The Hybrid Machine Learning Model Based on Random Forest Optimized by PSO and ACO for Predicting Heart Disease

Youness Khourdifi[1], Mohamed Bahaj[2]

[1,2] Faculty of Sciences and Techniques, Hassan 1st University, Settat, Morocco
[1,2]{ykhourdifi,mohamedbahaj}@gmail.com

Abstract. In this paper, we used the hybrid Machine Learning model, for proposed PA-RF, a classification based on Random Forest model, optimized by Particle Swarm Optimization (PSO) associated with Ant Colony Optimization (ACO), and we use Fast Correlation-Based Feature Selection (FCBF) method to filter redundant and irrelevant characteristics, in order to improve the quality of heart disease classification. The proposed mixed approach is applied to the heart disease dataset. The results demonstrate the effectiveness and robustness of the proposed hybrid method in processing various types of data for the classification of heart disease. Therefore, this study examines the different automatic learning algorithms and compares the results using different performance measures, i.e. Accuracy, Precision, Recall, F1-Score, etc. The data set used in this study comes from the UCI's automatic learning repository, entitled "Heart Disease" Data set. We can be concluded that PA-RF has demonstrated efficiency and robustness compared to other classification methods.

Keywords: Machine Learning; Heart Disease; Random Forest; Ant Colony Optimization; Particle Swarm Optimization.

1 Introduction

Intelligent optimization algorithms are developed by simulating or revealing certain natural phenomena and are widely used in many research fields because of their versatility [1], [2]. The Particle Swarm Optimization (PSO) algorithm has been successfully applied to heart disease because of its simplicity and generality[3]. However, PSO easily fell into the optimal local solution. In addition, the ACO algorithm was originally introduced for combinatorial optimization. Recently, ACO algorithms have been developed to solve continuous optimization problems. These problems are characterized by the fact that decision variables have continuous domains, unlike discrete problems [4]. Using a single optimization algorithm has the disadvantages of low accuracy and generalizability in solving complex problems. To further explore the application of intelligent optimization in bioinformatics, PSO and ACO are combined in this article, meaning that exploitation and exploration capacity are combined for binary

ICCWCS 2019, April 24-25, Kenitra, Morocco

DOI 10.4108/eai.24-4-2019.2284088

and multi-class heart disease. In this article, the Fast Correlation-Based Feature selection (FCBF) method[5] used to remove redundant and irrelevant features, the results of the PSO optimization are considered the initial values of the ACO, and then the classification model for heart disease is constructed after the parameters are adjusted. In this study, algorithms such as K-Nearest Neighbour (K-NN), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and Artificial Neural Network (MLP) are used. The results demonstrate that the hybrid optimized by FCBF, PSO and ACO method presented in this work is robust and provides more accurate classification results. This work aims to provide heart disease classification results for reference and to contribute to the clinical diagnosis and treatment of different types of heart disease.

The main objective of this article is the prediction of heart disease using the weka data-mining tool and its use for classification in the field of medical bioinformatics. It first classifies the data set and then determines the best algorithm for the diagnosis and prediction of heart disease. Prediction begins by identifying symptoms in patients, then identifying sick patients from a large number of sick and healthy patients. Thus, the primary objective of this paper is to analyze data from a heart disease dataset using a classification technique to accurately predict the class in each case. The main contributions of this paper are:

- **Extraction of classified accuracy useful for heart disease prediction**
- **Remove redundant and irrelevant features with Fast Correlation-Based Feature selection (FCBF) method.**
- **Optimizations with Particle Swarm Optimization PSO then we consider the result of PSO the initial values of Ant Colony Optimization ACO approaches.**
- **Comparison of different data mining algorithms on the heart disease dataset.**
- **Identification of the best performance-based algorithm for heart disease prediction.**

The rest of the paper is arranged as follows: Recent work in this area is discussed in Section 2. Section 3 describes the detailed description of the proposed methodology. Section 4 explains in detail the experiments using the proposed machine learning models. Section 5 presents conclusions and future research directions.

2 Related Works

Several experiments are conducted on medical data sets using multiple classifiers and features selection techniques. There is little research on the classification of the heart disease dataset. Many of them show good classification accuracy[6].

Tan et al. [7] Proposed a hybrid method in which two machine learning algorithms, Support Vector Machine (SVM) and Genetic Algorithm (G.A), are effectively combined by the wrapper approach. The LIBSVM and the WEKA data mining tool are used to analyze the results of this method. Five data sets (Iris, diabetes disease, breast cancer disease, heart disease and hepatitis) are collected from the Irvine UC machine learning repository for this experiment. After applying the hybrid GA and SVM approach, an accuracy of 84.07% is obtained for heart disease. For all diabetes data, 78.26% accuracy is achieved. The accuracy for breast cancer is 76.20%. The 86.12% accuracy is the result of hepatitis disease.

R. RanjaniRani et al. [8] Proposed a combination of Spider Monkey Optimization Algorithm along with the Support Vector Machine classification algorithm and employed for the microarray cancer gene expression data. It has two phases: first is to eliminate irrelevant and redundant genes and select the subset of genes from the large volume of genes using the Spider Monkey Optimization algorithm. The next phase is to classify cancer types using selected genes from the initial phase. The results conclude that the proposed algorithm outperforms the other existing methods in classification accuracy, and it selects a less number of genes.

Vembandasamy et al. [9] diagnosed heart disease using the Naive Bayes algorithm. Bayes' theorem is used in Naive Bayes. Therefore, Naïve Bayes has a powerful principle of independence. The data used are from one of the leading diabetes research institutes in Chennai. The data set consists of 500 patients. Weka is used as a tool and performs classification using 70% of the Percentage Split. Naive Bayes offers 86.419% accuracy.

3 Methodology.

3.1 Data Set and Attributes

The data is collected from the UCI machine learning repository. The data set is named Heart Disease DataSet and can be found in the UCI machine learning repository. The UCI machine learning repository contains a vast and varied number of datasets which include datasets from various domains. These data are widely used by machine learning community from novices to experts to understand data empirically. Various academic papers and researches have been conducted using this repository. This repository was created in 1987 by David Aha and fellow students at UCI Irvine. Heart disease dataset contains data from four institutions[10].

- Cleveland Clinic Foundation.
- Hungarian Institute of Cardiology, Budapest.
- V.A. Medical Centre, Long Beach, CA.
- University Hospital, Zurich, Switzerland.

For the purpose of this study, the data set provided by the Cleveland Clinic Foundation is used. This dataset was provided by Robert Detrano, M.D. Ph.D. Reason to choose this dataset is, it has less missing values and is also widely used by the research community [11].

Table 1. Attributes of the Heart disease dataset

Attribute	Representation	Information Attribute	Description
Age	Age	Integer	Age in years (29 to 77)
Sex	Sex	Integer	Gender instance (0 = Female, 1 = Male)
ChestPainType	Cp	Integer	Chest pain type (1: typical angina, 2: atypical angina, 3: non- anginal pain, 4: asymptomatic)
RestBloodPressure	Trestbps	Integer	Resting blood pressure in mm Hg[94, 200]
SerumCholestoral	Chol	Integer	Serum cholesterol in mg/dl[126, 564]
FastingBloodSugar	Fbs	Integer	Fasting blood sugar > 120 mg/dl (0 = False, 1= True)
ResElectrocardiographic	Restecg	Integer	Resting ECG results (0: normal, 1: ST-T wave abnormality, 2: LV hypertrophy)
MaxHeartRate	Thalach	Integer	Maximum heart rate achieved[71, 202]
ExerciseInduced	Exang	Integer	Exercise induced angina (0: No, 1: Yes)
Oldpeak	Oldpeak	Real	ST depression induced by exercise relative to rest[0.0, 62.0]
Slope	Slope	Integer	Slope of the peak exercise ST segment (1: up-sloping, 2: flat, 3: down-sloping)
MajorVessels	Ca	Integer	Number of major vessels coloured by fluoroscopy (values 0 - 3)
Thal	Thal	Integer	Defect types: value 3: normal, 6: fixed defect, 7: irreversible defect
Class	Class	Integer	Diagnosis of heart disease (1: Unhealthy, 2: Healthy)

3.2 Classification Task

From the perspective of automatic learning, heart disease detection can be seen as a classification or clustering problem. On the other hand, we formed a model on the vast set of presence and absence file data, we can reduce this problem to classification. For known families, this problem can be reduced to one classification only - having a limited set of classes, certainly including the heart disease sample, it is easier to identify the right class, and the result would be more accurate than with clustering algorithms. In this section, the theoretical context is given on all the methods used in this research. For the purpose of comparative analysis, five Machine Learning algorithms are discussed. The different Machine Learning (ML) algorithms are K-Nearest Neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes and Artificial Neural Network (ANN). The reason to choose these algorithms is based on their popularity [12].

3.3Particle swarm optimization (PSO)

Swarm intelligence is a distributed solution to complex problems which intend to solve complicated problems by interactions between simple agents and their environment[13]–[15]. In 1995, Russel Eberhart, electrical engineer and James Kennedy, socio-psychologist were inspired by the living world to set up a metaheuristic: optimization by particle swarm. This method is based on the collaboration of individuals between them: each particle moves and at each iteration, the one closest to the optimum communicates its position to the others so that they can modify their trajectory. This idea is that a group of unintelligent individuals may have a complex global organization.

Due to its recent nature, a lot of research is being done on P.S.O., but the most effective so far is the extension to the framework of combinatorial optimization.

Figure 1 show the flowchart of the PSO algorithm.

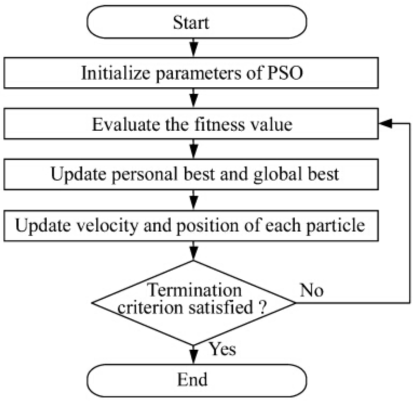
Start
Initialize parameters of PSO
Evaluate the fitness value
Update personal best and global best
Update velocity and position of each particle
Termination criterion satisfied ?
No
Yes
End

Fig. 1. The flowchart of the PSO algorithm

In particle swarm optimization, each individual of the population called particle. In standard PSO, after the initialization of the population, each particle update its velocity and its position in each iteration based on their own experience (pbest) and the best experience of all particles (gbest) as shown in Eql.(1 & 2). At the end of each iteration the performance of all particles will be evaluated by predefined cost functions.

$$v^i[\mathrm{t}+1] = \mathrm{w}.v^i[t] + c_1r_1(p^{i,best}[t] - p^i[t] + c_2r_2(p^{g,best}[t] - p^i[t] \quad (1)$$

$$p^i[\mathrm{t}+1] = p^i[t] + v^i[t+1] \quad (2)$$

Where, $i = 1,2,\dots,N$, N is the a number of swarm population. $v^i[t]$ is the velocity vector in $[t]th$ iteration. $p^i[t]$ represent the current position of the *i*th particle. $p^{i,best}[t]$ is the previous best position of *i*th particle and $p^{g,best}[t]$ is the previous best position of a whole particle. To control the pressure of local and global search, *w* has been used. $c_1$and c_2 are positive acceleration coefficients which respectively called cognitive parameter and social parameter. $r_1$and r_2 are random number between 0 and 1.

3.4Ant Colony Optimization (ACO)

Ant Colony Optimization method explores to find the optimal feature subset using some iterations [16]. The main objective of the Ant Colony Optimization method is to minimize redundancy between them by selecting a subset of feature. In this method, each ant in relation to the previously selected features selects the lowest similarity features. Therefore, if a feature is selected by most ants, it indicates that the features has the lowest similarity with the other features. The feature receive the largest number of pheromones, and the chances of its selection by other ants will be increased in subsequent iterations. Finally, by considering the similarity between the features, the selected main features will have high pheromone values. Thus, the ACO method selects the best features with a minimum of redundancy [17].
Figure 2 shows the illustration of the feature selection problem.

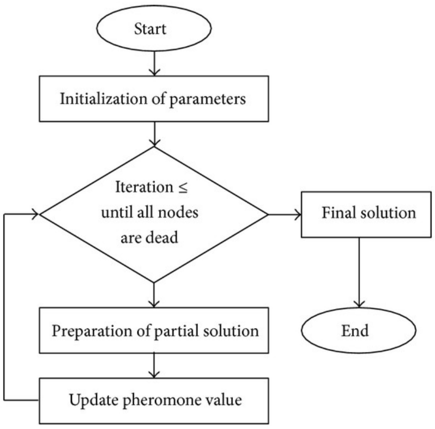

Fig. 2. The flowchart of the ACO algorithm

4 Experiments and Results

The aim of the entire project was to test which algorithm classifies heart disease the best with the proposed optimization methods.
The classification experiment in this paper was carried out under a Weka environment. In addition, due to the small number of selected features, 10-fold cross validation was used. For the purpose of avoiding instable operation results, each experiment was run 10 times, and the optimal classification accuracy was selected for comparison. We evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy according to 3 steps:

1. Optimizations with PSO and ACO approaches for Random Forest, and represented by PA-RF.

2. Comparison of different data mining algorithms on the heart disease dataset.
3. Identification of the best performance-based algorithm for heart disease prediction.

4.1 Effectiveness

In this section, we evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy. The results are shown in Table 2.

Table 2. Classifiers Performance

Evaluation criteria	K-NN	SVM	RF	NB	MLP	PA-RF
Time to build model (s)	0,01	0,07	0,16	0,01	0,89	0,03
Correctly classified instances	202	226	220	226	222	269
Incorrectly classified instance	68	44	50	44	48	1

In order to improve the measurement of classifier performance, the simulation error is also taken into account in this study. To do this, we evaluate the effectiveness of our classifier in terms of: Kappa as a randomly corrected measure of agreement between classifications and actual classes, Mean Absolute Error as the way in which predictions or predictions approximate possible results, Root Mean Squared Error, Relative Absolute Error, Root Relative Absolute Error, Root Relative Squared Error. The results are presented in Figures 3.

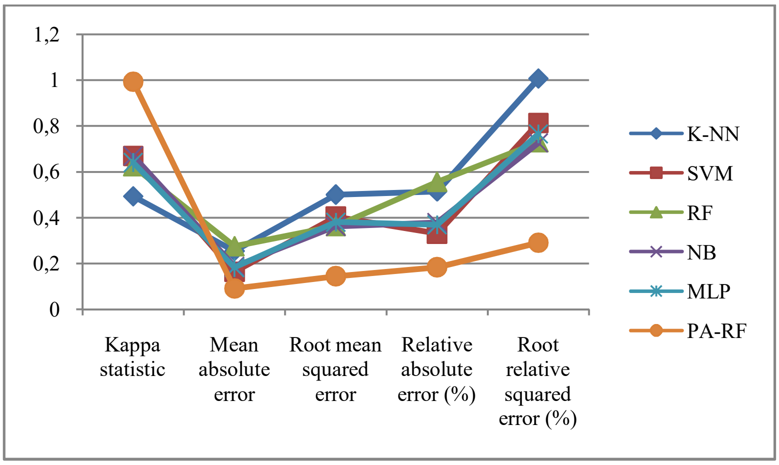

Fig. 3. Simulation Error

4.2 Accuracy Results

Once the predictive model is built, we can check how efficient it is. For that, we compare the accuracy measures based on precision, recall, TP rate and FP rate values for K-NN, SVM, RF, NB, MLP and PA-RF. The results are shown in Figure 4.

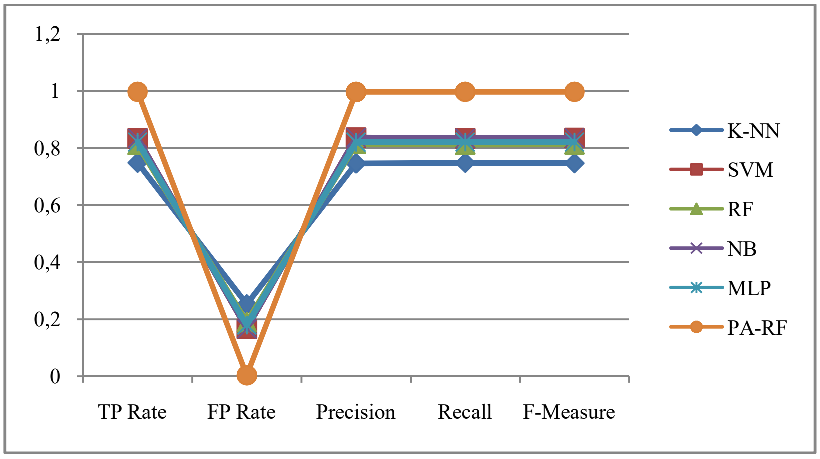

Fig. 4. Accuracy measured

From the different classifiers results presented in Figure 4. We can see that the best results are those generated by Classifiers optimized by PSO and ACO. The PA-RF model shows the best results in comparison with other classifiers algorithms.

Confusion Matrix

Confusion matrices represent a useful way of evaluating classifier; each row of Table 3 represents rates in an actual class while each column shows predictions.

Table 3. Confusion Matrix

	Absence	Presence	
K-NN	113	37	Absence
	31	89	Presence
SVM	130	20	Absence
	24	96	Presence
RF	127	23	Absence
	27	93	Presence
NB	130	20	Absence
	24	96	Presence
MLP	125	25	Absence
	23	97	Presence
PA-RF	149	1	Absence
	0	120	Presence

Results Discussion

In this paper, we applied machine-learning algorithms on heart disease dataset to predict heart disease, based on the data of each attribute for each patient. Our goal was to compare different classification models and define the most efficient one. From all the results above, different algorithms performed better depending upon the situation whether cross-validation, grid search, calibration and feature selection is used or not.
For the comparison of the dataset, performance metrics after feature selection, parameter tuning and calibration areused because this is a standard process of evaluating algorithms. The precision average value of the best performance without optimization it's for SVM and NB with 83,6% than RF with 81,35%. These shows SVM and NB are performing on average, after optimized by FCBF, PSO and ACO, we find the best one is PA-RF with 99,6 %.

Conclusion and Future work

The hybrid Machine Learning model proposed in this work was to optimize the Random Forest algorithm using Particle Swarm Optimization (PSO) associated with Ant Colony Optimization (ACO). The results of the Random Forest optimize (PA-RF) were compared to other algorithms with different performance measures. The results showed that the optimized Random Forest algorithm, PA-RF showed the better performance and effectiveness compared to K-Nearest Neighbour K-NN, Random Forest RF, Naïve Bayes NB, Support Vector Machine SVM and Artificial Neural Network MLP in all the data set used in this study.
This work can be the first step in learning the diagnosis of heart disease through automatic learning and can be extended for future research. There are several limitations to this study, mainly the tools used in this study such as the processing power of the computer and second the time limit available for the study. This type of study requires state-of-the-art resources and expertise in the respective fields.

For future work, we intend to conduct an in-depth study of these datasets by combining Machine Learning techniques with deep learning models on the application of more complex deep learning architectures to achieve better performance. In addition, we test our in-depth learning approach on larger data sets with more disease classes to achieve higher accuracy.

References

1. Kamkar, M. Akbarzadeh-T, and M. Yaghoobi, "Intelligent water drops a new optimization algorithm for solving the Vehicle Routing Problem," in 2010 IEEE International

Conference on Systems, Man and Cybernetics, 2010, pp. 4142–4146.

2. Z. Gazzaz, N. M., Yusoff, M. K., Ramli, M. F., Juahir, H., & Aris, “Artificial neural network modeling of the water quality index using land use areas as predictors,” Water Environ. Res., vol. 87, no. 2, pp. 99–112, 2015.
3. G. Zhang, Y., Wang, S., & Ji, “A comprehensive survey on particle swarm optimization algorithm and its applications,” Math. Probl. Eng., 2015.
4. H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, “Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification,” Comput. Biol. Chem., vol. 56, pp. 49–60, 2015.
5. H. Yu, L., & Liu, “Feature selection for high-dimensional data: A fast correlation-based filter solution,” Proc. 20th Int. Conf. Mach. Learn., pp. 856–863, 2003.
6. M. Fatima, M., & Pasha, “Survey of machine learning algorithms for disease diagnostic,” J. Intell. Learn. Syst. Appl., vol. 9, no. 1, p. 1, 2017.
7. K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, “A hybrid evolutionary algorithm for attribute selection in data mining,” Expert Syst. Appl., vol. 36, no. 4, pp. 8616–8630, 2009.
8. Rani, R. Ranjani, and D. Ramyachitra. "Microarray Cancer Gene Feature Selection Using Spider Monkey Optimization Algorithm and Cancer Classification using SVM." Procedia computer science 143 (2018): 108-116.
9. E. Vembandasamy, K., Sasipriya, R., & Deepa, “Heart Diseases Detection Using Naive Bayes Algorithm,” IJISET-International J. Innov. Sci. Eng. Technol., vol. 2, pp. 441–444, 2015.
10. M. Lichman, “UCI Machine Learning Repositry [Online],” Available: https://archive.ics.uci.edu/, 2013.
11. U. H. Dataset, “UCI Machine Learning Repository,” [online]. URLhttps//archive.ics.uci.edu/ml/machine-learning-databases/heartdisease/ Hear.
12. L. Van Cauwenberge., “Top 10 Machine Learning Algorithms,” Data Sci. Cent., 2015.
13. R. Poli, J. Kennedy, and T. Blackwell, “Particle swarm optimization,” Swarm Intell., vol. 1, no. 1, pp. 33–57, 2007.
14. J. Kennedy, “Particle Swarm Optimization,” in Encyclopedia of Machine Learning (Springer), C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 760–766.
15. Y. Shi, “Particle swarm optimization,” IEEE Connect., vol. 2, no. 1, pp. 8–13, 2004.
16. M. Dorigo and M. Birattari, “Ant Colony Optimization,” in In Encyclopedia of machine learning (Springer), C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 36–39.
17. S. Tabakhi, P. Moradi, and F. Akhlaghian, “An unsupervised feature selection algorithm based on ant colony optimization,” Eng. Appl. Artif. Intell., vol. 32, pp. 112–123, 2014.