CORONARY ARTERY DISEASE (CAD) PREDICTION AND CLASSIFICATION - A SURVEY

Rajkumar R.[1] and Anandakumar K.[2] and Bharathi A.[3]
[1]Department of Computer Applications, Sri Krishna Arts and Science College, Coimbatore, India
[2]Department of Computer Applications, Bannari Amman Institute of Technology, Tamil Nadu, India
[3]Department of IT, Bannari Amman Institute of Technology, Tamil Nadu, India
E-Mail: rajkumar.feb12@gmail.com

ABSTRACT

Among many major dangerous diseases, Coronary artery disease (CAD) is considered as an important disease, because it can lead to sudden cardiac death. Manual checking is highly impossible to diagnose for this disease. To predict CAD several approaches have been carried out. This comparative study paper presented a thorough reviews on various approaches made towards prediction of heart diseases. Several data mining and soft computing approaches are studied. This study concludes that the performance comparison of accuracy, sensitivity and specificity of several algorithms and approaches. This research can be done in risk assessment among diabetic patients those who are developing heart diseases.

Keywords: heart diseases, weighted fuzzy rule, K- nearest neighbor, genetic, scoring system, PRAA, SVM classifier, support vector machines, particle swarm optimization.

INTRODUCTION

Data mining is the methodology of discovering beforehand obscure patterns and patterns in databases and utilizing the data to fabricate prescient models. In social insurance, data mining is getting to be progressively prominent, if not if not progressively more essential. Medicinal services industry today produces expansive measure of complex Data about patients, medicinal centers assets, sickness analysis, electronic patient records, therapeutic gadgets, and so on. The extensive measure of data is a key asset to be handled and investigated for learning extraction that empowers help for expense funds and choice making. Data mining gives a set of devices and strategies that can be connected to this transformed data to find shrouded examples furthermore gives medicinal services experts an extra wellspring of learning for deciding.

According to the World Health Organization heart disease is the first leading cause of death in high and low income countries and occurs almost equally in men and women [World Health Organization., 2011]. By the year 2030, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs) [World Health Organization., 2005]. Cardiovascular diseases (CVDs), also on the rise, comprise a major portion of non communicable diseases. In the year 2010, of all projected worldwide deaths, 23 million are expected to be because of cardiovascular diseases.

This paper discusses about the data mining algorithms used to predict the heart diseases which was proposed by various authors.

K- Nearest Neighbor and Genetic Algorithm

M. Akhil jabbar *et al*., 2013 deals with classifying the heart disease. It combines the approach of KNN and genetic algorithm to improve the classification accuracy of heart disease data set.

The authors used genetic search as a goodness measure to prune redundant and irrelevant attributes, and to rank the attributes which contribute more towards classification. Least ranked attributes are removed, and classification algorithm was built based on evaluated attributes. This classifier was trained to classify heart disease data set as either healthy or sick. This algorithm consists of two parts.

Table-1. Accuracy comparison.

Data set name	Accuracy without GA (Knn only)	Accuracy with GA (Knn + GA)
Weather Data Set	85.71	100
Breast Cancer	90	94.35
Heart Stalog	100	100
Lympography	99.32	100
Hypothyroid	100	10
Primary Tumor	75	75.8
Heart Disease A.P	95	100
Averager	92.14	95.73

1. First part deals with evaluating attributes using genetic search

2. Part two deals with building classifier and measuring accuracy of the classifier is shown below:

a) Load the data set
b) Apply genetic search on the data set
c) Attributes are ranked based on their value
d) Select the subset of higher ranked attributes
e) Apply (KNN+GA) on the subset of attributes that maximizes classification accuracy
f) vi. Calculate accuracy of the classifier, which measures the ability of the classifier to correctly classify unknown sample

Step 1 to 4 comes under part 1 which deals with attributes and their ranking. Step 5 is used to build the classifier and step 6 records the accuracy of the classifier. Accuracy of the classifier is computed as

Accuracy= (No. of samples correctly classified in test data) / (Total no. of samples in the test data)

The performance results given by M. Akhil jabbar *et al*., 2013 is shown in Table-1.

Weighted fuzzy rules

P.K. Anooj, 2012 deals with weighted fuzzy rule-based clinical decision support system (CDSS) for computer-aided diagnosis of the heart disease. Authors of P.K. Anooj, 2012 used a weighted fuzzy rule to predict the heart disease. The high priority given to the task of generating fuzzy rules from the data described using numeric–symbolic values appears to be very difficult. Handling these types of values is extremely important because it is very close to human knowledge and rules with such values are normally more comprehensible and accountable when compared to rules with numerical values.

Table-2. Performance of the CDSS in risk prediction.

Data sets	Class	Metric	Proposed system	
			Training	Testing
Cleve-land	<50%	Ac	0.509901	0.623529
		Se	0.724771	0.765957
		Sp	0.258065	0.447368
	>50%	Ac	0.509901	0.623529
		Se	0.258065	0.447368
		Sp	0.724771	0.765957
Hung-arian	<50%	Ac	0.715045	0.469388
		Se	0.8	0.31746
		Sp	0.540984	0.742857
	>50%	Ac	0.715045	0.469388
		Se	0.540984	0.742857
		Sp	0.8	0.31746
Swizer-land	<50%	Ac	0.364706	0.512195
		Se	0.625	0.333333
		Sp	0.337662	0.526316
	>50%	Ac	0.364706	0.512195
		Se	0.337662	0.526316
		Sp	0.625	0.333333

Handling such values is permitted by the introduction of fuzzy set theory which by the construction of fuzzy leads to the generation of a set of fuzzy rules. The automatic method is based on the construction of fuzzy modalities that enables the generation of fuzzy values from a set of rules with numerical values. The decision rules obtained from the previous contain IF and THEN parts, in which IF part specifies the numerical variable and THEN part specifies the class label. At first, the numerical variable specified in the IF part of the decision rules is converted into the linguistic variable according to the fuzzy membership function and THEN part of the fuzzy rules is similar to that of the decision rules. The author have tested his weighted fuzzy rule on different data set and given the result (shown in Table-2).

For example, ‘‘IF β^((1)) is LOW, THEN the risk is less than 50 (class ’0’) and ‘‘IF β^((1)) is MEDIUM, THEN the risk is either less than 50 or greater than 50 (class ‘0’ or ‘1’) and ‘‘IF b(1) is HIGH, THEN the risk is greater than 50 (class ‘1’)’’. In a similar way, it processes the entire decision rules with numeric variable and they are converted into the fuzzy rules using membership function. A group of fuzzy IF–THEN rules obtained is belonging to one of the most popular, most effective, and user-friendly knowledge representations so as to provide the effective learning for the fuzzy system.

Scoring system

Nan Liu *et al.*, 2012 deals with risk score prediction system with HRV parameters and vital signs, in which geometric distance serves as the key component. They used prediction model named “scoring system” to compute a risk score on a patient’s clinical outcome, utilizing both HRV parameters and vital signs. The scoring system was built based on the calculation of geometric distances among a set of feature vectors obtained from the records of multiple patients. The score prediction algorithm is summarized below and the details are elaborated in the following.

Algorithm: Geometric distance based score prediction
Input

- HRV parameters and vital signs of N training patients and one testing patient xt where each patient is a sample in the database.
- Patients’ hospital records and characteristics.

(1) Variable selection

- Select HRV parameters and vital signs to form a feature vector to represent each patient’s health condition.
- Transform both training and testing feature vectors into the interval [−1, 1] by min-max normalization.

(2) Initial score calculation

- Obtain the cluster center Cp of the positive class (patients with cardiac arrest within 72 hours) in the Euclidean feature space.
- Calculate distance Dp, which is the distance between Cp and one positive training sample nearest to Cp.
- Calculate distance Dn, which is the distance between Cp and one negative training sample farthest to Cp.
- Calculate distance Dt between the testing sample xt and Cp and compute the initial score based on Dp and Dn.

(3) Classification based score updating

- Predict binary outcome with SVM classifier and obtain the number of positive samples Np within K neighbors.
- Implement predefined rules for score updating.

Output

- Predictive risk score on the clinical outcome.

Table-3. Risk prediction using feature extraction.

Measure	Proposed	SVM - LIN	SVM - RBF	GLM
Sens	78.80%	73.01%	61.05%	63.05%
Spec	80.80%	80.01%	80.08%	80.08%

Table-3 shows the comparison of their proposed algorithm with SVM with linear kernel (SVM-LIN), SVM with RBF kernel (SVM-RBF) and generalized linear model (GLM).

Genetic-SVM

E. Avci., 2009 deals with the wavelet entropy computing in the discrete wavelet transform layer of GSVM which can be performed for robust feature extraction against to noise from DHS signals.

An intelligent system based on genetic-support vector machines (GSVM) approach was proposed for classification of the Doppler signals of the heart valve diseases. This intelligent system deals with combination of the feature extraction and classification from measured Doppler signal waveforms at the heart valve using the Doppler ultrasound. GSVM is used in this study for diagnosis of the heart valve diseases. The GSVM selects of most appropriate wavelet filter type for problem, wavelet entropy parameter, the optimal kernel function type, kernel function parameter, and soft margin constant C penalty parameter of support vector machines (SVM) classifier. The performance of the GSVM system is evaluated in 215 samples. The test results show that this GSVM system is effective to detect Doppler heart sounds. The averaged rate of correct classification rate was about 95%.

A chromosome of GSVM represents a kernel function type, a value of related kernel parameter, and a value of C parameter. Thus, a chromosome of GSVM consists of 28 bits (genes). The kernel function type part of a chromosome in GSVM is represented by 3 bits.

a) The value of C parameter part of a chromosome in GSVM is represented by 3 bits.
b) The value of RBF kernel parameter (r) part of a chromosome in GSVM is represented by 7 bits.
c) The value of the polynomial kernel parameter (d) part of a chromosome in GSVM is represented by 2 bits.
d) The value of the sigmoid kernel parameter (d1) part of a chromosome in GSVM is represented by 2 bits.
e) The value of the bspline kernel parameter (d) part of a chromosome in GSVM is represented by 2 bits.
f) The value of the ERBF kernel parameter (r) part of a chromosome in GSVM is represented by 7 bits.
g) The value of the Fourier kernel parameter (d) part of a chromosome in GSVM is represented by 2 bits.

Table-4. Performance comparison of intelligent system.

The Kernel function types		Value of Kernel function parameter	Value of C parameter	The average recognition (%)	
				Normal	Abnormal
GSVM Model -1	RBF	36.8	0.1	90	93.15
GSVM Model -2	RBF	18.3	0.1	88	89.04
GSVM Model -3	RBF	24.5	0.1	92	94.52
GSVM Model -4	ERBF	2.2	0.1	96	94.52
GSVM Model -5	ERBF	2.2	10,000	94	94.52
WNN used in Turkoglu et al., (2003)				95.9	84

The author have compared the performance of Genetic SVM with Turkoglu *et al.* 2003

Least square support vector machines

Davut Hanbay., 2009 deals with the interpretation of the DHS signals using pattern recognition.

An expert system based on least squares support vector machines (LS-SVM) for diagnosis of valvular heart disease (VHD) was presented. Wavelet packet decomposition (WPD) and fast-Fourier transform (FFT) methods are used for feature extracting from Doppler signals. LS-SVM is used in the classification stage. Threefold cross-validation method is used to evaluate the expert system performance.

The performances of the developed systems were evaluated in 105 samples that contain 39 normal and 66 abnormal subjects for mitral valve disease. The results showed that this system is effective to detect Doppler heart sounds. The average correct classification rate was about 96.13% for normal subjects and abnormal subjects.

The objective of the classification is to demonstrate the effectiveness of the feature extraction method from the DHS signals. For this purpose, the feature vectors were applied as the input to an LS-SVM classifier. The KULeuven's LS-SVMlab MATLAB/C Toolbox was used for the purpose of training and testing. RBF kernel is used. Grid search algorithm is used to tune the c regularization constant and r width of RBF kernel parameters. The determined optimal c value is 2.8439 and optimal r value is 29.316 for predicting mitral valve diseases. Threefold cross-validation method was applied to the 105 experimental data sets for computing the validation of LS-SVM model. In k-fold cross-validation method, the data set is divided into k subsets, and the holdout method is repeated k times. At each time, k - 1 subsets are used for training and kth subset is used for testing. Then the average error across all k trials is computed. Therefore, every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times. Different evaluation methods were used for calculating the performance of the expert system. The best test performance of LS-SVM model is shown in the below table.

RESULTS

Table-5. Threefold test performance of LS-SVM model.

Data set (105)		Correct classified	Wrong classified	Performance (%)
Training sets	Test sets			
Set-1, set-2 (70)	Set-3 (35)	35	0	100
Set-1, set-2 (70)	Set-2 (35)	33	2	94.2
Set-1, set-2 (70)	Set-1(35)	33	2	94.2
Average Performance				96.13

LS-SVM model predicts the measured values at a high accuracy rate. Threefold test performance of LS-SVM model is shown in Table-5. The average correct classification rate is 96.13%

An enhanced version of binary particle swarm optimization, designed to cope with premature convergence of the BPSO algorithm. MBPSO control the swarm variability using the velocity and the similarity between best swarm solutions. It uses support vector machines in a wrapper approach, where the kernel parameters are optimized at the same time. The approach is applied to predict the outcome (survived or deceased) of patients with septic shock. Further, MBPSO is tested in several benchmark datasets and is compared with other PSO based algorithms and genetic algorithms (GA).

Their experimental results show that this approach can correctly select the discriminating input features and also achieve high classification accuracy, especially when compared to other PSO based algorithms.

When compared to GA, MBPSO is similar in terms of accuracy, but the subset solutions have less selected features. The modified binary PSO includes the SVM model parameters in the encoding of the particles and is described in the following.

The MBPSO combines the described reset swarm best mechanism with a local search operator that displaces the particle best position x^{pb} and uses an operator similar to the mutation mechanism, but instead of using a probability of mutation, the change in the particles are made controlling the value of v_{max}. These modified mechanisms were first introduced and are used by the MBPSO, once these mechanisms have shown better performance.

MBPSO algorithm is given below:

```
Initialize algorithm parameters
Normalize and divide data (train + test)
Randomly initialize swarm particles
while (number of iterations or stopping criteria are not met) do
    Evaluate the fitness of all the individuals
    for i = 1 to number of particles do
        if fitness of xi is greater than the fitness of $x^{pb}$ then
            $x_i^{pb}$ = $x_i$
        end if
        if fitness of $x_i$ is greater than the fitness of $x^{sb}$ then
            $x^{sb} = x_i$
        end if
        if $x^{sb}$ constant for $I_{max}$ iterations then
            reset $x^{sb}$
            update $x^{pb}$ using displacement rate dr
        end if
        for j = 1 to dimension of particle's position do
            $v_{ij} \leftarrow wv_{ij} + c_1 q \left(\frac{x_{ij}^{pb} - x_{ij}}{\Delta t}\right) + c_2 r \left(\frac{x_j^{sb} - x_{ij}}{\Delta t}\right) +$
            if $|v_{ij}| > v_{max}$ then
                $v_{ij} = v_{max} \times sgn(v_{ij})$
            end if
            $s(v_{ij}) = \frac{1}{1+e^{-v_{ij}}}$
            if (r < $s(v_{ij})$) then
                $x_{ij} = 1$ else $x_{ij} = 0$
            end if
        end for
        Introduce variability in the swarm
    end for
end while
```

Table-6. Performance comparison of MBPSO.

Method	Accuracy	NF	Sensitivity	Specificity
No-FS	89.0 ± 1.7	28	76.1 ± 5.4	95.6 ± 1.7
BPSO	94.0 ± 1.5	6 ± 1	89.5 ± 5.6	96.1 ± 2.0
IBPSO	94.2 ± 1.1	6 ± 1	90.4 ± 5.3	95.9 ± 1.8
GA	95.7 ± 1.4	7 ± 1	94.3 ± 1.2	96.5 ± 2.1
MBPSO	94.4 ± 1.2	6 ± 1	90.2 ± 5.1	96.5 ±1.9

The authors have compared their MBPSO with existing techniques and showed that MBPSO outperforms in feature selection used in the prediction of heart diseases.

CONCLUSIONS

Data mining in healthcare has significantly grown. The applications of data mining and several methodologies tend healthcare informatics a great way to owe its implications. Heart disease prediction among the patients is one of the plunge research areas in healthcare informatics. This survey paper presented a thorough study on various approaches made towards prediction of heart diseases. Several data mining and soft computing approaches are studied. This study concludes that the future scope of research can be done in risk assessment among diabetic patients those who are developing heart diseases.

REFERENCES

[1] Davut Hanbay. 2009. An expert system based on least square support vector machines for diagnosis of the valvular heart disease. Expert Systems with Applications. 36(3): 4232-4238.

[2] E. Avci. 2009. A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier. Expert Systems with Applications. 36(7): 10618-10626.

[3] M. Akhil jabbar, B.L. Deekshatulu, Priti Chandra. 2013. Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm. First International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), Procedia Technology. 10: 85-94.

[4] Nan Liu, Zhiping Lin, Jiuwen Cao, Zhixiong Koh, Tongtong Zhang. 2012. An Intelligent Scoring System and Its Application to Cardiac Arrest Prediction. IEEE Transactions on Information Technology in Biomedicine. 16(6): 1324-1331.

[5] P.K. Anooj. 2011. Clinical decision support system: Risk level prediction of heart disease using weighted

fuzzy rules. Central European Journal of Computer Science, Springer. pp. 482-498.

[6] Susana M. Vieiraa, Luís F. Mendonc, Gonc¸ J. Farinhaa, Joao M.C. Sousaa. 2013. Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. Applied Soft Computing. 13(8): 3494-3504.

[7] Turkoglu I., Arslan A. and ve Ilkay E. 2003. A wavelet neural network for the detection of the heart valve diseases. Computers in Biology and Medicine. 33(4): 319-331.

[8] V. Sree Hari Rao and M. Naresh Kumar. 2013. Novel Approaches for Predicting Risk Factors of Atherosclerosis. IEEE Journal of Biomedical and Health Informatics. 17(1): 183-189.

[9] 2005. World Health Organization Global Report, Preventing Chronic Disease: A Vital Investment.

[10] World Health Organization. 7-Febuary 2011; Available from: http://www.who.int/mediacentre/factsheets/fs310.pdf.