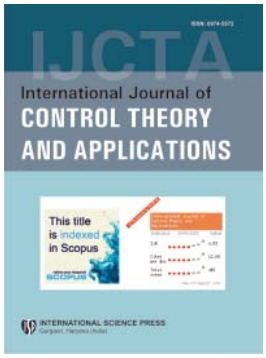

International Journal of Control Theory and Applications

ISSN : 0974–5572

Volume 9 • Number 40 • 2016

Diagnosis of Coronary Artery Diseases using Classification Algorithms Based on Wavelet Transforms

D. Haritha[a] T. Rajesh Kumar[a] and E. Rajesh Kumar[a]

[a]Department of Computer Science & Engineering, K L University, Guntur, Andhra Pradesh, INDIA.
E-mail: t.rajesh61074@gmail.com

Abstract: One of the primary drivers of the death in the world is cardiovascular diseases (CAD) which is a major threat in developing and developed countries. The fundamental drivers in CAD results in clogging of the coronary lumen consequently occlusion, and after that prompts to myocardial dead tissue (MI) or sudden heart attach which causes death. It is difficult to ascertain that a certain person has been affected by CAD, since there are bunch of parameters has been involved to ascertain the conclusion. Classification has been done using wavelet transform to classify the certain parameters. We analyzed following methods such as NB, Logistic, SMO, RBF Network, K-star, Multiclass Classifier, Conjunctive rule, Decision table, DTNB, LAD Tree, LMT, NB Tree, Random forest and Random Tree calculations has been associated with extensive fragment of the surveys. This database has been generated from UCI machine learning database. In this paper, we used 10-fold cross validation with 14 attributes and calculations of TP rate, FP rate, Precision, Recall, F-measure, ROC and Accuracy are analyzed practically. As a result, the Logistic, SMO and LMT algorithms has yield to improve the high accuracy rate of 77.0%.

1. INTRODUCTION

The ethical quality rate of the death due to diseases are much greater than those of accidents and natural disasters. The World Health Organization estimates that 17 million deaths worldwideeach year occur due to cardiovascular diseases [1]. A major type of such diseases is coronary artery disease (CAD, which is reported to account for 7 million deaths over the world per annum [1]. Mining is the extraction of knowledge from a set of data. Otherwise, data mining is a process that uses intelligent techniques whereby knowledge of a set of data can be extracted [2].

Angiography is the methodology of decision for the diagnosing of CAD. Angiography determines the location and extent of the stenotic arteries; nevertheless, its high costs and risks for the patient have prompted researchers to seek less expensive and more effective methods with the aid of data mining. Moreover, Cost-sensitive algorithms can be of huge value inthis field as misclassification of diseased or healthy patients has different costs. Pedreira*et al.*, [3] utilizing a neural system on UCI [4] datasets, achieved a precision rate of 80% for CAD conclusion. Das *et al.*[5] applied Neural Network on the datasets of Cleveland and reported an

accuracy rate of 89.01%. Babaoglu*et al.* [6] used the Support Vector Machine (SVM) Algorithm on an exercise test data and achieved an accuracy rate of 79.17%. Tsipouras*et al.* [7] used the Fuzzy Model to detect CAD. Itchhaporia*etal.* [8] drew upon the Neural Network to examine a practice test information for the conclusion of CAD.

The reason for the present review is to utilize information mining procedures, which is a managed learning calculation, in order to recognize CAD patients from solid people. The motivation behind the present review is to utilize arrangement calculations to be specific, NB, Logistic, SMO, RBF Network, Kstar, Multiclass Classifier, Conjunctive lead, Decision table, DTNB, LAD Tree, LMT, NB Tree, Random woodland and Random Tree methods. The informational index is taken from information mining storehouse of college of California, Irvine (UCI)[4]. At last the framework is approved utilizing informational indexes from Cleveland, Hungarian, Long Beach, Switzerland and from Ipoh Specialist Hospital, Malaysia.

1.1. Wavelet Transform

The Discrete Wavelet Transform (DWT) is a linear signal processing technique that, when applied to a datavector X, transforms it to a numerically different vector X', without affecting the energy of a signal. The discrete wavelet transform uses the idea of dimensionality reduction, which is a multivariate statistical method that stores the compressed approximation of the data under the premise of little loss of information. A compressed approximation of the data can be retained by storing only a small fraction of the user-specified thresholdwavelet coefficients and remaining data as zero. This technique also performs well to remove noise and abnormal datawithout smoothing out the main features of data [11]. DWT algorithmic complexity for an input vector of length *n* is having a worst case of *O(n)*. Discrete wavelet transform can be better applicable at handling data of high dimensionality. The fig.1 shows how wavelet transformation is applied for a sample set of four data.

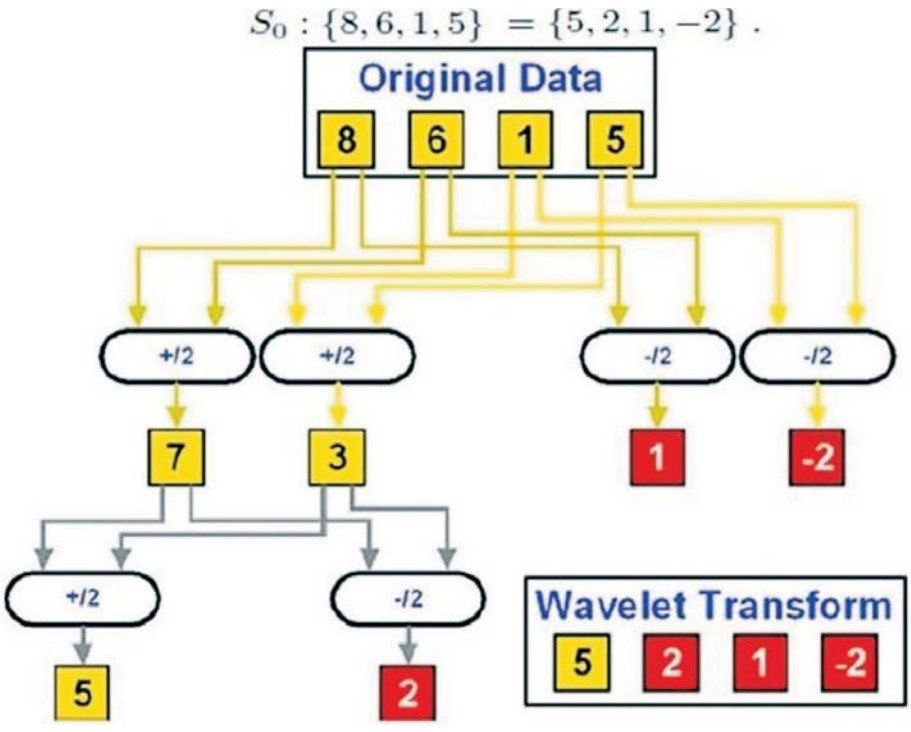

Figure 1: Wavelet transformation of four data points

1.2. Haar Wavelet Transform

A Haar wavelet transform is the simplest type of wavelet [11,12].In discrete form, Haar wavelet transform are related to a mathematical operation called the Haar Transform. The Haar transform serves as a prototype for all other wavelet transforms. A Haar transform decomposes an array into two halves of the originallength of the array. One half is a running average, and the other half is a running difference. Haar transform performs an average and difference on a pair of values.

1.3. Procedure

To Calculate the Haar transform of an array of n samples:

1. Find the average of each pair of samples. (*n*/2 averages)
2. Find the differences between each average and the sample it was calculated from. (*n*/2 differences)
3. Fill the first half of the array with averages.
4. Fill the second half of the array with differences.
5. Repeat the process on the first half of the array.

The rest of this paper is organized as follows: Section 2 describes the medical database for heart diseases problem. The data mining methodology are presented in section 3, In section 4 experimental results of all the classification algorithms are discussed and finally section 5 conclude this paper.

2. HEART DISEASE DATA BASE

In this paper, the fourteen algorithms namely NB, Logistic, SMO, RBF Network, Kstar, Multiclass Classifier, Conjunctive rule, Decision table, DTNB, LAD Tree, LMT, NB Tree, Random forest and Random Tree as tested in a medical datasets for Heart disease database from UCI repository (https://archive.ics.uci.edu/ml/datasets/Heart+Disea se) before applying it on wavelet transforms. This data set contains 303 patients and 54 features and introduces several effective features. Acharya et al. [13] employed gray scale features from left ventricle echo cardio graphic images to classify patients with coronary artery diseases.

2.1. Data Set Description

Coronary artery disease data sets are taken from Data Mining Repository of University of California, Irvine (UCI) [4]. The CAD data sets contain 920 instancescollected from Cleveland, Hungarian, VA Long Beach, andSwitzerland. Coronary angiography determines the result of CAD diagnosis. These data sets have 14 attributes of CADdata.These attributes are listed in given below and brief description of each of these database. 1. age: Age in years, 2.sex: sex (1 = male; 0 = female, 3.*cp*: Value 1: typical angina, Value 2: atypical angina, Value 3:non-anginal pain,Value 4: asymptomatic, 4.chol: serum cholestoral in mg/dl, 5.fbs:(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false), 6.restecg: resting electrocardiographic results, 7.thalach: maximum heart rate achieved, 8. exang: exercise induced angina (1 = yes; 0 = no, 9. ldpeak = ST depression induced by exercise relative to rest, 10. slope: the slope of the peak exercise ST segment, 11. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect, 12. Ca: Number of major vessels colored by fluoroscopy, 13. num: diagnosis of heart disease (angiographic disease status, 14. trestbps: resting blood pressure (in mm Hg on admission to the hospital

Cleveland Data : Cleveland data set was collected by Robert Detrano, M.D. and Ph.D. degrees holder at V.A. Medical Centre [4].

Hungarian Data : AndrasJanosi, M.D. degree holder, collected this data set at the Hungarian Institute of Cardiology, Budapest.The format of this data set is the same as that of the Cleveland data. Class distributions are 37.5%heart disease present and 62.5% heart disease absent [4].

Switzerland Data : This data setwas collecte data the University Hospital, Zurich, Switzerland, by William Steinbrunn, M.D. degree holder. Among four data sets related to CAD, the maximum number of missing value is related to Switzerland data set. It has 123 instances and class distributions in it are 6.5% heart disease absent and 93.5% with heart disease [4].

3. METHODOLOGY

In this paper, we have tested eleven tree based algorithms namely, NB, Logistic, SMO[9], RBF Network, Kstar, Multiclass Classifier, Conjunctive rule [9,10], Decision table, DTNB, LAD Tree, LMT, NB Tree [9], Random forest [9] and Random Tree [9] techniques. We calculated true positive, false positive, precision, recall, F- measure and ROC and selected above eleven classification algorithms to find the highest accuracy rate for cardiovascular diseases (CAD) database. It support all the mining process to get valid and clear visualization with accuracy results, 10 fold cross validation with 14 attributes was applied to the input datasets in this experiment.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this paper, we used experimental results analysis namely, NB, Logistic, SMO, RBF Network, Kstar, Multiclass Classifier, Conjunctive rule, Decision table, DTNB, LAD Tree, LMT, NB Tree, Random forest and Random Tree techniques era compared based on the application in the heart diseases database. Weak data mining tool is being used for research banking, education, weather and real database. We have tested 10 fold cross validation with 14 attributes was used to the input data set in the experiment.

4.1. Performance Measure

The accuracy,true positive, true negative, precision and F-measure are of great significance in the heart diseases medical field. Consequently, for measuring the performance of algorithms, accuracy,true positive, true negative, precision and F-measure were used.

4.2. Confusion matrix

A confusion matrix contains information on actual and predicted classification done by a classification system. The table.1 provides the results of confusion matrix.

Table 1
Confusion matrix

Known Class	*Predicted class*	
	A	*B*
A	True positive (TP)	False Negative (FN)
B	False positive (FP)	True Negative (TN)

1. TP is the number of correct predictions for positive instance.
2. FP is the number of incorrect predications for positive instance.
3. FN is the number of incorrect of predications for negative instance
4. TN is the number of correct of predications for negative instance

4.3. True positive and True negative

True positive and true negative are the ratio of correctly diagnosed CAD and normal samples [9].

$$\text{True Positive} = \frac{TP}{(TP + FN)}$$

$$\text{True Negative} = \frac{TN}{(TN + FP)}$$

4.4. Accuracy

Accuracy is the proportion of the total number of prediction that are correct. It is determined using the following equation [9].

$$\text{Accuracy} = \frac{TP + TN}{(TP + FN + FP + FN)}$$

4.5. Precision

Precision is the ration of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage [9].

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

4.6. Recall

Recall is the ration of the number or relevant records retrieved to the total number of relevant records in the data set. It usually expressed as a percentage [9].

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

4.7. F-measure

F-measure is evaluated by the harmonic mean of both precision and recall. Mathematically

$$\text{F} - \text{measure} = \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

4.8. Receiver operating Characteristics (ROC)

The receiver operating characteristic (ROC) plot is obtained for a binary class classifier by plotting the true -positive over the *y*-axis and false positive over *x*-axis.

Table 2
Confusion matrix and 10 fold cross validation with 14 attributes

Confusion Matrix			*Methods*	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-measure*	*ROC*	*Accuracy*
	AP	AN								
AP	422	78	Naive Bayes	0.76	0.30	0.75	0.76	0.76	0.81	0.76
AN	104	164								
AP	440	60	Logistic	0.77	0.32	0.76	0.77	0.76	0.83	0.77
AN	115	153								
AP	449	51	SMO	0.77	0.33	0.76	0.77	0.76	0.72	0.77
AN	123	145								

Confusion Matrix			Methods	TP Rate	FP Rate	Precision	Recall	F-measure	ROC	Accuracy
AP	434	66	RBF Network	0.75	0.34	0.74	0.75	0.74	0.78	0.75
AN	123	145								
AP	407	93	KStar	0.69	0.41	0.68	0.69	0.69	0.71	0.69
AN	144	124								
AP	440	60	MultiClass Classifier	0.77	0.32	0.76	0.77	0.76	0.83	0.77
AN	115	153								
AP	385	115	Conjunctive Rule	0.68	0.38	0.68	0.68	0.68	0.69	0.68
AN	125	143								
AP	405	95	Decision Table	0.71	0.37	0.70	0.71	0.70	0.77	0.71
AN	126	142								
AP	415	85	DTNB	0.73	0.34	0.73	0.73	0.73	0.79	0.73
AN	116	152								
AP	415	85	LAD Tree	0.74	0.33	0.73	0.74	0.73	0.78	0.74
AN	114	154								
AP	445	55	LMT	0.77	0.32	0.77	0.77	0.76	0.83	0.77
AN	114	154								
AP	409	91	NB Tree	0.73	0.33	0.73	0.73	0.73	0.78	0.73
AN	112	156								
AP	417	83	Random Forest	0.74	0.32	0.74	0.74	0.74	0.81	0.74
AN	110	158								
AP	373	17	Random Tree	0.68	0.37	0.68	0.68	0.68	0.68	0.68
AN	118	150								

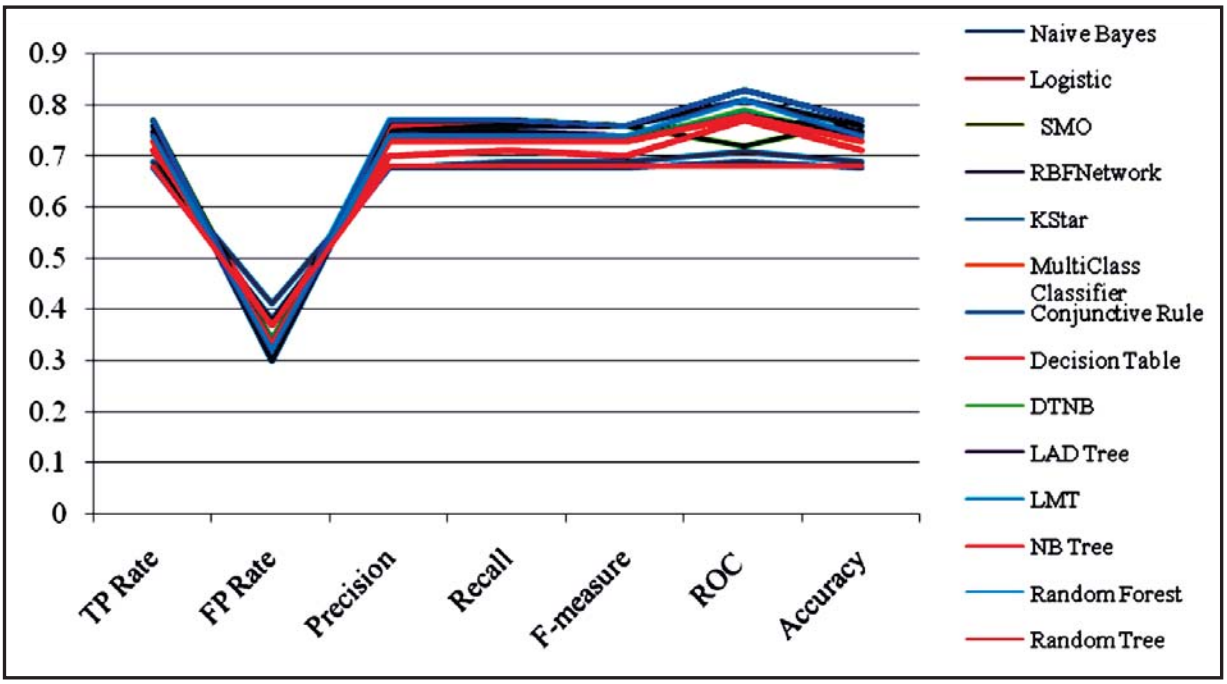

Figure 2: Comparison of 10 fold cross validation with 14 attributes

The Table-2 shows the Confusion matrix and 10 fold cross validation with 14 attributes. The plot is also drawn for the above table's data. Thus the fig.2 stipulates the comparison of 14 classification algorithm. Finally, after comparison of all the algorithm, the three algorithms such as Logistic, SMO and LMT algorithms provide the highest accuracy rate of 77.0% which has been experimentally proved.

5. CONCLUSION

In this paper, classification of 14 algorithms display in view of discrete wavelet transform was exhibited. We had utilized haar wavelet transform as the dimensionality reduction function. We implemented the UCI database for CAD diagnosis and yielded the highest accuracy rate when employed alongside the Logistic, SMO and LMT algorithms 77.0% used in 10 fold cross validation with 14 attributes. In future research, a large database, more attributes and apply the real time application could be used to achieve better results.

REFERENCES

[1] R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, "Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine", 9th edition: New York, Saunders, 2012.

[2] S. Bickel and T. Scheffer, "Multi-view clustering". In Proc. Of the IEEE Int'l Conf. on Data Mining, pp. 19–26, 2004.

[3] C. E. Pedreira, L. Macrini, and E. S. Costa, "Input and Data Selection Applied to Heart Disease Diagnosis", Proceedings of International Joint Conference on Neural Networks, IEEE, 2005.

[4] (https://archive.ics.uci.edu/ml/datasets/Heart+D isease)

[5] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with Applications, pp. 7675–7680, 2009.

[6] I. Babaoglu, O. Fındık, and M. Bayrak, "Effects of principle component analysis on assessment of coronary artery diseases using support vector machine", Expert Systems with Applications, pp. 2182–2185, 2010.

[7] M. Tsipouras, T. Exarchos, D. Fotiadis, A. Kotsia, K.Vakalis, K. Naka, L. Michalis, "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling", IEEE Transactions on information technology in biomedicine , Vol.12, NO.4, pp.447-458, 2008.

[8] D. Itchhaporia, R. Almassy, L. Kaufman, P. Snow, and W. Oetgen, "Artificial neural networks can predict significant coronary disease", J.Am. Coll. Cardiol., Vol.28, NO.2, pp.515-521, 1995.

[9] P.Ganesan, S.Sivakumar, and S.Sundar, An Experimental Analysis of Classification Mining Algorithm For Coronary Artery Disease, "International Journal of Applied Engineering Research",Volume 10, Number 6 (2015) pp. 14467-14477.

[10] P.Ganesan, S.Sivakumar, and S.Sundar, A Comparative Study on MMDBM Classifier Incorporating Various Sorting Procedure ," Indian Journal of Science and Technology " Vol 8(9), 868–874, May 2015.

[11] James S. Walker. 1999. A Primer on Wavelets and Scientific Applications. Jiawei Han, MichelineKamber Data mining: concepts and techniques: Second Edition illustrated. Morgan Kaufmann Publishers, Inc, 2006.

[12] Kiran Kumar Reddi, Ali MirzaMahmood, K.MrithyumjayaRao, "Generating Optimized Decision Tree Based on Discrete Wavelet Transform ", International Journal of Engineering Science and Technology Vol. 2(3), 2010, 157-164.

[13] A.Ben-Hur, and J. Weston, A User's Guide to Support Vector Machines, Methods in Molecular Biology, 2010, pp.223-239.