A DATA MINING MODEL TO PREDICT THE RISK OF HEART DISEASE USING MULTINOMIAL LOGISTIC REGRESSION (MLR)

Ms.M.K.Dharani,
Assistant Professor ,
Department of Computer science,
kongu engineering college
perundurai-638052
dharani.cse@kongu.ac.in

C.Poovitha ,
Department of Computer science,
Kongu Engineering College,
Perundurai-638052
Mail:cpoovitha1996@gmail.com

Abstract: **In this modern life style people are very much affected by various health issues. According to the survey of Indian medical council most of the people in India are affected by the heart disease. This is mainly due to their work nature and nature of food habits which leads to different level of pulse rate, cholesterol level, and stress rate. Even though it can't be completely eradicated, it can be predicted and treated using the clinical data. It can be effectively analyzed and predicted using the data mining algorithms and techniques. In our proposed work the clinical data set of 14 features and 303 instances are taken. The classification of the dataset is done by correlation based feature subset (CFS) selection with particle swam optimization (PSO) to segregate the attributes that are the necessary for the heart disease. Then the data are clustered using the data mining algorithm and multinomial logistic regression is being applied to clinical data .This technical approach provides the higher accuracy of heart disease prediction as compared with other techniques.**

KEYWORDS: data mining, classification, particle swarm optimization, MLR.

I. INTRODUCTION

Cardiovascular diseases (CVD) are caused by disorders of the heart and blood vessels and result in coronary heart disease, heart failure, cardiac arrest, ventricular arrhythmias and sudden cardiac death, ischemic stroke, transient ischemic attack, subarachnoid and intracerebral hemorrhage, rheumatic heart disease, abdominal aortic aneurysm, peripheral artery disease and congenital heart disease. According to World Health Organization (WHO), 17.5 million people died from CVD in 2012 amounting to 31 % of all global deaths . CAD is a type of CVD in which presence of atherosclerotic plaques in coronary arteries, leads to myocardial infarction or sudden cardiac death . In order to diagnose positive sign of heart disease and to assess the level of damage of heart muscles, certain tests may be prescribed by a medical practitioner including nuclear scan, angiography, echocardiogram, Electrocardiogram (ECG), exercise stress testing. ECG is a noninvasive technique used to identify CAD cases , though it could lead to undiagnosed symptoms of CAD. This limitation leads to angiography which is an invasive diagnosis to confirm CAD cases and is considered as the gold standard for disease detection and severity analysis. However, it is costly and requires high level of technical expertise . Researchers are, therefore, seeking less expensive and effective alternatives, say, using data mining for predicting CAD cases. During the past few decades, image processing, signal processing, statistical and machine learning techniques have been increasingly applied to assist medical diagnosis using ECG and echocardiogram.

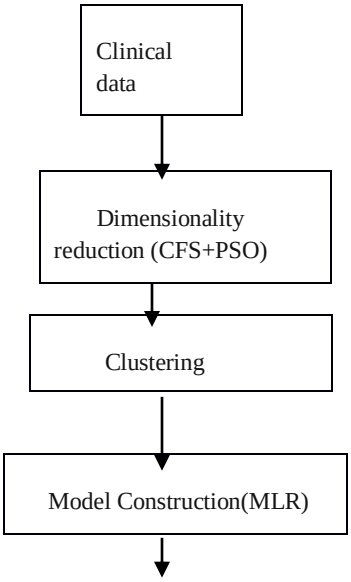
ECG and echocardiogram are specialized processes conducted by trained practioners. Sometimes ECG is not able to confirm CAD cases. This process is complex, costly, involves lot of time and effort. To overcome these limitations many researchers used

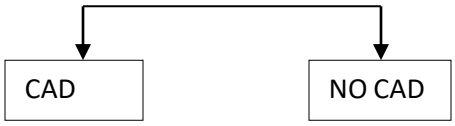
other risk factors excluding angiography to predict CAD cases. These methods are noninvasive, less complex, low cost, reproducible and objective diagnoses, can do automated detection of disease and can be used for screening large number of patients based on clinical data easily obtained at hospitals. we prose a method consisting of clinical data collection, dimensionality reduction with correlation based feature subset selection with PSO, followed by data clustering for identification of incorrectly assigned cluster data points. Finally, models were constructed with MLR.

II.EXISTING SYSTEM:

For diagnosis of CAD, angiography is used which is a costly time consuming and highly technical invasive method. This limitation leads to angiography which is an invasive diagnosis to confirm CAD cases and is considered as the gold standard for disease detection and severity analysis. However, it is costly and requires high level of technical expertise . Researchers are, therefore, seeking less expensive and effective alternatives, say, using data mining for predicting CAD cases

III.FLOW CHART FOR THE OUR SYSTEM





IV.DIMENTIONALITY REDUCTION

CFS:

Machine learning provides tools by which large quantities of data can be automatically analyzed. Fundamental to machine learning is feature selection . Feature selection, by identifying the most salient features for learning, focuses a learning algorithm on those aspects of the data most useful for analysis and future prediction. The hypothesis explored in this thesis is that feature selection for supervised classification tasks can be accomplished on the basis of correlation between features, and that such a feature selection process can be beneficial to a variety of common machine learning algorithms. A technique for correlation-based feature selection, based on ideas from test theory, is developed and evaluated using common machine learning algorithms on a variety of natural and artificial problems. The feature selector is simple and fast to execute. It eliminates irrelevant and redundant data and, in many cases, improves the performance of learning algorithms.

PSO:

Particle swarm optimization (PSO) is a global optimization strategy that simulates the social behavior observed in a flock (swarm) of birds searching for food. A simple search strategy in PSO guides the algorithm toward the best solution through constant updating of the cognitive knowledge and social behavior of the particles in the swarm. Feature selection is a process of selecting a subset of relevant features from a large number of original features to achieve similar or better classification performance and improve the computation efficiency. As an important data pre-processing technique, research into feature selection has been carried out over the past four decades. Determining an optimal feature subset is a complicated problem. Due to the limitations of conventional methods, evolutionary

computation (EC) has been proposed to solve feature selection problems. Particle swarm optimization (PSO) is an EC technique which recently has caught much interest from researchers in the field. This paper presents a review of PSO for feature selection in classification. Its purposes include reducing the amount of data needed for learning, shortening the running time, improving the system accuracy, and increasing the comprehensibility of the learned model.

V.ATTRIBUTE INFORMATION:

Only 14 attributes used:

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

A	B	C	D	E	F	G	H	I	J	K	L	M	N
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
44	1	2	120	263	0	0	173	0	0	1	0	7	0
52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
48	1	2	110	229	0	0	168	0	1	3	0	7	1

TABLE 1: DATASET FOR THE OUR PROJECT

VI.MODEL FOR CAD IDENTIFICATION:

MLR achieves highest prediction accuracy of 88.4 %.We tested this approach on benchmarked Cleave land heart disease data as well. In this case also, MLR outperforms other techniques. Proposed hybridized model improves the accuracy of classification algorithms from 8.3 % to 11.4 % for the Cleveland data.

Multinomial logistic regression model (MLR):

It is an extension of logistic regression with ridge estimator.MLR is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, MLR uses Maximum

likelihood estimation to evaluate the probability of categorical membership. Multinomial Logistic Regression is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels. Like all linear regressions, the multinomial regression is a predictive analysis. Multinomial regression is used to describe data and to explain the relationship between one dependent nominal variable and one or more continuous-level(interval or ratio scale) independent variables. Multinomial logistic regression is known by a variety of other names, including polytomous LR,multiclass LR, softmax regression, multinomial logit, maximum entropy (MaxEnt) classifier, conditional maximum entropy model.

Algorithm	Accuracy	Incorrectly classified instances	% improvement in accuracy (all features vs CFS + PSO feature selection
MLR	84.17	15.8	0.67

TABLE 2: CFS + PSO + classification

Algorithm	All the features	CFS + PSO	PSO + clustering	% improvement in accuracy with Accuracy our approach
MLR	85.47	83.16	91.36	8.2

Table 7 Performance of prediction models for Cleaveland heart disease
data set

VII.CONCLUSION:

Our work is to identify and confirm CAD cases at low cost by using clinical data that can be easily collected at hospitals. Complexity of the system is decreased by reducing the dimensionality of the data set with PSO. It provides reproducible and objective diagnosis, and hence can be a valuable adjunct tool in clinical practices. Results are comparably, promising and therefore the proposed method will be helpful in disease diagnostics. Experiment results demonstrate the superiority of the proposed method with regard to prediction accuracy of CAD with the features selected by CFS & PSO, we need only a few clinical data to apply this model. The accuracy can be further increased with more data instances.

References

1. Wong, N.D., Epidemiological studies of CHD and the evolution of preventive cardiology. Nat. Rev. Cardiol. 11(5):276–289, 2014.
2. http://www.who.int/mediacentre/factsheets/fs317/en/ (Accessed on January 2016).
3. Tsipouras, M.G., Exarchos, T.P., Fotiadis, D.I., Kotsia, A.P., Vakalis, K.V., Naka, K.K., and Michalis, L.K., Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Trans. Inf. Technol. Biomed. 12(4):447–458, 2008.
4. http://heartdiseaseonline.com (Accessed on November 2015).
5. Acharya, U.R., Faust, O., Sree, V., Swapna, G.,Martis, R.J., Kadri, N.A., and Suri, J.S., Linear and nonlinear analysis of normal and CAD-affected heart rate signals. Comput. Methods Prog. Biomed. 113(1):55–68, 2014.
6. Giri, D., Acharya, U.R., Martis, R.J., Sree, S.V., Lim, T.C., Ahamed, T., and Suri, J.S., Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. Knowl.-Based Syst. 37:274–282, 2013.
7. http://www.nhlbi.nih.gov/health/health-topics/topics/cad (Accessed on February 2016).
8. Alizadehsani, R., Hosseini, M. J., Sani, Z. A., Ghandeharioun, A., & Boghrati, R., Diagnosis of coronary artery disease using costsensitive algorithms. In Data Mining Workshops (ICDMW), 2012 I.E. 12th International Conference on (pp. 9–16). IEEE, 2012.
9. Arafat, S., Dohrmann, M., & Skubic, M., Classification of coronary artery disease stress ECGs using uncertainty modeling. In Computational Intelligence Methods and Applications, 2005 ICSC Congress on (pp. 4-pp). IEEE, 2005.
10. Lee, H. G., Noh, K. Y., & Ryu, K. H., A data mining approach for coronary heart disease prediction using HRV features and

carotid arterial wall thickness. In BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on (Vol. 1, pp. 200–206). IEEE, 2008.

11. Acharya, U.R., Sree, S.V., Krishnan, M.M.R., Molinari, F., Saba, L., Ho, S.Y.S., and Suri, J.S., Atherosclerotic risk stratification strategy for carotid arteries using texture-based features. Ultrasound Med. Biol. 38(6):899–915, 2012.

12. Acharya, U.R., Mookiah, M.R.K., Sree, S.V., Afonso, D., Sanches, J., Shafique, S., and Suri, J.S., Atherosclerotic plaque tissue characterization in 2D ultrasound longitudinal carotid scans for automated classification: a paradigm for stroke risk assessment. Med. Biol. Eng. Comput. 51(5):513–523, 2013.

13. Zhao, Z., & Ma, C., An intelligent system for noninvasive diagnosis of coronary artery disease with EMD-TEO and BP neural network. In Education Technology and Training, 2008. and 2008 International Workshop on Geoscience and Remote Sensing. ETT and GRS 2008. International Workshop on (Vol. 2, pp. 631–635). IEEE, 2008.

14. Acharya, U.R., Sree, S.V., Krishnan, M.M.R., Krishnananda, N., Ranjan, S., Umesh, P., and Suri, J.S., Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images. Comput. Methods Prog.
Biomed. 112(3):624–632, 2013.

15. Kim, W. S., Jin, S. H., Park, Y. K., & Choi, H. M., A study on development of multi-parametric measure of heart rate variability diagnosing cardiovascular disease. In World Congress on Medical Physics and Biomedical Engineering 2006 (pp. 3480–3483). Springer: Berlin Heidelberg, 2007.

16. A.Suresh and K.L. Shunmuganathan (2012), "Important Business Factor Analysis using Data Mining Approach", International Journal of Engineering Science and Technology, ISSN: 0975-5462, Vol. 4 No.02 February 2012, pp. 606-610. March 2011.