IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Review of Computer Engineering Studies
Vol. 5, No. 3, September, 2018, pp. 45-49
Journal homepage: http://iieta.org/Journals/RCES

Classification of heart disease using multiple classifiers

Marco Alfonse

Faculty of Computers and Information Sciences, Ain Shams University, Cairo 11566, Egypt

Corresponding Author Email: marco@fcis.asu.edu.eg

https://doi.org/10.18280/rces.050301

Received: 20 August 2018
Accepted: 28 September 2018

Keywords:
heart disease, classification, multilayer perceptron, K-Nearest Neighbor (K-NN), C4.5

ABSTRACT

Heart disease is amongst the most widely recognized diseases in the world. This research aims to consolidate the precision of heart disease classification/diagnosis by developing a system depending on multiple classifiers. The proposed system contains two phases, which are the preprocessing phase and the classification phase. The preprocessing phase includes data cleaning, normalization and accounting for missing values. In the classification phase, multiple classifiers are used as an ensemble technique based on the Multilayer Perceptron (MLP), K-Nearest Neighbor (K-NN) and C4.5. A heart disease dataset, which contains four databases and gathered from the UCI machine learning repository, was used for experiments. The proposed classification system gives 99.4% classification precision according to 10-fold cross-validation technique. The outcome obtained from the proposed system shows that its performance is better than that of already reported classification systems.

1. INTRODUCTION

Heart disease [1-2] is a term used for disorders that may influence the heart and it is also called cardiac disease. Also, "heart disease" is often named "cardiovascular disease". Cardiovascular disease refers to many conditions that include blocked or narrowed blood-vessels that may lead to a heart-attack, stroke, or angina (chest pain). The heart conditions influencing the heart's muscle, rhythm or valves are also considered a sort of heart disease. Also, the cardiovascular disease is considered a prime reason for disability. The WHO [2] and CDC [3] reported that the disease of the heart is the major source of death in the Great Britain, the USA, Australia and Canada. There are 26.6 million US adults that are diagnosed with a heart disease (11.3% of the population of the adult). The heart disease caused 23.5% of deaths in America nowadays.

In the context of computer science, classification is the issue of finding the category (class) of a new instance from a group of classes in accordance to a dataset (it is called a training set) that contains instances with their categories. Classification is a sort of supervised learning which is a learning where each item in its training set is associated with its category. Many classifiers are used in this context such as the Neural Networks (NN) [4], Decision Trees (DT) [5] and so on.

This paper presents a proposed system for categorizing the heart disease. In what follows, the various techniques of data mining and the different methods of machine learning used for heart disease classification are presented.

In [6], the author provides a study of heart disease diagnosis using several techniques of data mining. The author shows that the classification precision is not as intended when a single method of data mining is used separately so to boost the value of accuracy, it is better to combine various techniques of data mining together. When the Support Vector Machines (SVM) and Genetic Algorithms (GA) are combined together, the classification precision reaches 95%, which is higher than other techniques when used individually.

In [7], four classifiers (C5.0, SVM, K-NN and NN) are used to categorize the dataset of heart disease. The classifiers were implemented on a UCI dataset. The data used is partitioned into a testing set (30%) and a training dataset (70%). The training set is applied for constructing the classifier while the testing dataset is used for validating it. From the results obtained, the C5.0 decision tree gives the higher accuracy, which is 93.02%. The accuracies resulted from the remaining classifiers are lower than the C5.0. The K-NN gives 88.37% accuracy, SVM gives 86.05% accuracy and finally NN gives 80.23% accuracy.

In [8], numerous studies have been presented. These studies applied different techniques (which are Naïve Bayes (NB), NN and DT) for the categorization of heart disease and they achieved different accuracies. Also, the authors presented a prediction system of heart disease that is based on NB algorithm.

In [9], many techniques, which have been applied in the categorization of heart disease, have been discussed. Three algorithms (NB, NN and DT) are implemented on a dataset that contains a training dataset (303 records) and a testing set (270 records). From the results obtained, the NN gives better achievement for such a non-linear problem of categorization of heart disease.

In [10], the authors implemented a multi-stage categorization system of the heart disease database. First, the heart disease data is preprocessed by the removal of the repeated records. Second, the algorithm of K-means is applied on the data that are preprocessed with K=2. Third, the MAFIA algorithm is applied (usually used for mining the maximally frequent itemsets that are contained in a transactional database), and fourth, as an optional step, a decision tree is implemented as a prediction model.

In [11], the author develops a prediction system of the heart

disease that is rely on the Naive Bayesian classification and Jelinek-mercer smoothing technique. The author proves that the Naïve Bayes with Jelinek-mercer smoothing is more effective than the Naive Bayes for categorizing the patients with heart disease. The system allows for incorporating more records or attributes and new rules may be produced for categorizing the heart disease so as to enhance the precision of the classification.

The paper organization is as follows; section 2 contains the dataset used for classifying the heart disease, section 3 explains the proposed classification method, section 4 shows the obtained outcomes and discussions, and finally section 5 contains the conclusions.

2. THE HEART DISEASE DATASET

The dataset used in this research contains four databases concerning the categorization of heart disease. The data was gathered from the following locations:

- The Cleveland Clinic Foundation.
- The Cardiology Institute at Budapest.
- The VA Medical Center at Long Beach, CA.
- The University Hospital, Switzerland.

These databases are obtainable from the UCI machine learning warehouse [12]. All the databases have the same format of instances. Only 14 attributes, from the 76 databases' attributes, are rightly used for the categorization of heart disease. In table 1, these 14 attributes information are demonstrated.

Table 1. The 14 heart disease databases' attributes information

No.	Name	Description	Range/Value
1	age	The age of the patient in years	
2	sex	The gender of the patient	• 0: female • 1: male
3	cp	The kind of chest pain	• 1: typical angina • 2: atypical angina • 3: non-anginal pain • 4: asymptomatic
4	trestbps	The resting pressure of blood	in mm Hg
5	chol	The serum cholestoral	in mg/dl
6	fbs	The fasting sugar of blood > 120 mg/dl	• 0: false • 1: true
7	restecg	The resting electrocardiographic outcomes	• 0: normal • 1: abnormality of ST-T wave • 2: definite or probable left ventricular hypertrophy by Estes' criteria
8	thalach	The heart rate that is maximally achieved	
9	exang	The practice induced angina	• 0: no • 1: yes
10	oldpeak	The ST depression produced by practice relative to rest	
11	slope	The peak exercise ST segment slope	• 1: upsloping • 2: flat • 3: downsloping
12	ca	The no. of major vessels which is flourosopy colored	0 - 3
13	thal		• 3: normal • 6: fixed defect • 7: reversable defect
14	num	The predicted attribute, it is the diagnosis (the status of angiographic disease)	• 0: < 50% diameter narrowing • 1: > 50% diameter narrowing

The "num" field is the predicted attribute, which refers to the existence/absence of a heart disease in the patient. It is a numeric value (integer) valued from zero (no presence) to four. Almost all the experiments, that are done by the researchers, only seeking to distinguish the existence of the disease (values 1,2,3 and 4) from its absence (value 0).

This dataset has 920 instances; table 2 presents the databases' class distribution.

Table 2. The dataset class distribution

Database/Class	0	1	2	3	4	Total
Cleveland	164	55	36	35	13	303
Hungarian	188	37	26	28	15	294
Switzerland	8	48	32	30	5	123
Long Beach VA	51	56	41	42	10	200

The Cleveland dataset has 303 instances, the Hungarian dataset has 294 instances, the Switzerland dataset has 123 instances and the Long Beach VA dataset has 200 instances. This dataset has missing attribute values.

3. THE PROPOSED CLASSIFICATION APPROACH

This Proposed classification approach uses three classifiers, which are the MLP [13], K-NN [14] and C4.5 [15] to boost the classification precision of heart disease. This ensemble technique is used to categorize the heart disease dataset as either healthy or sick (this is the reason for which the values 1,2,3 and 4 of the "num" attribute in the used dataset are exchanged with the string value "sick" and the value 0 is replaced with the string value "healthy" in the experiments that are done by the author). The proposed methodology has three stages; data preprocessing, classification and validation (see figure 1). The first two stages (data preprocessing and classification) are expounded in the subsections (3.1 and 3.2) while the validation stage is expounded in the upcoming section (section 4).

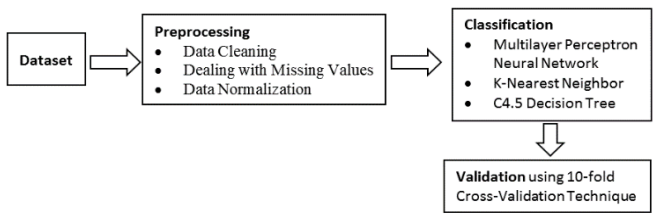

Figure 1. The proposed heart disease categorization system

The following subsections (3.1 and 3.2) explain the data preprocessing and the classification steps in details. The proposed classification methodology is stated in figure 2.

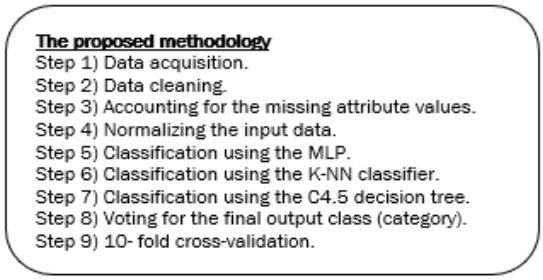

Figure 2. The proposed heart disease categorization methodology

The heart disease categorization methodology includes many stages; data acqusition, data cleaining, accounting for the missed attributes, normalizing the data, classification and evaluation.

3.1. Data preprocessing

The preprocessing of the dataset are accomplished through the following steps:

- Combining the four databases into a dataset and removing the duplicated records if exist.
- Accounting for missing data values.
- Normalizing the values that represent the attributes' information enclosed in the dataset.

Initially the four databases (Cleveland, Hungarian, Switzerland, and Long Beach VA) are combined into a dataset and any duplicated records are removed. The produced dataset has almost 900 records.

The produced dataset has some missing attribute values. These values are identified and replaced with appropriate values by scanning all the dataset records and substituting the missing values with the mean value [16].

For normalizing (scaling) the values that represent the attributes' information enclosed in the dataset, each attribute value is exchanged with a value into the range between 0 to 1 according to the following formula (Eqs. 1-4) [17]:

$$D = Xmax - Xmin \quad (1)$$

$$C = (X - Xmin) / D \quad (2)$$

$$M = 1 / D \quad (3)$$

$$Y = M X + C \quad (4)$$

where Xmax and Xmin are the maximum and the minimum values of the feature being normalized respectively, X is the input attribute value and Y is its corresponding output value. The data normalization (scaling) is a very important step for classification, it can fasten the process of training the NN and minimize the chances of being stuck in, what is called, local optima. After preprocessing the data, multiple classifiers are used for classification.

3.2. Classification

After preprocessing the data, multiple classifiers are used and a voting is applied to find what the output (category) should be. The classifiers used in this voting process are; MLP, K-NN and C4.5 decision tree. Voting is an ensemble algorithm used for classification. In voting, every classifier makes a vote (prediction) for each test record and the final category (class) is the one that is received by all (or at least 2) of these classifiers.

The first classifier is the MLP. The MLP is a sort of NN (a feedforward model) that has 3 or more layers; one input layer, a hidden layer or more and a single layer for output. Every layer consists of processing units, which are named neurons. The learning in the NN occurs by changing the weights of the processing units' connections after presenting each segment of data to the network, in accordance to the error estimate occurred in the output [18]. The error value in the output node j for the nth training example is calculated as in Eq. (5).

$$e_j(n) = d_j(n) - y_j(n) \quad (5)$$

where the desired (target) value is represented by the symbol d and y represents the value resulted by the perceptron. The node weights are changed, based on corrections, so as to make the error as low as possible in the output value, this error is calculated as in Eq. (6).

$$\mathcal{E}(n) = \frac{1}{2}\sum_j e_j^2(n) \quad (6)$$

The network developed in the developed system has a learning rate value 0.5, a momentum 0.2, 2 hidden layers with four hidden nodes in every layer and 5000 epochs are performed for training the NN through.

The second classifier is the K-NN. The K-NN classifier is a fundamental method of classification that is used if there is either no prior knowledge or little knowledge concerning the distribution of the input data. The classifier is ordinarily uses the Euclidean distance between the particular training samples and a test sample. Consider x_i be an input record that has p features (x_{i1},x_{i2},…,x_{ip}) and the input records' count is n (i=1,2,…,n), the Euclidean distance between sample x_i and x_l is calculated as in Eq(7).

$$d(\mathbf{x}_i, \mathbf{x}_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + (x_{ip} - x_{lp})^2} \quad (7)$$

The K-NN assigns to a test instance the majority category label of its k closest (nearest) training instances. In practice, k is defined to be odd so as to avoid ties. In the applied algorithm, the value chosen for K is five, p=13 and n=900.

The third classifier is the C4.5. The C4.5 is applied for constructing decision trees that are used for classification purposes; it is called a statistical classifier. C4.5 generates decision trees using a dataset (training data) by applying the entropy concept. The training dataset used consists of a set S= s_1, s_2 ,… of formerly classified instances. Each instance s_i consists of 2 parts; a vector of 13-dimension ($X_{1,i}$, $X_{2,i}$, ….$X_{13,i}$) where x_j is the feature value, and the category (class) to which s_i belongs. The algorithm chooses the data attribute that breaks the instances into subsets effectively. To divide the data, the entropy difference (the normalized information gain) is used. The feature that has the maximum value of normalized information gain, is selected to divide the

dataset. Then the steps are repeated for the smaller subgroups until all of the records in a subset have a similar class, in this case the algorithm creates a leaf node (the output class).

4. RESULTS AND DISCUSSION

The developed classification system performance has been examined with four medical databases. The source of these databases is the UCI repository. As a method to test the developed classification system, the 10 fold cross validation is applied. The cross validation is used to test the predictive systems' performance. The evaluation is done by partitioning the dataset into two sets; a training dataset for training the system, and a testing set for evaluating it. In 10 fold cross validation, the available data is partitioned randomly into ten subgroups (subsets) of equal size. From the ten subsets, a subset is retained for assessing the system (the validation data) and the remaining nine subgroups are used for system training. This operation is repeated ten times (the folds), in which every subset of the 10-subsets is used rightly once for validating the system. The 10-results (from the folds) are averaged for obtaining a single assessment. The cross-validation advantage is that all the data observations (instances) are used for the training procedure and the validation procedure of the system, and each data instance is used for the validation process exactly once. Table 3 manifests the comparison between the proposed classification system and other approaches. The classification precision of heart disease is improved by the developed classification system considering that only one database (the Cleveland database) is used by the authors discussed in this literature while the developed classification system uses four heart disease databases including the Cleveland database.

Table 3. The comparison between different approaches of heart disease prediction/classification

Approach/ Techniques used	Accuracy (%)
CART Classifier	83
SVM Classifier with Genetic Algorithms	95
C5.0	93.02
NN	89.4
SVM	86.05
K-NN	80.23
C4.5	84.1
Naïve Bayesian Classification Technique	81
Fuzzy Mechanism [6]	94.11
C5.0 Decision Tree	93.02
K-NN	88.37
SVM	86.05
NN [7]	80.23
NN	99
DT	98
NB [8]	90
NB	95
NN	99
DT [9]	97
K-Means with MAFIA	74
K-Means with MAFIA and ID3	83
K-Means with MAFIA, ID3 and C4.5 [10]	89
Classification by Naïve Bayes	78
Classification using Laplace Smoothing [11]	86
The Proposed Approach	99.4

From table 3, it is noticed that the DT and NN give the highest accuracy in classifying the heart disease but with the proposed approach, which uses the MLP, the K-NN and the C4.5, the accuracy is improved significantly.

5. CONCLUSIONS

In this paper, a (Multilayer Perceptron, C4.5 and K-NN) based categorization system is proposed for heart disease. As a method to validate the developed system, it is examined against databases that are collected from the UCI repository. The experimental results, performed on four databases, show that the proposed approach is a competitive method for the categorization of heart disease. The developed system gives 99.4% classification accuracy, which is outstanding comparing to the existing techniques that are explored in the literature. The results proved that the proposed classification system is a reliable alternative system for the categorization of heart disease.

REFERENCES

[1] The Mayo Foundation for Medical Education and Research (MFMER). http://www.mayoclinic.org/diseases-conditions/heart-disease/basics/definition/con-20034056, accessed on June 2018.

[2] The World Health Organization (WHO). http://www.who.int/en/, accessed on June 2018.

[3] The Centers for Disease Control and Prevention (CDC). https://www.cdc.gov/, accessed on June 2018.

[4] Haykin S. (2016). Neural networks and learning machines. Pearson Education Dorling Kindersley. 3rd edition.

[5] Rokach L, Maimon OZ. (2014). Data mining with decision trees: Theory and applications, series in machine perception and artificial intelligence. World Scientific Publishing Company, 2nd Edition.

[6] Manimekalai K. (2016). Prediction of heart diseases using data mining techniques. International Journal of Innovative Research in Computer and Communication Engineering 4(2): 2161-2168. https://doi.org/10.17485/ijst/2016/v9i39/102078

[7] Abdar M, Kalhori SRN, Sutikno T, Subroto IMI, Arji G. (2015). Comparing performance of data mining algorithms in prediction heart diseases. International Journal of Electrical and Computer Engineering (IJECE) 5(6): 1569-1576.

[8] Ratnaparkhi D, Mahajan T, Jadhav V. (2015). Heart disease prediction system using data mining technique. International Research Journal of Engineering and Technology (IRJET) 2(8): 1553-1555.

[9] Dewan A, Sharma M. (2015). Prediction of heart disease using a hybrid technique in data mining classification. Proceedings of IEEE 2nd International Conference on Computing for Sustainable Global Development, pp. 704-706.

[10] Karthiga G, Preethi C, Devi RDH. (2014). Heart disease analysis system using data mining techniques. International Journal of Innovative Research in Science Engineering and Technology 3(Sp.3): 3101-3105.

[11] Patil RR. (2014). Heart disease prediction system using naïve bayes and Jelinek-mercer smoothing. International

Journal of Advanced Research in Computer and Communication Engineering 3(5): 6787-6789.
[12] The UCI Machine learning Repository, Center for Machine Learning and Intelligent Systems, Heart Disease Data Set. http://archive.ics.uci.edu/ml/datasets/heart+Disease, accessed June 2018.
[13] Du KL, Swamy MNS. (2014). Neural networks and statistical learning. Springer.
[14] Harrington P. (2012). Machine learning in action. Manning Publications, 1st Edition.
[15] Hssina B, Merbouha A, Ezzikouri H, Erritali M. (2014). A comparative study of decision tree ID3 and C4.5. International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, pp. 13-19.
[16] Viswanathan S, Viswanathan V. (2015). R data analysis cookbook. Packt Publishing.
[17] Rani KU. (2011). Analysis of heart diseases dataset using neural network approach. International Journal of Data Mining and Knowledge Management Process (IJDKP) 1(5): 1-8. https://doi.org/10.5121/ijdkp.2011.1501
[18] Meda S, Bhogapathi RB. (2018). Identification of heart disease using fuzzy neural genetic algorithm with data mining techniques. Advances in Modelling and Analysis B 61(2): 99-105. https://doi.org/10.18280/ama_b.610208