International Journal of Engineering

Journal Homepage: www.ije.ir

RESEARCH NOTE

A Random Forest Classifier based on Genetic Algorithm for Cardiovascular Diseases Diagnosis

S. Kumar*, G. Sahoo

Department of Computer Science & Engineering, B.I.T,Mesra,Ranchi, India

P A P E R I N F O

Paper history:
Received 06 April 2017
Received in revised form 28 May 2017
Accepted 08 September 2017

Keywords:
Random Forest
Genetic Algorithm
Feature Selection
Cardiovascular Disease

A B S T R A C T

Machine learning-based classification techniques provide support for the decision making process in the field of healthcare, especially in disease diagnosis, prognosis and screening. Healthcare datasets are voluminous in nature and their high dimensionality problem comprises in terms of slower learning rate and higher computational cost. Feature selection is expected to deal with the high dimensionality of datasets in terms of reduced feature set. Feature selection improves the performance of classification accuracy particularly performing with less number of features in decision making process. In this paper, Random Forest (RF) is employed for the diagnosis of cardiovascular disease. The first phase of the proposed system aims at constructing various feature selection algorithms such as Principal Component Analysis (PCA), Relief- F, Sequential Forward Floating Search (SFFS), Sequential Backward Floating Search (SBFS) and Genetic Algorithm (GA) for reducing the dimension of cardiovascular disease dataset. The second phase switched to model construction based on RF algorithm for cardiovascular disease classification. The outcome shows that the combination with GA and RF delivered the highest classification accuracy of 93.2% by the help of six features.

***doi**: 10.5829/ije.2017.30.11b.13*

1. INTRODUCTION

Computer based diagnosis contributing a significant role to support physicians and medical professionals from many years. Medical diagnoses require not only the specific details of a patient dataset but also a past experience of the physicians [1]. Medical diagnoses based on machine learning algorithm are increasingly introduced and being applied to solve complex problems by many researchers in the area of bioinformatics and pattern recognition [2]. Medical datasets shows the high dimensionality and the complex relationship among different features of specific disease makes it difficult for the classification. The high dimensionality problem often leads to cause of lower classification accuracy. The accuracy of the specific classifier obtained in disease diagnosis is the vital issue to be considered by the researchers. Most of the datasets applied for the disease classification and prediction are of high dimensional in nature [3]. To deal with high dimensional data, descriptive features are extracted with the help of feature selection methods and hence dimension of datasets is reduced [4]. In this context, feature reduction plays an important role in achieving higher classification performance in terms of high accuracy and lower computational time. Dimension reduction is also useful to remove the irrelevant features of datasets and decreasing the datasets complexity. The current research work is focused to select an optimal feature subset from cardiovascular disease dataset in order to improve diagnosis accuracy. This work is focused towards trade-off between computational time and quality of produced feature subset solutions.

Cardiovascular disease (CVD) is a disease that concerned with heart and blood vessels. CVD includes coronary artery disease such as myocardial infarction, which is commonly known as heart disease [5]. Therefore, the highlight of this work is to examine the efficiency of Random Forest (RF) in accomplishing the CVD diagnostic problem. This work is based on RF classification to improve the classification accuracy for cardiovascular disease diagnosis. The difference

*Corresponding Author's Email: san77j@gmail.com (S. Kumar)

Please cite this article as: S. Kumar, G. Sahoo, A Random Forest Classifier based on Genetic Algorithm for Cardiovascular Diseases Diagnosis, International Journal of Engineering (IJE), TRANSACTIONS B: Applications Vol. 30, No. 11, (November 2017) 1723-1729

between this study and other existing work that depicts the superiority of Genetic Algorithm (GA) based RF classifier, where feature selection is performed by GA and the extracted features are taken as input to the RF classifier. Furthermore, the proposed method shows more efficient outcomes compared to the other existing methods being compared.

The framework of this paper is as follows: introduction and related work are discussed in sections 1 and 2. Feature selection method and random forest (RF) classifier are described in section 3. The evaluating procedure is illustrated in section 4. The experimental details and description of cardiovascular disease dataset are presented in section 5 and conclusive explanation are included in section 6.

2. RELATED WORK

Identifying relevant feature from the specific datasets has been considered as a high priority optimization problem among the research community. Feature selection is a well-known technique that applied to remove redundant and irrelevant features for disease classification. In this study, various feature selection techniques are studied which are used for cardiovascular disease classification. Shilaskar and Ghatol [6] applied the hybrid feature selection techniques for dimension reduction and enhancement in classification accuracy. In this work, forward selection techniques are used for cardiovascular disease diagnosis and their results are compared with forward inclusion and back-elimination techniques. Inbarani et al. [7] have presented the supervised feature selection techniques based on the hybridization of PSO (particle swam optimization) based Relative Reduct (PSO-RR) and PSO based Quick Reduct (PSO-QR) for disease diagnosis. The experimental result shows the efficiency of the proposed techniques as well as enhancement in accuracies. Liu et al. [8] have proposed the hybrid classification system based on Relief-F and Rough set (RFRS) techniques. In the RFRS system, Relief-F is employed for feature extraction, where heuristic Rough set method for feature reduction. A C4.5 based ensemble classifier is applied on feature set for heart disease classification. A maximum classification accuracy of 92.59% was achieved to validate the superiority of the classifier.

Classification of diseases in terms of cardiovascular disease is a widespread research area in the literature. In medicine, lack of information, imprecision and contradictory facts creates difficulties in decision process. Fuzzy logic based expert system are constructed to deal with imprecise information of problem specification. Fuzzy logic emerges as a suitable tool related to decision making problem in various real life problems [9]. Polat and Gunes [10] has applied the k-nearest neighbor (k-NN) based weighting scheme for a preprocessing task and artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism is used as a classifier for cardiovascular disease diagnosis. The obtained accuracy of proposed system was 87%. Shouman et al. [11] have proposed a decision tree (j4.8) based classification in the diagnosis of heart disease patients. The outperforming model achieves the 84.1% accuracy. Das et al. [12] have proposed a neural network ensemble method that creates new model by combining the posterior probabilities. This method obtained 89.01% classification accuracy on UCI heart dataset.

3. GENETIC ALGORITHM AND RANDOM FOREST: BASICS

This section presents a concise description of essential structure of genetic algorithm, random forest and their definitions.

3. 1. Genetic Algorithm Genetic algorithm is best known to solve optimization problems [13], such as [14] where genetic algorithm is used to optimize the generated rule structure produced by various classifiers and also produces the optimized feature set.

A genetic algorithm is an unconventional search or optimization procedure applied on a population of P individuals where individuals are categorized by chromosomes C_k, k = (1,...,P). The Chromosomes consists of multiple strings of symbols, which is known as genes $C_k = C_{k1},..,C_{kn}$ and we can write N as the length of string. Individuals are evaluated based on their respective fitness function. To grow further in uninterrupted manner, genetic algorithm accomplishes the task by three fundamental operators: selection, crossover, and mutation [14]. The role of selection operator is to select the best fitness valued individuals from current generation to get survive in successive generation. Crossover is the process of combining two parents to produce their children. Mutation function makes small alteration in a particular element of genes from the population, which provides more ability to produce the solution of optimization problems. Mutation is also helping to preserve the genetic diversity and restricted the search to fall in wrong direction.

$$fitness = \frac{number\ of\ correctly\ classified\ instances}{number\ of\ training\ samples} \quad (1)$$

Genetic algorithm searches the best optimized solution during chromosomes evolution, in terms of the specified fitness function [15]. The aim of fitness function is to trim down the number of random forest trees and minimize the rate of learning error simultaneously.

3. 2. Random Forest Random forest is emerging as a most popular and powerful ensemble techniques in pattern recognition application for high dimensional and complex problems. There is a limitation with tree classifier having their high variance. The methodology was presented by few researchers such as Amit and Geman [16] and Breiman [17], in an integrated form known as random forest. Random forest is ensemble of decision tree which contains several classification methods and different parameters settings. The principle of random forest based on the method of constructing a forest of uncorrelated trees using a CART procedure, combined with bagging and "random feature selected" technique. Breiman [17] presented the previously known and some novel techniques altogether for modern practice of random forest, in particular:

- Applying out of bag error to evaluate the generalization error
- Variables are measured through permutation

In particular decision trees has ability to grown very deep in terms of learn highly irregular patterns, that's lead a cause of over-fitting problem. Decision tree training sets have very high variance. Hence, Random forest are the solution of averaging multiple deep decision tree, trained on different part of the same training set, with promising goal of reducing the variance. This approaches works well at the expenses of small increase in the bias, but greatly boosts the performance of the classifier. Taking learning set K= ((X_1, Y_1).... (X_n, Y_n)), made of n vectors, $X_n \in M$ where M is a set of numerical observation and $Y_n \in N$, where N is a set of class labels. In solving the classification problem, a particular classifier is a mapping $M \rightarrow N$. The considered new input vectors are classified by each individual tree of the forest. Each tree gives up a definite classification results. The principle of random forests is to build binary sub-trees using the training bootstrap samples coming from learning samples S and selecting randomly at each node a subset of M. The most voted by all the trees in a forest is chosen as class of decision forest. The random forest principle works on Breiman bagging method. Bagging method is abbreviated as bootstrap aggregation an ensemble learning method introduced by the Breiman [17], employed to improve the accuracy of weak classifier by building a set of classifier. The bootstrapping method achieves the better performance because it reduced the variance of the classifier without increasing the bias. This shows the prediction of a single tree is highly sensitive to noise compared to the average of many trees, unless trees are not correlated.

Random forest can be applied to rank the variables in classification problem. In first phase evaluating the variable interest in datasets $D_n = (X_i, Y_i)$, $i=1$, n is fit a random forest to the data. During the fitting process the out-of-bag error for each data point is recorded and averaged over the forest. Typically, for a classification problem with p features, $\sqrt{p}$ (rounded down) features are used in each split.

4. PROPOSED METHOD

The proposed system for diagnosis of CVD is presented in Figure 1. The primary objective of this study is to design a method that improves the classification accuracy and also obtain the important features which are able to indicate the cardiovascular disease class. Here, genetic algorithm is employed to find optimal features for diagnosis of cardiovascular disease. The hybridized system comprises two phases that include feature selection and classification. The feature selection method employed the genetic algorithm to find out the optimal features from large and high-dimensional datasets. While for the classification phase random forest technique is applied to improve the prediction accuracy and decreases the variance. A chromosome consists of 13 genes and each represents an input variable. In general, genetic algorithm adopts bit coding with fixed length; the most common of use is the binary code; this method uses a string which is constituted by the symbol to denote an individual. Each code responding to a condition attribute and the attribute value will determine the encoding length.

For example, an attribute has kinds of value (the continuous attributes need discretized firstly), so the individual coding will distribute bits for it and each bit corresponds to the possible values.

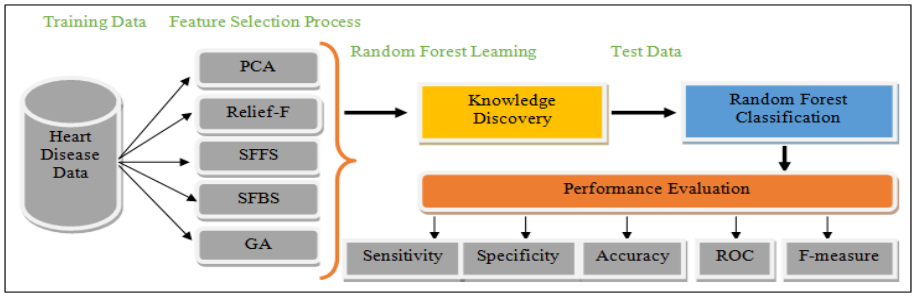


Figure 1. Random forest combined with feature selection algorithms for cardiovascular disease diagnosis

When the value is 0, it means that the individual will not take the attribute value. When the value is 1, the individual will take the attribute value. Transformation of this method is simple and each Chromosome has fixed length. However, the random forest has a feature which the node has not only discrete attributes, but also numerical attributes. The evaluation of the fitness of an input variable subset was performed by 10-fold cross validation based on the subset evaluator function. The initial population comprises of 60 chromosomes that iterated to grow through a maximum of 60 generations. The experimental parameter for crossover and mutation was set at default value of 0.9 and 0.05, respectively.

5. RESULTS AND DISCUSSION

In this section, performance of Random Forest classifier and feature selection method are presented. The experiments were carried out on Intel(R) i3 CPU 2.40 GHz personal computer with Window 8.1, 64 bit and simulation was done on Matlab R2017a.

5. 1. Dataset The datasets consist of 303 cases provided by the UCI repository including 76 attributes, but only 13 variables are of use [18]. Six cases are encountered as missing value. The task of classification is to detect the heart patient. There are 164 cases listed out of 303 in dataset having no heart disease, while rest 139 people having heart disease varying level from 1 to 4.Our experiment concentrated on trying to separate values 1,2,3,4 from absence (value 0). Moreover, all the published articles only refer to 13 attributes are such as age, sex, chest pain type (four values), resting blood pressure, serum cholestoral in mg/dl, fasting blood sugar>120 mg/dl, resting electrocardiographic results (values 0, 1 and 2), maximum heart rate achieved, exercise induced angina, old peak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, thal: 3 = normal; 6 = fixed defect and 7 = reversable defect.

5. 2. Performance Analysis The Popular performance indices such as accuracy, specificity, sensitivity, F-measure was considered to evaluate the performance of RF classifier. The formulations of these popular performance indices are as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

$$Specificity = \frac{TN}{FP+TN} \times 100\% \quad (4)$$

Precision is the proportion of positive test results that are true positives. It is a critical measure of the performance of a diagnostic method, as it reflects the probability that a positive test reflects the underlying condition being tested for.

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (5)$$

In Equations (1)-(4) the abbreviated short term TP, FN, TN and FP are used for number of true positive, number of false negative, number of true negative and number of false positive respectively. These terms are defined as a confusion matrix. Another performance matrice considered to evaluate the classification accuracy is Mathews Correlation Coefficient (MCC) [19], which characterizes to evaluate the imbalanced positive and negative samples in the datasets. It is given as follows:

$$MCC = \frac{TP.TN-FN.FP}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \quad (6)$$

5. 3. Feature Selection and Classification Accuracy Feature selection has essential role in construction of classification systems. Feature selection method reduces the dimension of data, but also reduce the computational time of a classifier. Feature selection algorithms improve the classifier in terms of accuracy and computational cost. In this work, we have employed six feature selection methods such as PCA [20], Relief-F [21], Sequential Forward Floating search (SFFS), Sequential Backward Floating Search (SBFS) [22] and genetic algorithm. Results of all five respective feature selection algorithms are listed in Table 1. As listed from Table 1, the reduced dimension of cardiovascular disease dataset required less storage space for smoothly program execution. The outcome of feature selection process is directly processed as an input to the random forest (RF) classifier.

In order to get consistent approximations of accuracy on five different classification task, a 10-fold cross validation has been performed. The shown accuracy is the average outcome of 10-fold.

TABLE 1. Selected features of cardiovascular disease dataset

Feature Selection method	Number of selected attributes	Selected attributes
PCA	13	(1-13)
Relief-F	7	1,3,5,8,9,12,13
SFFS	4	3,5,10,12
SFBS	3	3,5,13
GA	6	3,4,9,10,12,13

In this work, random forest classifier provide the greater predictive accuracy than single-tree model but also have a disadvantage as a black box model which cannot visualize the decision tree forest structure. The accuracy of prediction depends upon the size of the decision tree forest. The larger the size of decision tree forest, more have the predictive accuracy. Random forest associated with two types of size control mechanism (1) total number of trees in the random forest and (2) the size of the respective tree. Some of the research reveals that it is best to grow very large size of trees, so the maximum level should be set large trees and minimum node size control would limit the size of the trees. Hence, maximum levels of trees were adjusted at 50 trees.

As shown in Table 2, the random forest classifier get the 84.8, 85.4, 79.1.85.8% of accuracies for PCA, Relief-F, SFFS, SBFS. The proposed classifier GA-RF achieves the 93.2, 90.5, 91.9% for accuracy, sensitivity, specificity respectively. Therefore, the outcomes validate the efficiency of proposed GA-RF model. Similarly the value of AUC and MCC for GA-RF obtained 96.8, 90.7% respectively. Hence, these competent values establish the usefulness of the GA-RF strategy. Figure 2 illustrates the performances of RF model with different feature selection strategy and comparative analysis depicted the clear difference among them.

The performance of classifier often depends upon the quality of feature selected by the individual feature selection method. The Cohen's kappa statistics (KS) is one of the performance analysis metrics, which is used for performance analysis of classifiers. Kappa statistic is used to measure the agreement between predicted and actual value of a datasets, while correcting the agreement that occurs by chance [23].

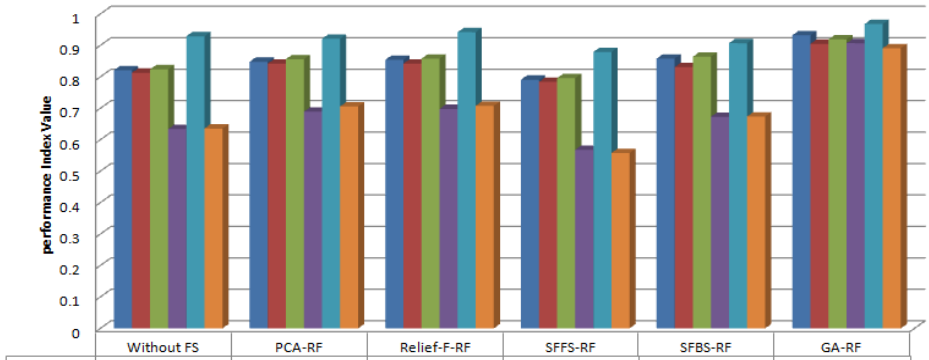
$$k_s = \frac{p_o - p_c}{1 - p_c} \quad (7)$$

where, P_o is total agreement probability, and P_c is the hypothetical probability of chance agreement.

The results obtained shows that the GA-RF method has produced very promising outcomes on the classification of two-class datasets in classifying the possible cardiovascular disease patents. Table 3 shows the accuracy comparison of cardiovascular disease classification by various classical data mining techniques. Table 4 shows the accuracies comparison of previous studies. The proposed classifier GA-RF has highest accuracy among other previous works by various researchers (Table 4).

TABLE 2. Performance evaluation of RF classifier combined with different feature selection algorithm

Performance Index	Without FS	PCA-RF	Relief-F-RF	SFFS-RF	SFBS-RF	GA-RF
Accuracy	0.821	0.848	0.854	0.791	0.858	**0.932**
Sensitivity	0.813	0.842	0.842	0.784	0.831	0.905
Specificity	0.824	0.856	0.858	0.796	0.864	0.919
MCC	0.634	0.689	0.698	0.568	0.673	0.907
ROC	0.929	0.921	0.942	0.879	0.907	0.968
KS	0.6355	0.7055	0.7079	0.5578	0.6739	0.891



	Without FS	PCA-RF	Relief-F-RF	SFFS-RF	SFBS-RF	GA-RF
Accuracy	0.821	0.848	0.854	0.791	0.858	0.932
Sensitivity	0.813	0.842	0.842	0.784	0.831	0.905
Specificity	0.824	0.856	0.858	0.796	0.864	0.919
MCC	0.634	0.689	0.698	0.568	0.673	0.907
ROC	0.929	0.921	0.942	0.879	0.907	0.968
KS	0.6355	0.7055	0.7079	0.5578	0.6739	0.891

Figure 2. Performance indices comparisons for random forest classifiers combined with the feature selected algorithm

TABLE 3. Accuracy comparison for cardiovascular disease dataset

Methods	Accuracy (%)
Naive bayes	79.45
C4.5	80.64
Decision Table	84.58
Neural Network	84.87
Our approach	**93.20**

TABLE 4. Comparison of GA-RF method with studies in literature

Author	Method	Accuracy (%)
Shilaskar and Ghatol [6]	Feature selection with forward selection techniques and comparing results with forward inclusion and back elimination techniques, SVM as a classifier	85
Liu et al. [8]	Feature selection and reduction methods with Relief-F and Rough set (RFRS), and C4.5 decision tree classifiers	92.59
Polat and Gunes [10]	Fuzzy-Artificial immune recognition system (AIRS)	87
Nguyen et al. [24]	Feature selection with wavelet and fuzzy SAM, Genetic algorithm is applied for classification	78.78
Bashir and Qamar [25]	Novel classifier ensemble framework based on enhanced bagging approach with multi-objective weighted voting scheme	84.16
Jabbar et al. [26]	RF and GA	83.70
Elyan and Gaber [27]	RF and GA	83.96
This study	GA-RF	**93.20**

Table 1 examined the selected features by the various feature selection methods applied on the cardiovascular disease dataset. CVD achieved the highest accuracy based on the optimum features selected with GA. The outcome illustrates the efficiency of selected features to produce the cardiovascular dataset's class information. Random forest retains many features of decision tree while achieving the better results through the usage of bagging on samples, majority voting schemes and random subsets of variables. It works well in presence of missing values, and shows the efficiency with the variety of variables (discrete, continuous, and binary) makes it ideal for high dimensional data modeling. Unlike decision trees, there is no need to prune the trees in random forest classification while bootstrapping techniques helps it to overcome on over-fitting issues. It is believed that the proposed optimized RF techniques can be helpful to the physician in cardiovascular disease treatment.

6. CONCLUSION

Classification of two-class classification problems such as cardiovascular disease dataset is a significant issue in pattern recognition applications. The medical data mining is applied to extract hidden information using data mining techniques. Data mining techniques applied on heart disease dataset to extract useful information and RF is employed for better classification. RF is ensemble method that combines the prediction of many individual tree models to provide more accurate prediction than individual classifier. In this work, a hybrid classification system for diagnosis of cardiovascular disease called GA-RF is proposed. Therefore, GA is utilizing for dimension reduction for cardiovascular disease and RF is employed for intelligent classification. The primary goal of this system is to employ the RF classification on unique features including fast learning speed, better classification accuracy and simple to implement for the diagnosis of cardiovascular disease. The proposed GA-RF system has been compared with the other combination with RF such as PCA, Relief-F, SFFS, and SFBS. The performance of classifier is evaluated in terms of accuracy, sensitivity, specificity and AUC. The outcome shows that RF achieves 84.8, 85.4, 79.1, 85.8% classification accuracy for PCA, Relief-F, SFFS and SBFS respectively. The proposed model has achieved a significant accuracy of 93.2, 90.5, 91.9% for accuracy, sensitivity, specificity respectively. Hence, these accuracy is verified the efficiency of GA-RF by the AUC, and MCC which achieves the 0.968 and 0.907, respectively. The features of cardiovascular dataset have reduced significantly from 13 to 6 using Genetic Algorithm. Experimental outcomes show the efficiency of proposed model for the cardiovascular disease. These results illustrated that Genetic Algorithm can be apply for dimension reduction and GA-RF model can be implemented for other disease diagnosis. The future direction of the proposed model is its application for other medical diagnosis problems. In addition, instead of RF, other classification algorithm can be used and tested with other optimization techniques.

7. REFERENCES

1. Koh, H.C. and Tan, G., "Data mining applications in healthcare", ***Journal of Healthcare Information Management***, Vol. 19, No. 2, (2011), 65-73.
2. Dietterich, T.G., "Ensemble methods in machine learning", ***Multiple Classifier Systems***, Vol. 1857, (2000), 1-15.
3. Van Der Maaten, L., Postma, E. and Van den Herik, J., "Dimensionality reduction: A comparative", ***The Journal of Machine Learning Research,*** Vol. 10, (2009), 66-71.
4. Guyon, I. and Elisseeff, A., "An introduction to variable and feature selection", ***Journal of Machine Learning Research***, Vol. 3, No. Mar, (2003), 1157-1182.

5. Organization, W.H., "Prevention of cardiovascular disease: Guidelines for assessment and management of cardiovascular risk, World Health Organization, (2007), ISBN: 9789241547178
6. Shilaskar, S. and Ghatol, A., "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases", ***Expert Systems with Applications***, Vol. 40, No. 10, (2013), 4146-4153.
7. Inbarani, H.H., Azar, A.T. and Jothi, G., "Supervised hybrid feature selection based on pso and rough sets for medical diagnosis", ***Computer Methods and Programs in Biomedicine***, Vol. 113, No. 1, (2014), 175-185.
8. Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q. and Wang, Q., "A hybrid classification system for heart disease diagnosis based on the rfrs method", ***Computational and Mathematical Methods in Medicine***, Vol. 2017, (2017).
9. Shafiee-Chafi, M. and Gholizade-Narm, H., "A novel fuzzy based method for heart rate variability prediction", ***International Journal of Engineering-Transactions A: Basics***, Vol. 27, No. 7, (2014), 1041.
10. Polat, K., Sahan, S. and Gunes, S., "Automatic detection of heart disease using an artificial immune recognition system (airs) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing", ***Expert Systems with Applications***, Vol. 32, No. 2, (2007), 625-631.
11. Shouman, M., Turner, T. and Stocker, R., "Using decision tree for diagnosing heart disease patients", in Proceedings of the Ninth Australasian Data Mining Conference-Volume 121, Australian Computer Society, Inc. (2011), 23-30.
12. Das, R., Turkoglu, I. and Sengur, A., "Effective diagnosis of heart disease through neural networks ensembles", ***Expert Systems with Applications***, Vol. 36, No. 4, (2009), 7675-7680.
13. Holland, J.H., "Genetic algorithms", ***Scientific American***, Vol. 267, No. 1, (1992), 66-73.
14. Azar, A.T., Elshazly, H.I., Hassanien, A.E. and Elkorany, A.M., "A random forest classifier for lymph diseases", ***Computer Methods and Programs in Biomedicine***, Vol. 113, No. 2, (2014), 465-473.
15. Elsayed, S.M., Sarker, R.A. and Essam, D.L., "A new genetic algorithm for solving optimization problems", ***Engineering Applications of Artificial Intelligence***, Vol. 27, (2014), 57-69.
16. Amit, Y. and Geman, D., "Shape quantization and recognition with randomized trees", ***Neural Computation***, Vol. 9, No. 7, (1997), 1545-1588.
17. Breiman, L., "Random forests", ***Machine Learning***, Vol. 45, No. 1, (2001), 5-32.
18. Newman, D., Hettich, S., Blake, C., Merz, C. and Aha, D., "Uci repository of machine learning databases. Department of information and computer science, university of california, irvine, ca", in 1998 of Conference, http://archive. ics. uci. edu/ml/datasets. html., (1998).
19. Powers, D.M., "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation", Vol., No., (2011).
20. Yang, T.-N. and Wang, S.-D., "Robust algorithms for principal component analysis", ***Pattern Recognition Letters***, Vol. 20, No. 9, (1999), 927-933.
21. Kira, K. and Rendell, L.A., "A practical approach to feature selection", in Proceedings of the ninth international workshop on Machine learning., (1992), 249-256.
22. Pudil, P., Novovicova, J. and Kittler, J., "Floating search methods in feature selection", ***Pattern Recognition Letters***, Vol. 15, No. 11, (1994), 1119-1125.
23. Donner, A., Shoukri, M.M., Klar, N. and Bartfay, E., "Testing the equality of two dependent kappa statistics", ***Statistics in Medicine***, Vol. 19, No. 3, (2000), 373-387.

A Random Forest Classifier based on Genetic Algorithm for Cardiovascular Diseases Diagnosis

RESEARCH NOTE

S. Kumar, G. Sahoo

Department of Computer Science & Engineering, B.I.T,Mesra,Ranchi, India

P A P E R I N F O

Paper history:
Received 06 April 2017
Received in revised form 28 May 2017
Accepted 08 September 2017

Keywords:
Random Forest
Genetic Algorithm
Feature Selection
Cardiovascular Disease

چکیده

روش های طبقه بندی مبتنی بر یادگیری ماشین، از فرآیند تصمیم گیری در زمینه مراقبت های بهداشتی، به ویژه در تشخیص بیماری، پیش آگهی و غربالگری حمایت می کند. مجموعه داده های مراقبت های بهداشتی به طور طبیعی در مقیاس وسیع هستند و مشکل بزرگ بودن آنها شامل نرخ یادگیری کمتر و هزینه های محاسباتی بالاتر است. انتظار می رود که انتخاب ویژگی با ابعاد بالاتری از مجموعه داده ها از لحاظ تنظیم ویژگی های کاهش یافته باشد. انتخاب ویژگی عملکرد دقت طبقه بندی را به ویژه با انجام تعداد کمتر از ویژگی های در روند تصمیم گیری بهبود می بخشد. در این مقاله، فارست تصادفی (RF) برای تشخیص بیماری قلبی عروقی مورد استفاده قرار می گیرد. هدف فاز اول سیستم پیشنهادی، ساخت الگوریتم های انتخابی گوناگون مانند تجزیه و تحلیل مولفه های اصلی (PCA) ، Relief-F، جستجو شناور متوالی مستقیم (SFFS) ، جستجو به صورت شناور متوالی بازگشت به عقب (SBFS) و الگوریتم ژنتیک (GA) برای کاهش بعد مجموعه داده های بیماری های قلبی عروقی است. فاز دوم، ساخت مدل بر اساس الگوریتم RF برای طبقه بندی بیماری های قلبی عروقی تغییر یافت. نتیجه نشان می دهد که ترکیب با GA و RF بالاترین ضریب طبقه بندی ۹۳/۲٪ را با کمک شش ویژگی ارائه می کند.

***doi**: 10.5829/ije.2017.30.11b.13*