and Milan and Godara (2011). According to the World Health Organization (WHO), the USA's Centers for Disease Control and Prevention (CDC) and the World Heart Federation (WHF), heart disease and stroke are a leading cause of death for about 20 million human in the world and this number will reach 24 million in year 2030 Moloud (2015). Misdiagnosis of heart disease is one of the major factors that causes the rising number of deaths. In order to avoid misdiagnosis, the physicians have focus more in designing effective techniques which can be used to diagnose heart diseases with high accuracy Masethe and Masethe (2014). Diagnosis of heart diseases is a significant in the medical field. There is a multi layered issue to detect heart disease from different symptoms which is not free from false hypothesis. Thus, experience and knowledge of specialists and clinical screening data of patients are used to facilitate the diagnosis process. The main constraint encountered by the health care organizations is giving quality services with an affordable costs. Quality service provides the patients with appropriate treatment and indicates the accurate diagnosis. As a result, different data mining techniques have

ToCite ThisArticle: Lamiaa M. El Bakrawy.,Grey Wolf Optimization and Naive Bayes classifier Incorporation for Heart Disease Diagnosis. ***Aust. J. Basic & Appl. Sci.,*** *11(7): 64-70, 2017*

Australian Journal of Basic and Applied Sciences, 11(7) May 2017, Pages: 64-70

become critical to the medical healthcare world in achieving clinical tests with a reduced cost Subbalakshmi*et al*. (2011) and Hany and Desuky (2014).

In recent years, data mining techniques have been used for diagnosing heart diseases. Moloud(2015) found a better decision tree algorithm and then used it for extracting rules in predicting heart disease. He used Cleveland data, for his study. In his paper, C5.0 algorithm with accuracy value of 85.33% has a better performance compared to the rest of the algorithms used in this study. On the other hand, Jyoti *et al*. (2011) designed a GUI based Interface to enter the patient record and predict whether the patient is having Heart disease or not using Weighted Association rule based Classifier. Experimental results showed that Weighted Associative Classifier was providing improved accuracy as compare to other already existing Associative Classifiers. Moreover, Rupali and Patil (2014) proposed approach that had developed a Decision Support in Heart Disease Prediction System (HDPS) using data mining technique called Nave Bayes. They used medical profiles such as blood pressure, age, sex. The algorithm can predict the likelihood of patients getting a heart disease. They implemented their algorithm in Matlab as an application which takes medical tests parameter as an input. Their algorithm can be used as a training tool to diagnose patients with heart disease.

In this paper we propose Grey Wolf Optimization algorithm to improve the classification accuracy of the traditional Naive Bayes classifier in the field of medical data especially diagnosis of heart disease. The rest of this paper is organized as follows: Brief introduction of Naive Bayes classifier and Grey Wolf Optimization algorithms are introduced in Section 2. Section 3 describes in details the proposed algorithm. Section 4 presents data set and experimental results. Finally, conclusions are discussed in Section 5.

2. Preliminaries:

This section provides a brief overview of Naive Bayes classifier and Grey Wolf Optimization algorithms with some of the key definitions.

2.1. Naive Bayes classifier:

Naive Bayes classifier is widely used in many different fields such as medical diagnosis, weather forecasting and classification because of its robustness, elegance and simplicity Shweta and Soni (2016) and Amirjahan and Sujatha (2016). Naive Bayes classifier gets its name from being depend on Naive and Bayes. Naive comes from the assumption that all conditional probabilities are mutually independent and Bayes' theorem is used for the Bayes rule of conditional probability Omar *et al*. (2013).

Naive Bayes classifier was designed considering the fact that there is no relationship between the presence or absence of a particular feature of a class and the presence or absence of any other feature. According to this specific nature of the probability model, naive Bayes classifiers can be used efficiently in a supervised learning Smita and Kumar (2015).

The probability model for a naive Bayes classifier is a conditional model, given a problem instance to be classified, represented by a vector $F = (F_1, F_2, \dots, F_n)$ and n represents the independent features, then the instance probabilities is $p\ (C_k|F_1, F_2, \dots, F_n)$, k is possible classes C_k.

The conditional probability can be written by using Bayes' theorem as

$$p\ (C_k|\ F)\ = \frac{p(C_k).p\ (F\ |C_k)}{p\ (F)} \tag{1}$$

The above equation can be written as

posterior = prior × likelihood / evidence (2)

This means that the denominator (evidence) does not depend on *C* and the values of the features, so that the evidence is constant and the conditional distribution over the class variable can be calculated as

$$p\ (C_k|F_1, F_2, \dots, F_n) = \frac{1}{z}\ p(C_k)\ \prod_{i=1}^{n} p\ (F_i|C_k) \tag{3}$$

Where the evidence z is a scaling factor dependent only on $(F_1, F_2, \dots, F_n)$ which means that it is a constant if the values of the features are known.

2.2. Grey Wolf Optimization:

Grey Wolf Optimization (GWO) was proposed by Mirjalili *et al*. (2014). It is a new metaheuristic technique which can be applied for solving optimization problems. The algorithm simulates the social leadership and hunting behavior of grey wolves. The population of grey wolves includes four groups: alpha (α), beta (β), delta

(δ), and omega (ω). The alpha is making decisions about sleeping place, hunting, time to wake, and so on. The beta helps the alpha in decision-making. The beta wolf is the most likely to replace the alpha wolf in case one of the alpha wolves becomes very old to lead. The delta wolf is called subordinate and consists of sentinels, hunters elders, caretakers and scouts. Scouts observe the boundaries of region and warning the pack from any risk may occur.

The omega wolf is the lowest ranking grey wolf which considers the scapegoat. Alpha, beta and delta are considered the best search agents who guide omega wolves toward to promising regions of the search space Esraa *et al*. (2015). During the hunt process (optimization), the grey wolves encircle their prey and update their positions around α, β, and δ as follows:

$$\mathrm{D} = \left|C.x_p(t) - x(t)\right| \tag{4}$$

$$X(t+1) = X_p(t) - A.D \tag{5}$$

Where t is the current iteration, X_p is the vector of the prey position, and X indicates the vector of the grey wolf position. $A = 2a.r_1 - a$, $C = 2.r_2$, a is linearly decreased from 2 to 0, over the course of iterations and r_1 , r_2 are random vectors in [0, 1].

In the hunting process of grey wolves, the alpha is the best candidate solution, beta, and delta are assumed to have superior knowledge about the potential location of prey.

The three best solutions obtained so far and compel the other search agents to update their positions according to the position of the best search agents. The mathematical model proposed to update the grey wolves positions are as follows

$$D_\alpha = |C_1.X_\alpha - X| \tag{6}$$
$$D_\beta = \left|C_2.X_\beta - X\right| \tag{7}$$
$$D_\delta = |C_3.X_\delta - X| \tag{8}$$
$$X_1 = X_\alpha - A_1.(D_\alpha) \tag{9}$$
$$X_2 = X_\beta - A_2.(D_\beta) \tag{10}$$
$$X_3 = X_\beta - A_3.(D_\beta) \tag{11}$$

$$X(t+1) = \frac{(X_1 + X_2 + X_3)}{3} \tag{12}$$

where X_α, X_β, X_δ are the positions of the alpha, beta and delta respectively. C_1,C_2, C_3, A_1,A_2, A_3 are random vectors and X is the position of the current solution.X_1, X_2, X_3 are the distance between the current solution and alpha, beta and delta respectively and $X(t+1)$ is the final position of the current solution

3. The proposed algorithm (GWO-NB):

In this paper, we use Grey Wolf Optimization (GWO) technique to learn the weight in Naive Bayes classifier (NB). The mechanism of GWO helps us to obtain the optimal weight automatically so we do not need any information about the weight. The aim of this paper is to improve the classification accuracy of the traditional Naive Bayes classifier based on GWO in the field of medical data especially heart database which can be used by healthcare professionals in the diagnosis of heart disease. The proposed algorithm can simulate like human diagnostic expertise for curative of heart ailment.

In the proposed algorithm, GWO is used to promote population of candidate solutions which represent the weights for attributes. Alpha is the best candidate weight which is the one getting the highest fitness function (classification accuracy) used in NB. The major steps are shown in Fig. 1 and the detailed process is described as follows:

1. Initially, the attributes of heart disease data should be discretized. Discretization is a process which converts continuous numeric values into discrete ones to make the attributes values efficient for the learning process Marcin et al. (2008).
2. Cross-validation (5 folds) is used for validating the results. The heart disease data is randomly partitioned into 5 equal sized subsets. One subset of them is used as the test set and the remaining 4 subsets are used as training set. Cross-validation is repeated 5 times and in each time another subset from 5 subsets is used as testing set. Then the validation results over the 5 times is computed in the final phase.
3. Initialize the population for N grey wolves positions (the weight individuals) randomly. Assume that, the number of attributes for the weight individual is K.

A single grey wolf position is $X_i = \{a_{i,j}, j = 1,2, \dots, k\}$ where $a_{i,j}$ is j th weight for the i th individual. For each individual, the value of the weight is k random number in the range from 1 to 10.

4. Initialize the parameters C, A, and a.
5. Calculate the corresponding fitness function for each wolf and select the first three best grey wolves positions and store them as alpha (α), beta (β), and delta (δ). The fitness function is the classification accuracy which is obtained by NB to differentiate the patient and healthy cases correctly.
6. Update the position of omega (ω) wolves using equations from (6) to (12)
7. Update C, A, and a.
8. Go to step 5 if t is less than maximum number of iterations
9. Return the position of α which is the individual vector of best weight
10. The best individual vector is the group of the best weight vector acquired by NB classifier.
11. After acquiring the best weight, we use the GWO-NB algorithm with the best weight learned by GWO to classify the test data

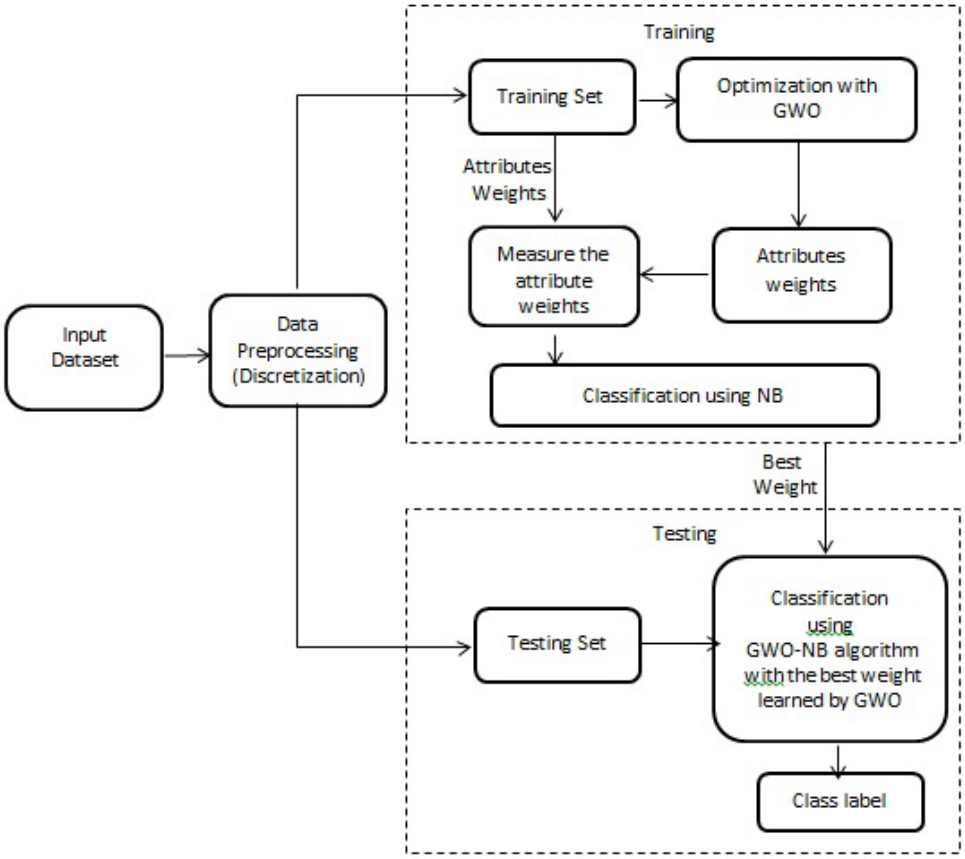

Fig. 1: The proposed algorithm

4. Data set and experimental results:

In this work, we run our experiments using Matlab 15, on a system with a 2.40 GHZ Intel(R) Core(TM)i7 processor and 16 GB of RAM running Microsoft Windows 8 Professional.

In order to evaluate our algorithm we have used the Cleveland Heart Database obtained from the UCI Machine Learning (http://archive.ics.uci.edu/ml/datasets/Heart+Disease). The dataset consists of 303 instances and 14 attributes including the predicted attribute. All attributes are numeric valued and the detailed description of the attributes is shown in Table 1.

In our algorithm, we first apply the discretization CAIM (Class-Attribute Interdependence Maximization) method on Cleveland Heart Data to preserve the highest interdependence between discretized attributes and target class which improve the performance of the NB classifier Marcin *et al*. (2008). Then, the discretized

attributes are randomly divided into training and testing sets by using 5-folds cross validation. For the training, the initial population size was 100 and the maximum number of iterations was 500.

The efficiency of our algorithm is determined using four statistical measures - accuracy, Sensitivity, F-measure and G-mean Karlijn *et al.*(2009). Accuracy is defined as the percentage of observations correctly classified by the algorithm. The ratio of true positive (TP) and true negative (TN) should be calculated in order to estimate the classification accuracy. It can be calculated according to the following equation:

$$\text{Accuracy} = (TN+TP)/(TN+TP+FN+FP) \quad (13)$$

Where FP is false positive and FN is false negative

Sensitivity or true positive rate (TPR) refers to the probability of correctly identifying the existence of heart disease in sick people described by

$$\text{Sensitivity} = TP/(TP + FN) \quad (14)$$

F-measure (F1 score), is defined as the harmonic average of recall and precision. It is measured by

$$\text{F-measure} = 2TP/(2TP + FP + FN) \quad (15)$$

Finally, G-mean is the geometric mean calculated by:

$$\text{G-mean} = \sqrt{TPR \times TNR} \quad (16)$$

Where TNR is the true negative rate is measured by:

$$TPR = TN/(TN + FP) \quad (17)$$

In this part, we show the experiments results when Grey Wolf Optimization technique is applied on Naive Bayes classifier (GWO-NB). Table 2 shows that applying (GWO-NB) without using discretization methods gives better accuracy, Sensitivity, F-measure and G-mean than using traditional Naive Bayes classifier, but, GWO-NB with discretization gives the best accuracy, Sensitivity, F-measure and G-mean.

Table 1: Attributes of Cleveland heart disease dataset

No.	Attribute Name	Description
1	Age	Age in years
2	Sex	1 = male, 0 = female
3	Cp	Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
4	Trestbps	Resting blood sugar (in mm Hg on admission to hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar > 120 mg/dl(1 = true, 0 = false)
7	Restecg	Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy)
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST11 Slope segment (1 = up sloping,2 = flat, 3 = down sloping)
12	Ca	Number of major vessels colored by fluoroscopy
13	Thal	3 = normal, 6 = fixed defect, 7 = reversible defect
14	Num	Class (0 = healthy, 1 = have heart disease)

Table 2: Classification measures for optimized Naive Bayes (GWO-NB)

Algorithm Name	Accuracy	Sensitivity	F-measure	G-mean
NB	75.47 %	63.57 %	72.94 %	74.80 %
GWO-NB without discretization	85.79 %	86.06 %	86.91 %	85.56 %
GWO-NB with discretization	87.45 %	89.70 %	88.59 %	87.09 %

Table 3 shows the comparison between the proposed algorithm and previous traditional algorithms such as RIPPER Milan and Godara (2011), Decision Tree Milan and Godara (2011), ANN (MLP) Milan and Godara (2011)., SVM Milan and Godara (2011), WAC Jyoti*et al*. (2011), Laplace Smoothing Rupali and Patil (2014), LAD Boros *et al*. (2000), C5.0 Moloud (2015), C&R Tree Moloud (2015), CHAID Moloud (2015), and QUEST Moloud (2015). This comparative shows that Grey Wolf Optimization technique applied with Naive Bayes classifier after discretization of all attributes of Cleveland Heart Data gives the best accuracy.

Table 3: Classification Accuracy Comparison for different algorithms

Algorithm Name	Accuracy
RIPPER	81.08 %
Decision Tree	79.05 %
ANN (MLP)	80.06 %
SVM	84.12 %
WAC	81.51 %
Laplace Smoothing	86 %
LAD	83.8 %
C5.0	85.33 %
C & R Tree	60.82 %
CHAID	59 %
QUEST	59.36 %
NB	75.47 %
GWO-NB without Discretization	85.79 %
GWO-NB with Discretization (proposed)	87.45 %

5. Conclusions:

Due to the importance of heart disease diagnosis in the medicine, a new classification approach (GWO-NB) is proposed to detect whether the patient has heart disease or not. In this paper, concentration is on the method to improve the accuracy of diagnosis in the field of Medical Data especially Heart Disease. Initially, Cleveland Heart Database is discretized, and then it is divided into training and testing sets by using 5-folds cross validation. Grey Wolf Optimization is used to identify the best weights of attributes of Naive Bayes classifier. The experimental results showed that the proposed algorithm without discretization gives much higher accuracy, Sensitivity, F-measure and G-mean than the classification technique NB, but the proposed algorithm with discretization accomplished accuracy rate of 87.45% with high value of Sensitivity, F-measure and G-mean. The experiments emphasize that our proposed algorithm has superior efficiency compared with other techniques, considering the accuracy which can help the doctors in heart disease diagnosis.

REFERENCES

Amirjahan, M and N.Sujatha, 2016.Framework of Classification Algorithms. International Journal of Innovative Research in Computer and Communication Engineering, 4: 6173-6178.

Boros, E., P.L. Hammer, T.Ibaraki and A.Kogan,2000.An Implementation of Logical Analysis of Data. IEEE transactions on knowledge and data engineering, 12: 292-306.

Esraa, E., N. El-Bendary, AHassanien and A Abraham, 2015. Grey Wolf Optimization for One-Against-One Multi-class Support Vector Machines, 2015 Seventh International Conference of Soft Computing and Pattern Recognition (SoCPaR 2015), 7-12.

Hany, M.H and A.S. Desuky, 2014. Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization. International Journal of Computer Applications, pp: 94.

Jyoti, S., U. Ansari and D. Sharma, 2011.Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers.International Journal on Computer Science and Engineering (IJCSE),3: 2385-2392.

Karlijn, J., v.Stralen, V.S. Stel, J.B. Reitsma, F.W. Dekker, C.Zoccali and K.J. Jager,2009. Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. Kidney International., 75: 1257-1263.

Marcin, M., L. Kurgan and M.Ogiela, 2008.Comparative analysis of the impact of discretization on the classification with Naive Bayes and semi-Nave Bayes classifiers.Seventh International Conference on Machine Learning and Applications.

Masethe, H and M.A. Masethe, 2014.Prediction of Heart Disease Using Classification Algorithms.Proceedings of theWorld Congress on Engineering and Computer Science, II.

Milan, K and S.Godara, 2011.Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. International Journal of Computer Science and Technology, 2: 304-308.

Mirjalili, S., M.Seyed and L. Andrew, 2014 .Grey wolf optimizer. Advances in Engineering Software, 69: 46-61.

Moloud, A., 2015. Using Decision Trees in Data Mining for Predicting Factors Influencing of Heart Disease . Carpathian Journal of Electronic and Computer Engineering, 8: 31-36.

Omar, S., A.Hassanien, A.Darwish and R.Faraj,2013.A Survey of Machine Learning Techniques for Spam Filtering. IJCSNS International Journal of Computer Science and Network Security, 13: 103-110.

Rupali, M and R.Patil, 2014.Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing.International Journal of Advanced Research in Computer and Communication Engineering, 3.

Shweta, K and S.Soni, 2016.Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection. International Journal of Computer Applications,133: 33-37.

Smita, M and S. Kumar, 2015.Survey on Types of Bug Reports and General Classification Techniques in Data Mining. International Journal of Computer Science and Information Technologies, 6: 1578-1583.

Subbalakshmi, G., K. Ramesh and M. ChinnaRao, 2011. Decision Support in Heart Disease Prediction System using Naive Bayes. Indian Journal of Computer Science and Engineering (IJCSE), 2: 170-176.

Vijayashree, J and N.Ch, SrimanNarayanaIyengar, 2016.Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review. International Journal of Bio- Science and Bio-Technology, 8: 139-148.

http://archive.ics.uci.edu/ml/datasets/Heart+Disease