Cluster Computing
https://doi.org/10.1007/s10586-018-2416-4

An optimized feature selection based on genetic approach and support vector machine for heart disease

Chandra Babu Gokulnath[1] · S. P. Shantharajah[1]

Received: 13 January 2018 / Revised: 7 February 2018 / Accepted: 6 March 2018

Abstract
Heart disease diagnosis is found to be a challenging issue which can offer a computerized estimate about the level of heart disease so that supplementary action can be made easy. Thus, heart disease diagnosis has expected massive attention worldwide among the healthcare environment. Optimization algorithms played a significant role in heart disease diagnosis with good efficiency. The objective of this paper is to propose an optimization function on the basis of support vector machine (SVM). This objective function is used in the genetic algorithm (GA) for selecting the more significant features to get heart disease. The experimental results of the GA–SVM are compared with the various existing feature selection algorithms such as Relief, CFS, Filtered subset, Info gain, Consistency subset, Chi squared, One attribute based, Filtered attribute, Gain ratio, and GA. The receiver operating characteristic analysis is performed to evaluate the good performance of SVM classifier. The proposed framework is demonstrated in the MATLAB environment with a dataset collected from Cleveland heart disease database.

Keywords Heart disease diagnosis · Support vector machine · Genetic algorithm · Roulette wheel selection · Receiver operating characteristic · Crossover · Mutation · Elitism

1 Introduction

Nowadays, heart diseases have found to be a major issue in human health. Recently, a report state that the root causes of heart attacks [1]. Cardiovascular disease plays an important role in the sudden death of people in industrialized countries [2, 3]. Cardiovascular disease not only affects the human health but also economics and cost of the countries. The major risk factors for the cardiovascular disease include obesity, diabetes, family history, smoking and high cholesterol [4]. As newborn babies have the chance of heart disease, checking of cardiovascular diseases for a newborn baby is more common. Fatigue and chest pain are the most common symptoms for getting the heart disease. In general, treatment for these heart patients is given on the basis of lab test results, patient history and patient's answers to questions [5].

Nowadays, a number of data mining algorithms and machine learning algorithms are developed for predicting the early stage of heart disease [6–10]. In many cases, data mining algorithms and machine learning algorithms are used over a huge size of heart disease data [11, 12]. This would create high computation complexity and less efficiency. In order to overcome this issue, a range of feature selection algorithms is developed for identifying more significant features to predict the heart disease [13–15]. This paper provides an introduction to various types of feature selection algorithms. More commonly, feature selection algorithms are classified as following types it includes correlation-based feature selection and consistency filter-based feature selection [16–18].

Correlation-based feature selection is one of the types of simple filter algorithm. In correlation-based feature selection, a heuristic evaluation function is used for ranking the features subsets [19]. The more significant features are identified on the basis of high correlation while low correlated features are ignored for the training and testing process of the prediction model [20]. Moreover, redundant features are also screened out from the prediction model. In

✉ Chandra Babu Gokulnath
Gokulnath.c@vit.ac.in

[1] School of Information Technology and Engineering, VIT University, Vellore, India

Published online: 16 March 2018

consistency filter-based feature selection, the more important features are selected on the basis of consistency values of each feature [21–23]. In this algorithm, a random subset S is identified for each iteration from the number of features. If the current best features are greater than a number of features of S, S converts the new current best.

Once the significant features are identified from the dataset, then these features are ranked on the basis of various feature ranking techniques such as ReliefF and Information Gain. ReliefF is extended on the basis of the original Relief algorithm. The vital role of the Relief algorithm is to rank the multiclass features with partial and noisy data. The ReliefF algorithms are used in various applications where interaction among features is low and capturing of features with local dependencies. Information Gain is one of the types in feature evaluation methods widely used in diverse applications. This method orders all the features on the basis of a user-defined threshold value. The concept of entropy is embedded with the Information Gain method to rank the features. The joint evidence of the objective variable is used for calculating the expected value for the Information Gain method [24–26].

Nowadays, a number of machine learning algorithms and artificial intelligence techniques were developed for various disease diagnosis. For example, artificial intelligence techniques used in for the prediction of heart disease, machine learning method is presented in [27] for identifying the early stage of heart disease. Case-based reasoning (CBR) method is proposed in [28] for diagnosis, prognosis, and recommendation in healthcare. Similarly, genetic algorithm (GA) with Binary Particle Swarm Optimization (BPSO) is used in [19] for the prediction of Coronary Artery Disease. Moreover, a range of classification and regression methods are used in [29] for identifying the hidden values and useful information from the healthcare dataset.

2 Related work

Ordonez [30] has presented a novel algorithm for reducing the number of rules in the association mining process. This method removes the irrelevant or less significant association rules from the entire association rule test result. Once the more significant rules are identified from the dataset, the selected rules are evaluated on the basis of various validation metrics such as accuracy, precision, sensitivity, specificity, true positive rate and false positive rates. Independent test cases are used to evaluate the proposed algorithm. These rules are applied to the real-time medical records to identify the early stage of heart attacks.

Tan et al. [31] have presented a novel feature selection approach by using the genetic algorithms (GAs) with wrapper approach. The vital role of the GA is to identify the best attributes from the raw data set. The fitness function used in this GA originally identifies the best features from the entire dataset. Nahar et al., [4] have presented a novel intelligent method based on the three different meta-heuristic methods such as Apriori, Predictive Apriori and Tertius. The vital role of this algorithm is to extract the most significant rules that classify the heart disease. The experimental results of this paper prove the good performance of the proposed algorithm for feature selection of heart disease.

Polat and Gunes [32] have presented a Kernel F-score Feature Selection (KFFS) method not only for identifying the most significant features but also removing the redundant and irrelevant (less significant) features from the medical records of the heart disease patients. The working principle of the KFFS method is classified into two major categories it includes Radial Basis Function (RBF) kernel functions and F-score method. The objective of the RNF kernel function is to converts the features of medical datasets into a kernel space. Moreover, the F-score method is originally used to calculate the F-score values with high dimensional feature space.

Luukka and Lampinen [33] have used the Principal Component Analysis (PCA) with differential evolution classifier for generating the classification rules for the heart disease. The objective of the PCA method is to preprocess the raw medical data whereas differential evolution classifier is originally used for developing the classification rules. The demonstration of this algorithm is done with the help of classical Electronic Medical Record (EMR) which contains a range of patient demographical dataset it includes patient symptoms, patient lab results, patient details about angina and coronary infarction, patient clinical observations, Electrocardiography (ECG), heartbeat rate, blood pressure, glucose level and insulin level and so on.

Yan et al. [5] have presented a GA based feature selection method for identifying the more significant features to get heart disease. The proposed algorithm effectively demonstrated on five different heart diseases it includes congenital heart disease, coronary heart disease, hypertension, chronic pulmonale and rheumatic valvular heart disease. Ozcift and Gulten [34] have presented a novel classification algorithm on the basis of Rotation Forest (RF). The objective of this algorithm is to improve the accuracy of the existing machine learning algorithms. This algorithm is successfully implemented in a medical decision support system for predicting the heart diseases as well as Parkinson's and diabetes.

Chi et al. [35] have presented an Optimal Decision Path Finder (ODPF) for improving the accuracy of disease diagnosis as well as reducing the money and time spent on

diagnostic testing. Park et al. [28] have presented a new expert system to overcome the unequal misclassification issues. The number of neighbor and best classification of boundary point are used to improve the accuracy of the disease diagnosis. In order to adjust the cut-off distance point and optimal cut-off classification point, a novel classifier is introduced on the basis of best neighbors.

Nahar et al. [29] have presented a reviewed the various feature selection algorithms and intelligent methods for the prediction of heart diseases. The authors of this paper are also used Cleveland heart disease dataset to perform the feature selection and classification task. The experimental results generated from this paper are compared with other existing machine learning algorithms such as SVM, random forests, logistic regression and KNN and son on. Similarly, Laercio et al. [36] have used the Fuzzy Binary Space Partitioning method for partitioning of the input space. This would give the reduced test cases and optimal features for the prediction of heart diseases. Kemal [37] has developed a hybrid method called Artificial Immune Recognition System (AIRS) for preprocessing, feature selection and classification. The workflow of the proposed framework is classified as two stages namely dimensionality reduction for reducing the features from 19 to 9 and preprocessing based on fuzzy weighted pre-processing (0,1) for normalizing the feature values from 0 to 1.

3 Proposed framework for feature selection

The goal of this paper is to propose an optimization function on the basis of support vector machine (SVM). This objective function is used in the GA for finding the important features to get heart disease. The overall workflow of this paper is presented in Fig. 1.

3.1 Genetic algorithm based feature selection

A wrapper based GA is used in this paper to select the more significant features. The vital role of the wrappers used in the GA is to find the space and evaluate each subset by executing a model on the subset. A search algorithm used in the GA is used to find a population of candidate solution (individual) and an optimization problem is established for an improved solution. Figure 2 represents the working principle of a GA. More commonly, a candidate solution is

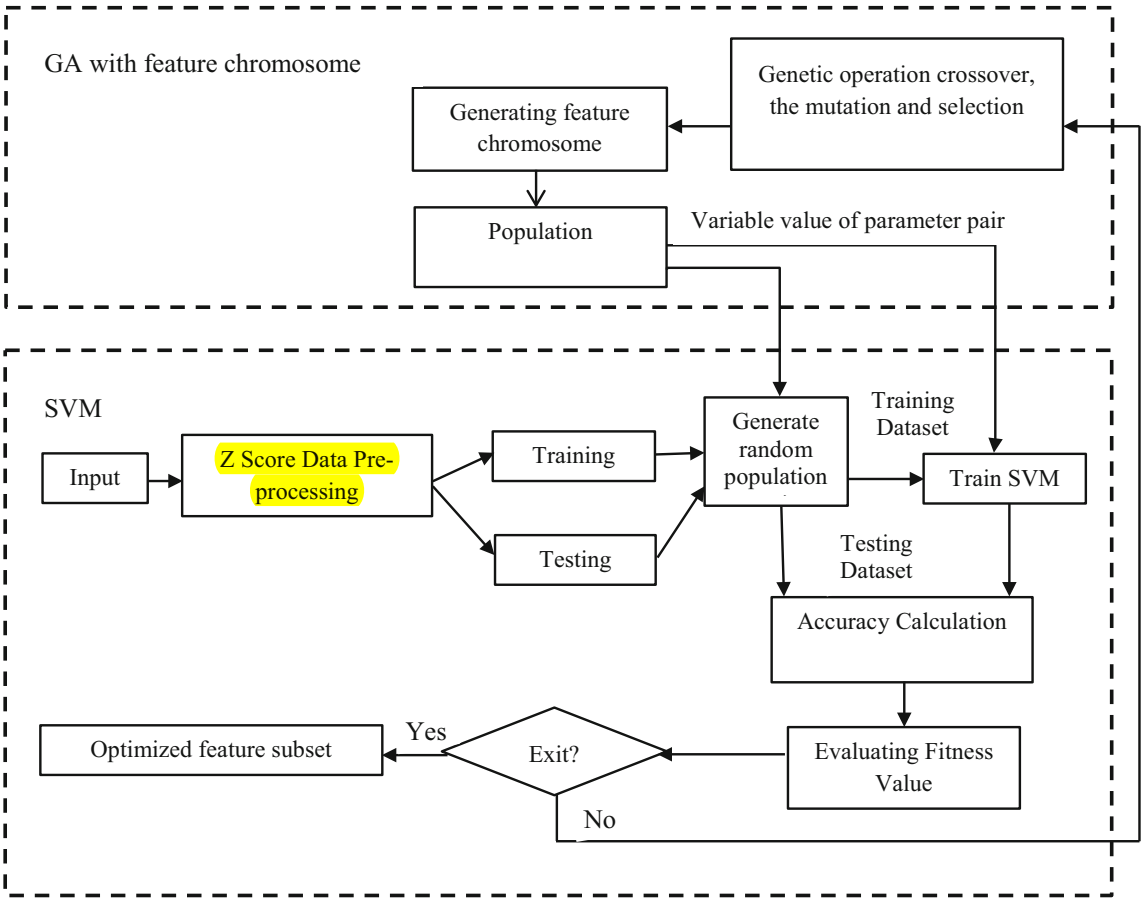


Fig. 1 Workflow of the proposed framework

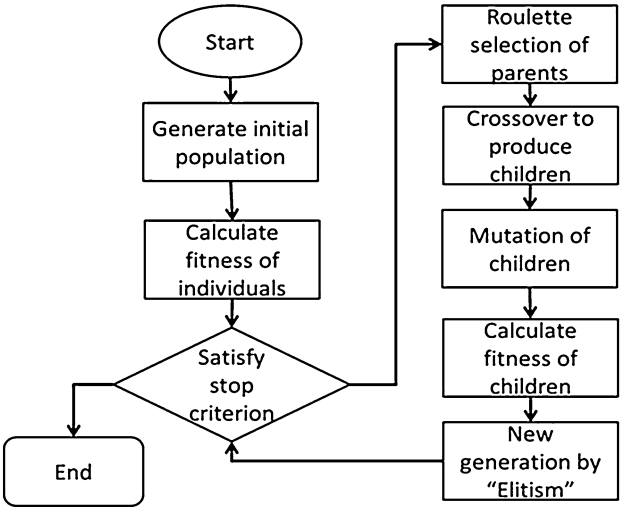


Fig. 2 Working principle of genetic algorithm

referred to an element of potential solutions to a given problem. The total population comprises a number of candidate solutions being developed by GA. Moreover, a candidate solution consists of a set of parameters which are also called as genotypes or chromosomes. These set of parameters are represented as genomes. Moreover, genomes are represented as binary strings whereas the candidate solutions are represented as a numerical form. In addition, the evolution procedure of GAs is started from an initial population of randomly created genomes. The fitness function is used in all the iterations for calculating the fitness value. The overall best fitness value is computed on the basis of comparison between the current populations. In this paper, SVM is used for classifying the features; these results are employed in the proposed framework for calculating the fitness function. Genotypes are used to represents the feature vector of the original dataset. The phenotypes of the chromosomes are used representing the masks of the feature vector. Hence, the phenotypes labeled as '0' represent the removed features while phenotypes labeled as '1' represent the selected features. Thus, phenotypes labeled as '0' known as less significant features while phenotypes labeled as '1' known as highly significant features. On the basis of phenotypes, a group of subsets is created by each genotype. These subsets are used as training sets for the proposed framework. Figure 3 represents the working principle of genetic operations.

3.1.1 Roulette wheel selection

In roulette wheel selection, the expected value of an individual's fitness is divided by the actual fitness of the population. The size of the roulette wheel is fixed on the basis of size being proportional to the individual's fitness. Hence, each individual is allocated a part of the roulette wheel. The roulette wheel is rotated on the basis of the number of individuals in the population.

On each rotation, an individual is selected as a pool of parents for the next generation. Figure 4 represents the conceptual view of roulette wheel selection.

Roulette Wheel Selection Algorithm

Step 1: Addition of total expected value 't' of the individuals in the population.

Step 2: Up to N times:

i. Select a random number 'r' between o and 't'.

ii. For each individual in the population

Calculate the sum of the expected values, until the sum is greater than or equal to 'r'.

The individuals (features) are selected whose expected value greater than the sum of the expected values

3.1.2 Crossover (recombination)

In Crossover, two-parent solutions are taken to produce a child. Afterwards, the selection (reproduction) process is done followed by the development of better individuals. Moreover, the reproduction process is used for developing replicas of significant strings but does not produce new ones.

Crossover Algorithm
Step 1: A pair of two individual strings is selected for the mating with the help of reproduction operator
Step 2: A cross-site is selected on the basis of random manner beside the string length.
Step 3: Swapping of position values between the two strings

In this paper, a single point crossover is used to perform mating of two chromosomes. Here, two chromosomes are cut once and the cuts are exchanged between two chromosomes. Figure 5 shows the working principle of single point crossover.

3.1.3 Mutation

Once the crossover process is finished, the strings are subjected to perform mutation process. Mutation of a bit encompasses flipping a bit from 0 to 1 and 1 to 0. The vital role of mutation process is to recover the randomly troubling genetic information and missing genetic materials. The working principle of the mutation process is more similar to a traditional search operator. The mutation process is used for the examination of the entire search space when crossover abuses the current solution to find better ones. In addition, the objective of the mutation process is to preserve genetic diversity in the population. Hence, the mutation process produces new genetic structures in the population with the help of changing its original structure. Figure 6 shows the working principle of mutation.

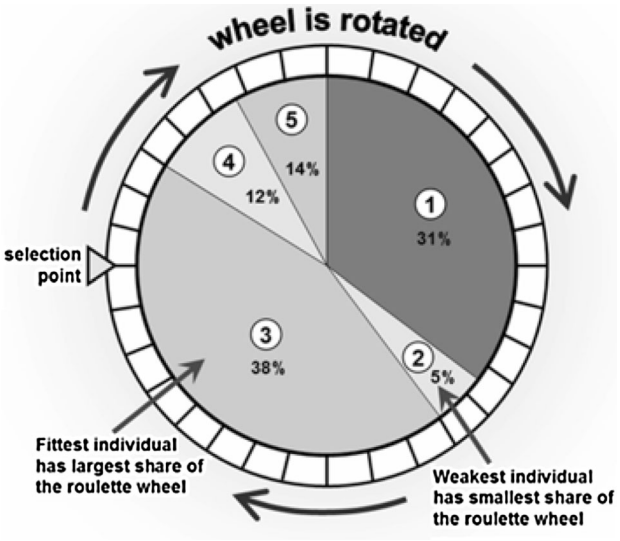
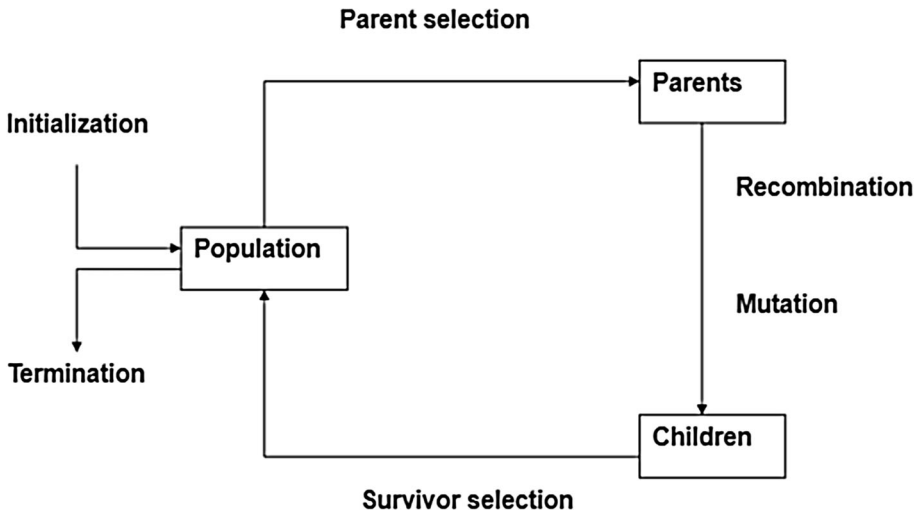


Fig. 4 Roulette wheel selection

3.1.4 Elitism

Finally, elitism is performed for improving the performance of the GA. In this procedure, the few best chromosomes are copied to the new population. The other chromosomes can be lost if they are not nominated to reproduce or if mutation or crossover terminates them.

Fig. 3 Genetic operations



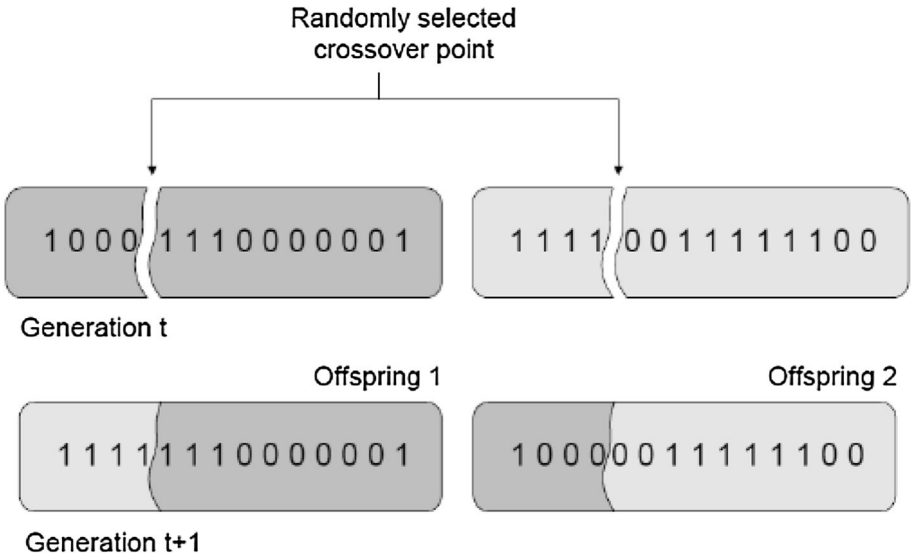


Fig. 5 Single point crossover

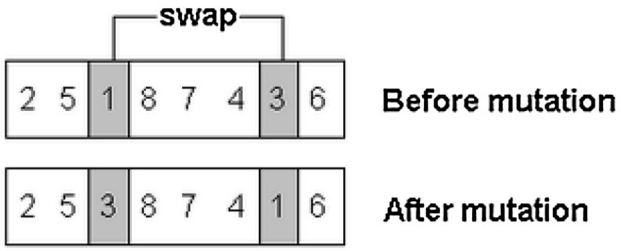


Fig. 6 Mutation process

3.2 Support vector machine (SVM) classifier

Vapnik [38] has proposed SVM for density estimation, classification and regression problems. The objective of SVM is to separates the data points xi based on the hyperplane $w.x + b = 0, x_i \in \mathbb{R}^n$, that corresponding to a given decision rule: $g(x) = sign(w.x + b)$.

The SVM identifies this hyperplane $w.x + b = 0$ (i.e.) utmost away from the data points xi. The basic principle behind the SVM is that a hyperplane faraway from any experimental data points should decrease the risk of creating incorrect results when classifying new data.

Let a pattern contains of a pair$\{x_i, y_i\}_{i=1}^{N}$

where $x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$.

Let us assume that the space of patterns is $X \subset \mathbb{R}^n$ and the space of labels is $Y \subset \{-1, 1\}$.

The objective of SVM is to find a classifier $y(x)$

$$y(x) = sign\left[\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right], \tag{1}$$

where $\alpha_i =$ represents the positive real constants, b = represents a real constant.

More commonly, the kernel function $K(x_i, x) = \phi(x_i), \phi(x),$

where $\langle \cdot, \cdot \rangle$ = inner product operation, $\phi(x)$ = representing a nonlinear high dimensional mapping.

Let us assume that, the given data set is divided by a linear hyperplane. The separation of this process on high dimensional space is defined by,

$$y_i\left[w^T\phi(x_i) + b\right] \geq 1, \quad i = 1, \ldots, N. \tag{2}$$

A slack variable n is introduced when splitting hyperplane does not exist,

$$\begin{aligned} &y_i\left[w^T\phi(x_i) + b\right] \geq 1 - \xi_i, \quad i = 1, \ldots, N \\ &\xi_i \geq 0, \quad i = 1, \ldots, N \end{aligned} \tag{3}$$

On the basis of the structural risk minimization theory, the minimization is defined by,

$$\min_{w,\xi} J_1(w, \xi) = \frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i \tag{4}$$

Using the Eq. (3), lagrangian function is developed as follows,

$$\begin{aligned} L_1(w, b, \xi, \alpha, \beta) = J_1(w, \xi) \\ - \sum_{i=1}^{N} \alpha_i \left\{ y_i \left[w^T \phi(x_i) + b \right] - 1 + \xi_i \right\} \\ - \sum_{i=1}^{N} \beta_i \xi_i \end{aligned} \tag{5}$$

where $\alpha_i > 0, \beta_i > 0, (i = 1, \ldots, N)$ = represents the Lagrangian multipliers of Eq. (3).

```
@relation heart-statlog
@attribute age real
@attribute sex real
@attribute chest real
@attribute resting_blood_pressure real
@attribute serum_cholestoral real
@attribute fasting_blood_sugar real
@attribute resting_electrocardiographic_results real
@attribute maximum_heart_rate_achieved real
@attribute exercise_induced_angina real
@attribute oldpeak real
@attribute slope real
@attribute number_of_major_vessels real
@attribute thal real
@attribute class { absent, present}
@data
70,1,4,130,322,0,2,109,0,2.4,2,3,3,present
67,0,3,115,564,0,2,160,0,1.6,2,0,7,absent
57,1,2,124,261,0,0,141,0,0.3,1,0,7,present
64,1,4,128,263,0,0,105,1,0.2,2,1,7,absent
74,0,2,120,269,0,2,121,1,0.2,1,1,3,absent
65,1,4,120,177,0,0,140,0,0.4,1,0,7,absent
56,1,3,130,256,1,2,142,1,0.6,2,1,6,present
59,1,4,110,239,0,2,142,1,1.2,2,1,7,present
```

Fig. 7 Heart disease dataset

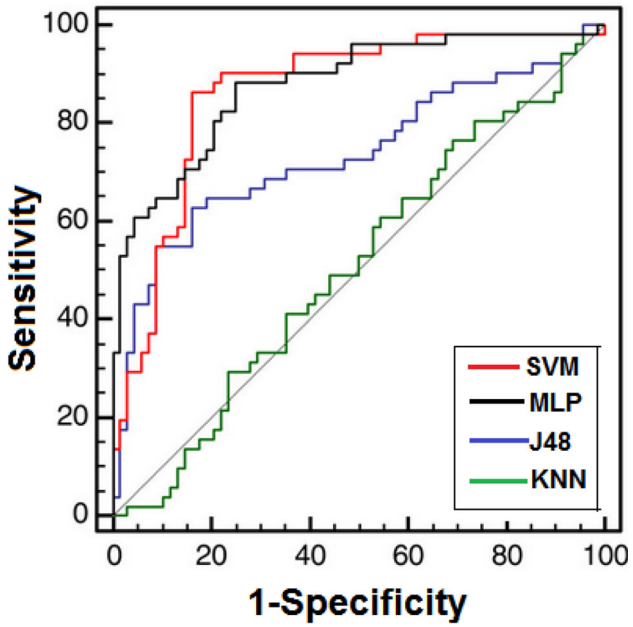


Fig. 8 ROC curve for different classifiers (original features)

The optimal point is computed on the basis of Lagrangian function, i.e.

$$\max_{w,\beta}\ \min_{w,b,\xi} L_1(w, b, \xi, \alpha, \beta) \tag{6}$$

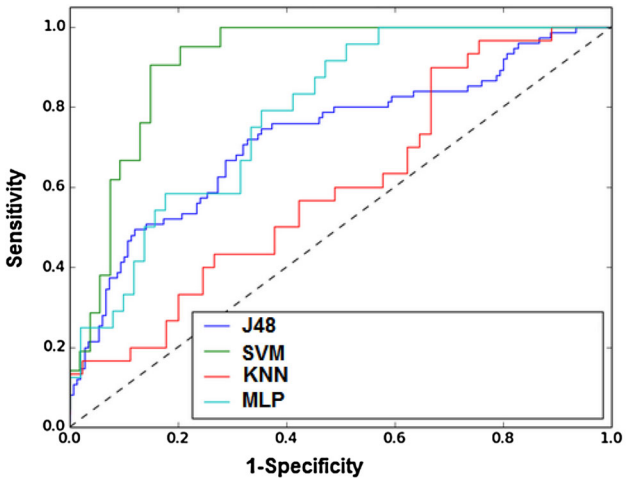


Fig. 9 ROC curve for different classifiers (reduced features)

The partial differentiation with zeros are applied in Eq. (6),

$$\begin{aligned}
&\frac{\partial L_1}{\partial w} = 0, w = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) \\
&\frac{\partial L_1}{\partial b} = 0, w = \sum_{i=1}^{N} \alpha_i y_i = 0 \\
&\frac{\partial L_1}{\partial \xi_i} = 0, 0 \le \alpha_i \le c, i = 1, 2, \ldots, N
\end{aligned} \tag{7}$$

The quadratic programming (QP) problem will obtain when applying Eq. (7) in (5),

$$\min_{\alpha} Q_1(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{8}$$

where $K(x_i, x_j) = \phi(x_i), \phi(x_j)$ = represents the kernel function.

A hyperplane in the high dimensional space and the classifier in the original space as in (1) are found by explaining the above QP problem Eq. (8) subject to the given constraints in Eq. (7).

Proposed genetic algorithm with SVM classifier for feature selection

Step 1: Supply input dataset includes training dataset and testing dataset.

Step 2: Data preprocessing based on z score normalization

$z = (x - \mu) / \sigma$

Where,

μ = mean

σ = standard deviation

Step 3: Generate random population of *n* chromosomes

Step 3.1: Evaluate the fitness function *f(x)* of each chromosome *x* in the population

Step 3.2: Train the SVM classifier by training set

Step 3.3: Calculate the Fitness function

Classification accuracy of SVM and the number of selected features are used to construct a fitness function.

$$\text{classification accuracy of SVM } f_1(I) = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

Where,

TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative

$$\text{Number of selected genes } f_2(I) = \left(1 - \frac{m_\tau}{p}\right) \quad (10)$$

Where,

$m_\tau =$ denotes the number of bits having the value "1"

p = represents the length of the chromosome

The fitness function f is defined by:

$f(I) = \alpha f_1(I) + (1 - \alpha) f_2(I)$ subject to $0 < \alpha < 1$

Let α greater than 0.5 for high classification accuracy and lesser than 0.5 for small sized gene subsets

Step 4: Perform the genetic operations such as crossover, the mutation and the selection

Step 5: Stop the algorithm if termination criterion is satisfied; (Up to 10N iteration, N the number of features), return to Step 2 otherwise.

Table 1 Confusion matrix

Test result	Original result		Row total
	P	N	
P	TP	FP	TP + FP (total number of positive cases)
N	FN	TN	FN + TN (total number of negative cases)
Column total	TP + FN	FP + TN	TP + TN + FP + FN (total number of subjects in study)

Table 2 Comparison of classifiers (original features)

S. no	Classification algorithm	No. of correctly classified instances	No. of wrongly classified instances	Accuracy (%)
1	SVM	226	44	83.70
2	Multilayer perception	211	59	78.148
3	J48	207	63	76.66
4	KNN	203	67	75.18

Table 3 Comparison of feature selection algorithms

Feature selection algorithms	Selected features
Relief	13 (1,2,3,4,5,6,7,8,9,10,11,12,13)
Info gain	13 (1,2,3,4,5,6,7,8,9,10,11,12,13)
Chi squared	13 (1,2,3,4,5,6,7,8,9,10,11,12,13)
Filtered subset	6 (3,8,9,10,12,13)
One attribute based	13 (1,2,3,4,5,6,7,8,9,10,11,12,13)
Consistency based	10 (1,2,3,7,8,9,10,11,12,13)
Gain ratio	13 (1,2,3,4,5,6,7,8,9,10,11,12,13)
Filtered attribute	13 (1,2,3,4,5,6,7,8,9,10,11,12,13)
CFS	8 (3,7,8,9,10,11,12,13)
Genetic algorithm	6 (3,7,8,9,10,13)
Genetic algorithm with SVM	7 (3,7,8,9,10,12,13)

4 Performance evaluation

In this section, the experimental results of the GA–SVM are compared with the various existing feature selection algorithms such as Relief, CFS, Filtered subset, Info gain, Consistency subset, Chi squared, One attribute based, Filtered attribute, Gain ratio, and GA. The receiver operating characteristic (ROC) analysis is used for evaluating the good performance of SVM classifier. The proposed GA–SVM framework is demonstrated in the MATLAB environment with a dataset collected from Cleveland heart disease database [39] (Fig. 7). Figure 8 represents the ROC curve for different classifiers (original features) while Fig. 9 represents the ROC curve for different classifiers (reduced features). The ROC analysis proved that the good performance of the proposed GA–SVM framework for feature selection. Table 1 depicts the confusion matrix and

Table 4 Comparison of the accuracy of classifiers (selected attributes) (in %)

Feature selection algorithms	Multilayer perceptron (accuracy in %)	KNN (accuracy in %)	J48 (accuracy in %)	SVM (accuracy in %)	Average
Relief	78.14	75.18	76.66	83.70	78.42
CFS	82.22	78.14	81.11	85.5	81.74
Filtered subset	78.88	80	79.60	85.18	80.91
Info gain	80.37	75.18	76.66	83.70	78.97
Consistency subset	81.11	78.14	78.88	84.07	80.55
Chi squared	80.37	75.18	76.66	83.70	78.97
One attribute based	79.25	75.18	76.66	83.70	78.69
Filtered attribute	80.37	75.18	76.66	83.70	78.97
Gain ratio	78.88	75.18	76.66	83.70	78.60
Genetic algorithm	79.81	81.4	80.63	86.22	82.01
Genetic algorithm with SVM	81.32	84.3	83.64	88.34	84.40

Table 2 depicts the comparison of classifiers (original features) while Table 4 depicts the comparison of the Accuracy of classifiers (selected attributes). As shown in Table 3, the selected features are varied on the basis of different feature selection algorithms (Table 4).

Sensitivity and specificity are defined by,

$$\text{Specificity} = \frac{\text{True Negative (TN)}}{\text{False Positive (FP)} + \text{True Negative (TN)}} \quad (11)$$

$$\text{Sensitivity} = \frac{\text{True Positive(TP)}}{\text{True Positive(TP)} + \text{False Negative (FN)}} \quad (12)$$

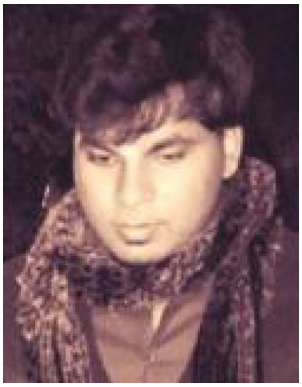
5 Conclusion

An optimization function on the basis of SVM is proposed in this paper for improving the performance of GA. GA with SVM based feature selection algorithms identifies 7 features (3,7,8,9,10,12,13) to get heart disease. The resultant features are supplied to SVM for finding the accuracy. The SVM produces 83.70% of accuracy when classifying the heart disease with the whole features. However, the SVM classifier produces 88.34% of accuracy when classifying the heart disease with the selected features. It is comparatively high when compared with Relief, CFS, Filtered subset, Info gain, Consistency subset, Chi squared, One attribute based, Filtered attribute, Gain ratio, and GA. In addition, the ROC analysis proved the good performance of SVM classifier.

References

1. Tsipouras, M.G., et al.: Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. J. IEEE Trans. Inform. Technol. Biomed. **12**(4), 447–458 (2008)
2. Kusiak, A., Caldarone, C.A., et al.: Hypo plastic left heart syndrome knowledge discovery with a data mining approach. J. Comput. Biol. Med. **36**(1), 21–40 (2006)
3. Huang, M.-J., et al.: Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. J. Expert Syst. Appl. **32**, 856–867 (2007)
4. Nahar, J., et al.: Association rule mining to detect factors which contribute to heart disease in males and females. J. Expert Syst. Appl. **40**, 1086–1093 (2013)
5. Padma, T., Mir, S.A., Shantharajah, S.P.: Intelligent decision support system for an integrated pest management in apple orchard. In: Intelligent Decision Support Systems for Sustainable Computing, Springer, pp. 225–245 (2017)
6. Das, Resul, Turkoglu, Ibrahim, et al.: Effective diagnosis of heart disease through neural networks ensembles. J. Expert Syst. Appl. **36**, 7675–7680 (2009)
7. Das, Resul, Turkoglu, Ibrahim, et al.: Diagnosis of valvular heart disease through neural networks ensembles. J. Comput. Methods Progr. Biomed. **93**, 185–191 (2009)
8. Gokulnath, C., Priyan, M. K., Balan, E. V., Prabha, K. R., Jeyanthi, R.: Preservation of privacy in data mining by using PCA based perturbation technique. In: Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015 International Conference on (pp. 202–206). IEEE (2015)
9. Babaoglu, et al.: Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization. J. Expert Syst. Appl. **36**, 2562–2566 (2009)
10. Rajeswari, K., et al.: Feature selection in ischemic heart disease identification using feed forward neural networks. Int. Symp. Robot. Intell. Sens. **41**, 1818–1823 (2012)
11. Park, Y.-J., et al.: Cost-sensitive case-based reasoning using a genetic algorithm: application to medical diagnosis. J. Artif. Intell. Med. **51**, 133–145 (2011)
12. Nahar, J., et al.: Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. J. Expert Syst Appl. **40**, 96–104 (2013)
13. Polat, K., Güneş, S.: A hybrid approach to medical decision support systems: combining feature selection, fuzzy weighted pre-processing and AIRS. J. Comput. Methods Progr. Biomed. **88**, 164–174 (2007)
14. Polat, K., et al.: Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbor) based weighting preprocessing. J. Expert Syst. Appl. **32**, 625–631 (2007)
15. Uma, S., Shantharajah, S.P., Rani, C.: Passive Incidental alertness-based link visualization for secure data transmission in manet. J. Appl. Secur. Res. **12**(2), 304–322 (2017)
16. Kahramanli, H., Allahverdi, N.: Design of a hybrid system for the diabetes and heart diseases. J. Expert Syst. Appl. **35**, 82–89 (2008)
17. Khatibi, V., Montazer, G.A.: A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. J. Expert Syst. Appl. **37**, 8536–8542 (2010)
18. Priyan, M.K., Devi, G.U.: Energy efficient node selection algorithm based on node performance index and random waypoint mobility model in internet of vehicles. Clust. Comput. **9**, 1–15 (2017)
19. Varatharajan, R., Manogaran, G., Priyan, M.K., Sundarasekar, R.: Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm. Clust. Comput. **35**, 1–10 (2017)
20. Anooj, P.K.: Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. J. Comput. Inform. Sci. **24**, 27–40 (2012)
21. Nawi, N.M. et al.: The development of improved back-propagation neural networks algorithm for predicting patients with heart disease. In: Proceedings of the First International Conference ICICA, vol. 6377, pp. 317–324 (2010)
22. Varatharajan, R., Manogaran, G., Priyan, M.K.: A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing. Multimed. Tools Appl. (2017). https://doi.org/10.1007/s11042-017-5318-1
23. Paredesa, S., et al.: Long term cardiovascular risk models' combination. J. Comput. Methods Progr. Biomed. **101**, 231–242 (2011)
24. Shilaskar, S., et al.: Feature selection for medical diagnosis: evaluation for cardiovascular diseases. J. Expert Syst. Appl **40**, 4146–4153 (2013)

25. Pu, L.N., et al.: Investigation on cardiovascular risk prediction using genetic information. J. IEEE Trans. Inform. Technol. Biomed. **16**(5), 795–808 (2012)
26. Manogaran, G., Vijayakumar, V., Varatharajan, R., Kumar, P.M., Sundarasekar, R., Hsu, C.H.: Machine learning based big data processing framework for cancer diagnosis using hidden markov model and gm clustering. Wirel. Person. Commun. **22**, 1–18 (2017)
27. UCI Machine Learning Repository: Heart Disease Data Set.: Archive.ics.uci.edu. http://archive.ics.uci.edu/ml/datasets/Heart+Disease (2017). Accessed 22 Oct 2017
28. Balan, E.V., Priyan, M.K., Gokulnath, C., Devi, G.U.: Fuzzy based intrusion detection systems in MANET. Proc. Comput. Sci. **50**, 109–114 (2015)
29. Babaoglu, I., et al.: A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine. J. Expert Syst. Appl. **37**(4), 3177–3183 (2010)
30. Ordonez, C.: Association rule discover with the train and test approach for the heart disease prediction. IEEE Trans. Inform. Technol. Biomed. **10**(2), 334–343 (2006)
31. Tan, K.C., et al.: A hybrid evolutionary algorithm for attribute selection in data mining. J. Expert Syst. Appl. **36**, 8616–8630 (2009)
32. Polat, K., Gunes, S.: A new feature selection method on classification of medical datasets: Kernel F-score feature selection. J. Expert Syst. Appl. **36**, 10367–10373 (2009)
33. Luukka, P., Lampinen, J.: A classification method based on principal component analysis and differential evolution algorithm applied for prediction diagnosis from clinical EMR heart data sets. J. Comput. Intell. Optim. Adapt. Learn. Optim. **7**, 263–283 (2010)
34. Yan, H., et al.: Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm. J. Appl. Soft Comput. **8**, 1105–1111 (2008)
35. Ozcift, A., Gulten, A.: Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. J. Comput. Methods Progr. Biomed. **104**, 443–451 (2011)
36. Varatharajan, R., Manogaran, G., Priyan, M.K., Balaş, V.E., Barna, C.: Visual analysis of geospatial habitat suitability model based on inverse distance weighting with paired comparison analysis. Multimed. Tools Appl. **24**, 1–21 (2017)
37. Gonçalves, L.B., et al.: Inverted hierarchical neuro-fuzzy BSP system: a novel neuro-fuzzy model for pattern classification and rule extraction in databases. J. IEEE Trans. Syst. Man Cybernet. **36**(2), 236–248 (2006)
38. Austin, P.C., et al.: Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes". J. Clin. Epidemiol. **66**, 398–407 (2013)
39. Turan, R.G., et al.: improved functional activity of bone marrow derived circulating progenitor cells after intra coronary freshly isolated bone marrow cells transplantation in patients with ischemic heart disease. J. Stem Cell Rev. Rep. **7**, 646–656 (2011)



Chandra Babu Gokulnath is pursuing a Ph.D in the School of Information Technology and Engineering, Vellore Institute of Technology University. He received is Bachelor of Engineering and Master of Engineering degree from VelTech University and Vellore Institute of Technology University, respectively. His current research interests include Big Data Analytics, IoT, IoE, IoV in Healthcare. He has published number of international journals and conferences.

S. P. Shantharajah is working as a professor in the School of Information Technology and Engineering, Vellore Institute of Technology University. His current research interests include Big Data Analytics and Wireless Networks. He has published number of international journals and conferences.