IMPROVISED NOVEL FUZZY CLUSTERING FOR CARDIAC DIAGNOSIS USING CROSS AMALGAMATION APPROACH

R. KrishanKumar, K. S. Ravichandran and K. R. Sekar
School of Computing, SASTRA University, Thanjavur, Tamil Nadu, India
E-Mail: krishankumar@sastra.ac.in

ABSTRACT

Diseases are growing stronger every now and then coping with the Dalton's theory of Survival of Fittest. To compensate such a drastic growth in vulnerability, medicines are also being well equipped. Scientist and doctors work tirelessly to fix the disease that pop every now and then. One such chronic, life threatening disease is the Cardiac Disease which is more prevalent in all parts of the world. Medical world is working a whole lot of ways to get remedy for this chronic disease. Last decade enormous amount of research had been set on track to find the cause and remedy to heart diseases. The literature signifies that not only is advancement in medical field efficient but also research models in computing field drive more accurate results. In this paper efforts are made to bring computers into medical field by proposing a statistical model. This model is wrapped using an Improvised Novel Fuzzy Clustering (INFC) which is a cross amalgamation of Prism classifier and Fuzzy Clustering methodology. This duo fold scheme drives the model for better accuracy. The methodology is incorporated to a real time dataset from Cleveland Hospital. The attributes taken up for study have a near optimal reach of the expert defined thresholds. The proposed INFC approach is compared with the classical Prism Classifier, Fuzzy Clustering and Actual expert inference. The inferences are graphically depicted based on match status of experimental and actual results.

Keywords: cardiac disease, statistical model, improvised novel fuzzy clustering.

INTRODUCTION

Background

The present era of medical world has seen drastic improvement in disease getting resistant to remedies and vice versa. According to Mendel's genetic study species evolve with time to get accustomed to the dynamically changing environment, so are diseases getting stronger every now and then. To compensate such negative growth the medical world strives hard to find eventual cure by incorporating both pharmacology and computing together as an integrated package. Many research works show the efficient use of computing techniques in disease prediction. Many computing models serve as a key factor for medical base to be built upon for future scope.

There are so many vulnerable, life taking prognoses that evolve during the period. One such chronic malady is the Cardiac Disease. Cardiac arrhythmia is a current study focus. Many math models and statistical methods are proposed to predict its efficacy. Different algorithms for speed, direction and time are estimated for cardiac mapping and its significance is analyzed (C.D. Cantwell *et al.*, 2015). A new approach to classify cardiac arrhythmia disease is proposed using Incremental Back propagation Neural Network and LM classification supported by CFS and linear FSS. The results show better accuracy when tested for UCI data base (Malay Mitra *et al.*, 2013)

This disease is quite common with all living beings under all age groups. Some literatures even show the origin of this cardiac disease as a hereditary trait. Heart is one of the most important and rudimentary building block of human life. It pumps blood to and fro the entire human body. The heart is divided into two parts namely the left heart and the right heart. These two segments has four chambers in total. A block in any one of these chambers may cause intensified effects to human life. There are several reasons for heart failure which scientists have figured out. Some of the important factors for heart diseases are *Age*: As people grow old so does their heart. It sets itself to wear and tear. *Blood Pressure*: Abnormal BP level causes short circuiting in the heart. *Cholesterol:* It is divided as good and bad cholesterol. The increase in bad cholesterol puts heart under immense pressure causing fatality. Cholesterol in blood is a major cause for death. A math model with differential equations is proposed for checking cholesterol rate in liver and in blood streams. Results from analytical relationship, shows the efficacy of diagnosis of high blood cholesterol levels (Hrydziuszko *et al.*, 2014). *Smoking and Alcoholism:* These are external factors that corrode the heart valves causing choking and death. *Heredity:* There are certain heart ailments that are heredity and major contributions of research are going on in order to diagnose and cure such maladies. The two typical catch phrases in CVD are the HFREF and HFPEF. The differences between these two are analysed using 712 samples of FHS when tested in EFFECT. The results show that CHD had low EF (Jennifer E *et al.*, 2012). Gene based research needs active study. Classical gene studies have drawbacks which are overcome using network topologies. Results show that the network topologies gives better gene set understanding on prognoses analysis (Hung JH *et al.*, 2010). The cardio electro physiology is a key aspect of

CVD. Numerical discretisation using high order Galerkin/hp spectral is employed over cardiac electrophysiology. The introduction of high order schemes reduces cost factor and results prove to be significant (ChrisD *et al.*, 2014). Traditional approaches to read through the HD symptoms are ECG and CMRI. ECG is a key mechanism to find cardiac disease. Multivariate maximal time series motif using navie bayes is proposed and is tested for Holter Monitor. Results prove effective when compared with other models (S. Padmavathi *et al.*, 2015). Segmentation of CMRI is a critical and key aspect in heart ailments. An automatic algorithm for segmentation is devised and tested for about 1008 samples. Results prove effective with reduced error rates (Yossi Tsadok *et al.*, 2013). The separation of dysfunction from AT is critical. The major problem is the decoupling of myocardial dysfunction from AT. A mechanical model with Dynamic G meshes are proposed and tested for two cases and results prove to be significant (Jiahe Xi *et al.*, 2013)

One such effort is made in this paper that proposes an innovative statistical model that is wrapped by an INFC methodology. This INFC is a two-fold scheme that sets Prism classifier to be cross amalgamated with Fuzzy clustering. The combination of a classifier with a clustering method increases the level of accuracy in a better manner. The model concentrates on three major attributes from the heterogeneous collection of 14 attributes for study. The 3 attributes are namely Age, BP and cholesterol. Based on expert advice these three attributes are marked as solid factors are all major categories of heart diseases.

RELATED WORKS

Long term research in medical field has been going on in order to predict various newly poping diseases. Scientist and researchers have solved many problems in medical field with the help of computing approaches. Medical data sets are huge sets of data which requires efficient treatment for predicting better results. Data mining and machine learning concepts are used for this purpose. A novel rough set method is proposed and is compared with classical rough sets, SVM and KNN methods and results show better accuracy in regards to back propagation and perceptron methods (S. Udhaya kumar *et al.*, 2015). A Hospital Information System is used to classify patients treated in emergency section. A statistical analysis is devised for over 12 months and results show that the major diseases are cardiac, respiratory and neurological (Yu Xu *et al.*, 2013)

There are a variety of standards that adhere to the CVD. EuroSCORE is one typical standard to measure survival rate after heart surgery. A novel fuzzy EuroSCORE is proposed with three classes for prediction and eight attributes for input from experts. Results show better prediction than the classical model and also a defect density function to find defect rate is proposed (Sina Khanmohammadi *et al.*, 2013). Many machine learning approaches help in diagnosing the HD. An effective scheme to predict HD is to use ECG. A combination of SVM and ANN is used to put ECG into healthy or MI category. A LIBSVM and backpropagation ANN is used for the prediction and results show better sets of classification of ECG signals (Nitin Aji Bhaskar *et al.*, 2015). Automatic classification is the first step of auto wall movement and CADD. Machine learning approach for 200 ECG samples are classified based on PSAX, PLAX, A2C and A4C. Results show effective results (G.N.Balaji *et al.*, 2015). The heart pulse rate is signified by its beat rhythm. Cardiac Sound Signals are analyzed using tuneable Q wavelet transform. FHS and murmurs are reconstructed using CSCW. The least square SVM is used frequency domains and results prove effective (Shivnarayan Patidar *et al.*, 2013). Patient's survival chance after a cardiac surgery is of key interest and is more risk prone. Length of stay prediction of cardiac surgery patients is a key aspect. A scoring system is designed using a Naive Bayes classifier and performance is assessed using cross validation. The method is compared with Bootstrap and results prove to be effective (Paolo Barbini *et al.*, 2014; Sindoori et al., 2013). The psychosocial risk factor is addressed in parts of South Asia. Questionnaires are analysed for 1065 members from South Asia and 818 members from UK and London based on cross sectional design for CHD. The results prove to be better (Emily D. Williams *et al.*, 2010). There are many risk factors that cause CVD. Markers of endothelial dysfunction identified some of the C reactive proteins to cause CVD. Test samples are taken from 2017 patients and a Z score is devised and results show significant relations (Nienke J. Wijnstok *et al.*, 2010).

There are several parameters that govern the CVD. Some of them are critical and some are non-critical. The parameters for metabolic syndrome that caused CVD and PVD are identified and analysed. The test sample is collected from Ibhramin Hospital and people who underwent angiogram are asked to fill questionnaires. Based on the statistical analysis results shows significant improvement (Faria Afsana *et al.*, 2010). The CVD can be cured if detected early. Arterial stiffness is a major cause for CVD. A new approach based on SVM and DWT is used to predict arterial stiffness and help identify CVD at earaly stages. Results show effective results when tested for real samples. (Yingying Zheng *et al.*, 2010)

STATISTICAL MODEL FOR MEDICAL DIAGNOSIS

The Figure-1 depicts A Statistical Model for Cardiac Diagnosis. The model sets itself with three major blocks starting with the Feed Data Set block which is used to feed real time data into the system. The data fed may be in the form of historical, observed or experimental data. For the instance sake a historical data set from Cleveland Hospital is being considered. The next block contains the methodology that is used to typically train the model. The Improvised Novel Fuzzy Clustering (INFC) is an amalgamation of Prism Classifier and Fuzzy Clustering that forms an empirical terminology called *Classering* which is classifier plus clustering methods.

The INFC starts by typically forming rules from the raw data set using Prism Classifier. These rules eventually reduce the huge data sets into finite set matrix. These rules thus generated are taken as centroids for the Fuzzy Clustering. These Centroids are used to find the best match for the test sample patterns in order to cluster them into various clustering bags. The final block shows the Expected Output which is the cluster bag to which the test pattern is more significantly clustered to. The model is trained with the data set and then is validated for its efficacy by feeding test samples. The system is optimally accurate and is also reliable for prediction of prognoses.

The remainder of this article consist of Section 2 for Overview on Methodologies, Section 3 for Results and Discussions and Section 4 for Conclusions.

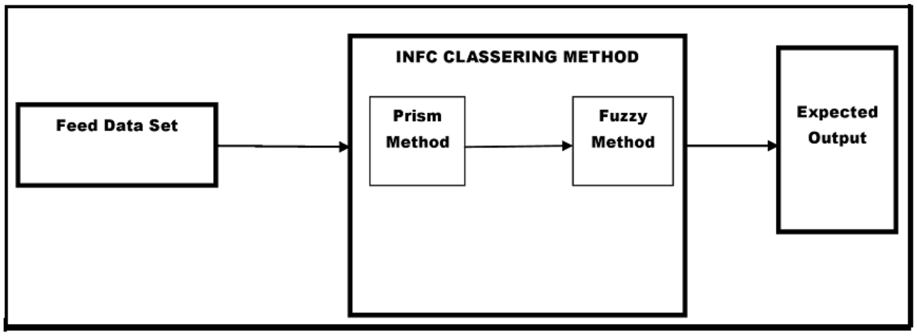

Figure-1. A statistical model for cardiac diagnosis.

OVERVIEW ON METHODOLOGIES

Description of data set

The historical data has been taken for study. Data sets from Cleveland Hospital are formulated for analyzing cardiac diseases. A total of 500 instances are taken up for training the statistical model using INFC method. The trained model is then used for predicting unknown patterns as a part of test phase. There are a total of 14 attributes taken for study out of 76 randomly chosen medical attributes based on metric analysis and expert opinion. Table-1 depicts the Historical Data Sample from Cleveland Hospital. The descriptions of the attributes are made as table legends.

Table-1. Historical data sample from Cleveland hospital.

Age	Sex	Chol	Tresbps	CP	Fbs	Thalach	Exang	Rest ECG	Old peak	Slope	CA	Thal	Num
>55	M	<248	>135	T	T	<156	NO	LVT	>2	D	<1	FD	NH
<55	M	>248	>135	A	F	>156	NO	N	<2	UP	>1	RD	NH
>55	M	>248	<135	NA	F	<156	YES	LVT	<2	FLAT	<1	FD	HD
>55	F	>248	>135	A	F	<156	NO	LVT	<2	FLAT	>1	N	HD
>55	M	<248	>135	AC	F	<156	NO	N	>2	FLAT	>1	FD	NH
>55	M	<248	<135	A	F	>156	NO	N	<2	UP	<1	N	NH
>55	F	>248	>135	AC	F	>156	NO	LVT	>2	D	>1	N	HD
>55	F	>248	<135	AC	F	>156	YES	N	<2	UP	<1	N	NH
>55	M	>248	<135	AC	F	<156	NO	LVT	<2	FLAT	<1	RD	HD
<55	M	<248	>135	AC	T	<156	YES	LVT	>2	D	>1	RD	HD

Chol=Cholesterol (threshold 248), Trestbps=Blood Pressure at Rest (threshold 135), CP=Chest Pain Type (T- Typical, A-Angina, NA- Non Angina, AC- Asymptomatic), Fbs= Fasting blood sugar (T- True, F- False), Thalach= Achieved Cardiac Rate (threshold 156), Exang= Exercise on angina, RestECG= ECG at rest (LVT- Left Ventricular Hypertrophy, N-Normal), Slope (D- Down), CA= vessels colored by fluoroscopy (Threshold 1), Thal= Heart rest state (FD- Fixed Defect, RD- Reverse Defect), Num= Class variable (NH- No Heart Disease, HD- Heart Disease), < signifies less than or equal to, > signifies greater than.

An outlook on prism classifier (Rule based classifier)

The rule based classifier is typically used for large instances of data. Most often it is used for medical data as the data content is enormous for resolving. The classifier summarizes the huge data set into finite set of rules which are used for prediction of the test sample data. The pseudo code for prism classifier is stated below:

```
a) Select a class instance whose probability is high.
b) Select attribute instances corresponding to that class instance by probability.
c) Set it to the condition and proceed to other attributes in the same way by reducing data
d) Cover all attributes for that instance of class
e) Proceed with the uncovered attribute instance for that class
f) Repeat from step 1 for next high probability class
g) All these form a vector that represents the Rule set
```

An outlook on fuzzy clustering (FC)

The fuzzy clustering is more predominant technique in medical field as the data in medical study are typically ambiguous and non-crisp in nature. So there is an emerging need for a fuzzy based approach. This method has the property of resolving ambiguity. The pseudo code and formula for FC is given below.

The formula for Euclidian distance is given in (1).

$$EuclidianDistanceED = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2} \quad (1)$$

Where u- instance of test sample input v- instance of centroid (a vector set)

The formula for Total Distance is given in (2)

$$TotalDistanceD = \sum_{i=1}^{n} 1/EDi \quad (2)$$

Where ED- Euclidian distance instance

The formula for Fuzzy Cluster is given in Equation (3)

$$FuzzyClusterF = \left(\frac{EDi}{D}\right)^{b-1} \quad (3)$$

Where b- bias state, set to value 2 for medical data sets.

```
a. Select centroid as a vector group to form a matrix
b. Select an input vector.
c. Evaluate the Euclidian distance
d. Sum all distances to find total distance
e. Use (3) to evaluate membership value
f. Depending on threshold club the test sample
g. Repeat step 1 for next sample
```

An outlook on improvised novel fuzzy clustering (INFC)

The statistical model is trained using the INFC approach. This technique combines the features of a classifier and a clustering method. The classifier used here is Prism Classifier and the clustering used is the Fuzzy clustering. The amalgamation of these two approaches forms a *"classering approach"* i.e. classification plus clustering approach. The INFC is set to an empirical classering category. The pseudo code for INFC is stated below.

```
a) Feed raw data into Prism classifier for generating rules.
b) The generated rules are taken as centroid values.
c) The test sample is taken as input for validation.
d) Apply fuzzy clustering to club test sample to suitable cluster bags either HD or NH bag.
e) Train the model using INFC for sample data sets
f) Once trained, the model is ready to predict test samples.
g) Stopping condition is set once optimality is gained
h) Repeat step 1 for another set of test sample
```

The Table-2 depicts the Comparative Prediction of Actual and Experimental Results. The actual value is gained from expert opinion and experimental value is obtained from the mining methodologies.

Table-1. Comparative prediction of actual and experimental results.

Patterns	Experimental inference	Actual inference	Status (actual to INFC)
Pattern 1	PC= NH		
	FC=HD	HD	Match
	INFC=HD		
Pattern 2	PC=HD		
	FC=NH	HD	Match
	INFC=HD		
Pattern 3	PC=HD		
	FC=NH	NH	No Match
	INFC=HD		
Pattern 4	PC=NH		
	FC=NH	NH	Match
	INFC=NH		
Pattern 5	PC=NH		
	FC=HD	HD	No Match
	INFC=NH		

PC= Prism Classifier, FC= Fuzzy Classifier, INFC= Improvised Novel Fuzzy Classifier, NH= No Heart Disease, HD= Heart Disease.

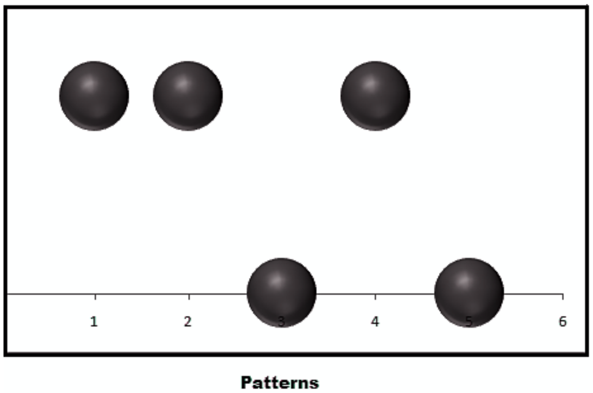

Figure-2. Bubble chart for heuristic pattern prediction.

RESULTS AND DISCUSSIONS

The world of medical zone has seen more advancement in the recent years. The prognoses grow stronger so do the medical cure for it. To add efficacy to the medical treatment an innovative effort has been formulated. A statistical model has been devised and it is trained via the INFC method to predict future cardiac maladies with accuracy and reliability. The Table-2 shows the Comparative Prediction of Actual and Experimental Results. This table takes actuals from expert opinion and experimental value from dredging methodologies. It is clear from Table-2 that the efficacy of INFC depends on the Prism Classifier as that acts as a centroid to the INFC. The INFC is compared with actuals to validate the accuracy for five randomly chosen patterns. The Figure-2 depicts the Bubble Chart for Heuristic Pattern Prediction. It shows that the patterns 3 and 5 are ground valued with bubbles signifying No Heart Disease and patterns 1, 2 and 4 are peak valued with bubbles signifying Heart Disease. The proposed method proves to be efficient and effective by offering optimal reliability and accuracy.

CONCLUSION

The study on medical sector has always been at its peak. Diseases evolve every now and then and so scientists stay of heels to find remedies for them. To add a supplement to those remedies efforts has been made in this paper by deploying a statistical model. This model cannot be taken as a rudimentary diagnosing agent but in turn be taken as a supplement to diagnosing. The future scope of the proposed model would be to include more maladies prediction and also improve upon the centroid construction by incorporating different rule based classifiers. The drawback of the proposed method is that, its efficacy totally depends on the rule that is generated as a part of centroid construction.

REFERENCES

[1] C.D. Cantwell, C.H.Roney, F.S.Ng, J.H.Siggers, S.J.Sherwin, N.S.Peters. 2015. Techniques for

automated local activation time annotation and conduction velocity estimation in cardiac mapping. Computers in Biology and Medicine.

[2] ChrisD. Cantwell, SergeyYakovlev, RobertM. Kirbyb, NicholasS.Peters, SpencerJ.Sherwin. 2014. High-orderspectral/hpelementdiscretisationforreaction diffusion problems on surfaces: Application to cardiac electrophysiology. Journal of Computational Physics. 257: 813-829.

[3] Emily D. Williams, James Y. Nazroo, Jaspal S. Kooner, Andrew Steptoe. 2010. Subgroup differences in psychosocial factors relating to coronary heart disease in the UK South Asian population. Journal of Psychosomatic Research. 69: 379-387.

[4] Faria Afsana, Zafar Ahmed Latif, M. Maksumul Haq. 2010. Parameters of metabolic syndrome are markers of coronary heart disease - An observational study. International Journal of Diabetes Mellitus. 2: 83-87.

[5] G.N.Balaji, T.S.Subashini, N.Chidambaram 2015. Automatic classification of Cardiac Views in Echocardiogram using Histogram and Statistical Features. Procedia Computer Science. 46: 1569-1576.

[6] Hrydziuszko, Artur Wrona, Joanna Balbus, Krystian Kubica. 2014. Mathematical Two-compartment Model of Human Cholesterol Transport in Application to High Blood Cholesterol Diagnosis and Treatment, Electronic Notes in Theoretical Computer Science. 306: 19-30.

[7] Hung JH, Whitfield TW, Yang TH, Hu Z, Weng Z, DeLisi C. 2010. Identification of functional modules that correlate with phenotypic difference: the influence of network topology. Genome Biol. 11:R23.

[8] Jennifer E. Ho, PhilimonGona, Michael J. Pencina5, Jack V. Tu, Peter C. Austin, Ramachandran S. Vasan,William B. Kannel, Ralph B. D'Agostino, Douglas S. Lee and Daniel Levy. Discriminating clinical features of heart failurewith preserved vs. reduced ejection fraction.

[9] Jiahe Xi, Pablo Lamata, Steven Niederer, Sander Land, Wenzhe Shi, Xiahai Zhuang, Sebastien Ourselin, Simon G. Duckett, Anoop K. Shetty, C. Aldo Rinaldi, Daniel Rueckert, Reza Razavi, Nic P. Smith, The estimation of patient-specific cardiac diastolic functions fr.

[10] Malay Mitra and R. K. Samanta. 2013. Cardiac Arrhythmia Classification Using Neural Networks with Selected Features. Procedia Technology. 10: 76-84.

[11] Nienke J. Wijnstok, Jos W.R. Twisk, Ian S. Young, Jayne V. Woodside, Cheryl McFarlaned, Jane McEnenyd, Trynke Hoekstra, Liam Murray and Colin A.G Boreham. Inflammation Markers are Associated with Cardiovascular Diseases Risk in Adolescents: The Young Hearts Project.

[12] Nitin Aji Bhaskar. 2015. Performance Analysis of Support Vector Machine and Neural Networks in Detection of Myocardial Infarction. Procedia Computer Science. 46: 20-30.

[13] Paolo Barbini, Emanuela Barbini, Simone Furini and Gabriele Cevenini. 2014. A straightforward approach to designing a scoring system for predicting length-of-stay of cardiac surgery patients, Barbini *et al*. BMC Medical Informatics and Decision Making. 14:89.

[14] S. Padmavathi, E. Ramanujam. 2015. Naïve Bayes Classifier for ECG abnormalities using Multivariate Maximal Time Series Motif. Procedia Computer Science. 47: 222-228.

[15] S. Udhaya kumar, H. Hannah Inbarani. 2015. A Novel Neighborhood Rough set Based Classification Approach for Medical Diagnosis. Procedia Computer Science. 47: 351-359.

[16] Shivnarayan Patidar, Ram Bilas Pachori. 2013. Constrained Tunable-Q wavelet Transform based Analysis of Cardiac Sound Signals. AASRI Procedia. 4: 57-63.

[17] Sina Khanmohammadi, Hassan Sadeghpour Khameneh, Harold W. Lewis III, Chun-An Chou. 2013. Prediction of Mortality and Survival of Patients after Cardiac Surgery Using Fuzzy EuroSCORE System and Reliability Analysis. Procedia Computer Science. 20: 368-373.

[18] Sindoori, R., Ravichandran, K.S., and Santhi, B.2013. AdaBoost technique for vehicle detection in aerial

surveillance, International Journal of Engineering and Technology, 5(3): 765-769

[19] Yingying Zheng, Yongliang Zhang, Zuchang Ma, Yining Sun. 2010. Predicting Arterial Stiffness from radial Pulse Waveform using support vector machines. Procedia Engineering. 7: 458-462.

[20] Yossi Tsadok, Yael Petrank, Sebastian Sarvari. 2013. Thor Edvardsen, Dan Adam, Automatic segmentation of cardiac MRI cines validated for long axis views. Computerized Medical Imaging and Graphics. 37: 500-511.

[21] Yu Xu, Hong Sun, Su Zhang. 2013. The Statistical Analysis of Patients' Clinical Data in Emergency Department by Using Hospital Information System. AASRI Procedia. 4: 334-339.