ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/105707, December 2016* ISSN (Online) : 0974-5645

A Data Mining Model for Coronary Artery Disease Detection using Noninvasive Clinical Parameters

Luxmi Verma[1] and Sangeet Srivastava[2*]

[1]The NorthCap University, Gurugram - 122017, Haryana, India; luxmi.verma@gmail.com
[2]Department of Applied Sciences,The NorthCap University, Gurugram - 122017, Haryana, India; sangeetsrivastava@ncuindia.edu

Abstract

Coronary Artery Disease (CAD) is one of the major cause of death as well as disability worldwide. Suspected cases of CAD go through invasive and non-invasive tests to get CAD detected. Angiography is a noninvasive method for detection. Not only it is costly, time consuming and risky, but it also needs technical expertise and not suitable for the screening of large population. Hence researchers are looking for better alternatives using non-invasive clinical tests. **Objectives:** To construct Artificial Neural Network based model for CAD identification and adjudging its accuracy with respect to other models. **Method:** Data mining techniques are being employed to identify CAD cases based on non-invasive clinical tests. Early detection of disease is necessary in order to avoid the risk being exaggerated further. Benchmark Cleave land heart disease data is collected from UCI machine repository and Probabilistic Neural Network is employed and trained and tested using non-invasive clinical parameters. **Finding:** Neural Network based model is presented that uses non-invasive clinical parameters of the subjects to model CAD cases and achieves the diagnosis accuracy of 96% and misclassification error rate of 4%.The models' performance is also compared with other classifiers such as RBF Network, AD Tree. **Improvement:** Neural network based model showed the superiority over other methods in terms of accuracy. Results are promising and reproducible and therefore the model can be valuable adjunct tool in clinical practices.

Keywords: Coronary Artery Disease, Data Mining, Decision Tree, Probabilistic Neural Network

1. Introduction

Data mining is defined as extraction of valuable and interesting patterns from enormous data[1–6]. Data mining is used in various fields namely crime data analysis[7], steganography[8], education[9], weather forecasting[10], traffic management[11], product quality management[12], retail business[13]and health sectors[14–18].During the past few decades' researchers have applied data mining methods into health sectors for making clinical decision for prognosis, health care management, treatment planning, prediction of the effectiveness of surgical procedures and identification of various diseases such as cancer, diabetes, cardiovascular diseases etc. CAD is a category of heart disease. It is one of the major reason for death as well as disability worldwide as per WHO[19]. CAD is chronic disease in which accumulation of plaque in coronary arteries gradually hardens and narrowing of coronary artery can lead to heart attack and death. The diagnose of CAD is a complex clinical procedure in which number of factors needs to be considered such as evaluation of risk factors, results of laboratory test and physical examination of the patients. Moreover, diagnosis consumes enormous amount of time, cost, equipment's' and requires highly skilled physicians having experience in the field. The cost of care and follow-up of patients is very high making it imperative to identify CAD cases with high order of accuracy. The early diagnosis and prediction helps in reducing the mortality rate and morbidity rate of the disease. There are number of non-invasive methods

* *Author for correspondence*

available for screening CAD such as Electrocardiogram, Echo cardiogram,Magnetic resonance imaging, Computer aided tomography. Most of the noninvasive methods are costly, not widely available and moreover the result of these methods are not as accurate as angiography[20–26]. Coronary Angiography is a noninvasive method for CAD detection. It is one of the gold standard to diagnose heart diseases which needs an extensive technical knowledge and expertise. Due to these limitations researcher is exploiting intelligent computation techniques for disease diagnosis to aid the clinician in the process of decision making. Predictive mining is intelligent computation technique used to construct model by using patient's clinical features to find potentially valuable patterns. Predictive mining methods ranges from simple to complex and powerful ones like linear regression and ANN respectively. Our main objective is to apply ANN for CAD identification and adjudge its accuracy with respect to other predictive mining techniques.

2. Material and Methods

We consider a data set for heart disease which is collected from UCI Machine repository contributed by Cleveland Clinical Foundation. Data consists of 14 features namely Age, Gender, type of Chest pain, Trestbps, Chol, thalach, Restecg, Fbs, Exang, Old peak, Slop, Ca, Thal and numas the result of coronary angiography. Total 303 subjects are considered in the data set with the possibility of having CAD. Table 1. illustrates the detailed description of the above subjects.

2.1 Model Construction

Data set is preprocessed by removing the instances with missing values. We used 70% of the data for training of the model and tested it with remaining 30%.

Artificial Neural Network (ANN)model is a mathematical representation which is enthused by human brains. It consists of highly interconnected neurons organized into layers such as input, processing and output layers[17].To model complex relationships among various parameters i.e. both input and output,ANN architecture can be used. Such models are trained for exploring relationship between data using training dataset. ANN is tested using testing data. Figure 1shows the model we developed using simple ANN for the identification of CAD.

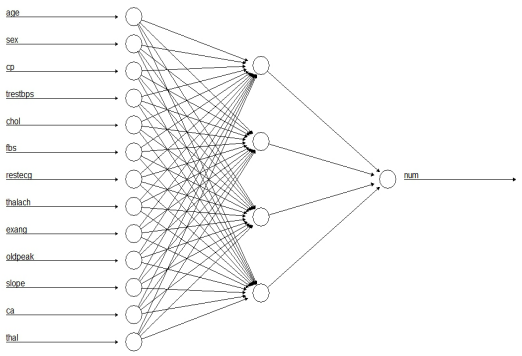

Figure 1. Neural network model for coroanry artery disease.

2.1.1 Probabilistic Neural Network (PNN)

It is a type of feed forward neural network implementation of statistical method kernel discriminant analysis and Bayesian decision rule. It is used to solve data classification problem of automatic learning. It consists of multilayers namely input, pattern, summation and output layer respectively. It was proposed by Specht[27]. As per Bayesian decision rule: xis a data pattern and c is the class labels $c=_{1,\ldots n}$. The probability of x ∈ c is P_n, and the classifying cost of x to class c is $D_{n.}$. The Probability density function for the target class is (PDF)$y_1(x),y_2(x)\ldots.y_n(x)$. Then according to Bayes theorem, when c≠h, the pattern x is label ledas class c, if $P_cD_ny_n(x)>P_hD_hy_h(x)$. Usually $P_c = P_h$ and $D_n = D_h$ thus if $y_n(x)>y_h(x)$ pattern x is labeled as class c[28].Figure 2 show the architecture of probabilistic neural network.

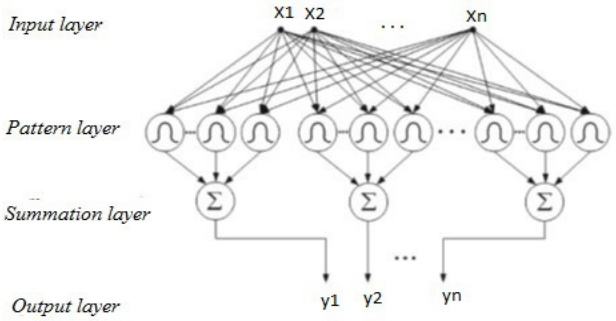

Figure 2. Architecture of probabilistic neural network[28].

2.1.2 ADTree

Alternating Decision tree is a supervised learning tree based method. It consists of decision nodes which specify

Table 1. Description of the data set

Features	Description	Mean ± StDev	CAD (n=83)	Without CAD (n=214)
Age	Age	55.43±9.03	57.506±7.862	53.393±9.234
Gender	0-female, 1-male	0.68±0.46	0.807±0.397	0.626±0.485
Chol	Serum cholesterol mg/d	246.6±51.77	253.614 ±54.698	244.921 ±50.836
Cpt	Chest pain type 1- Typical Angina 2- Atypical angina 3- Non-angina pain 4- Asymptomatic	3.15 ±0.96	3.747±0.641	2.93±0.974
Frbs	fasting blood sugar 0 - no 1- yes	0.149±0.356	0.205±0.406	0.121 ±0.327
Trestbp	Resting blood pressure at admission	131.69±17.6	135.518±19.397	130.21±16.903
Thalch	Maximum heart rate achieved	149.60±22.8	134.639±21.462	155.402±20.81
Rstecg	Resting electrocardiographic outcome 0-normal 1- having ST-T wave abnormality 2- showing probable or definite left ventricular hypertrophy	0.99±0.995	1.193±0.969	0.921 ±0.997
Exang	exercise induced angina 0 = no 1 = yes	0.327±0.47	0.602 ±0.492	0.22±0.415
Old_peak	ST depression induced by exercise related to rest	1.04±1.161	1.958 ±1.355	0.706±0.862
Slope	Slope of the peak exercise ST Segment 1-upsloping 2- flat 3: down sloping	1.601±0.616	1.958 ±0.539	1.467 ±0.594
Ca	Number of fluoroscopy colored vessels	0.672±0.937	1.41±1.025	0.393 ±0.728
Thall	3 – normal 6 -fixed defect 7 -reversible defect	4.73±1.94	6.169± 1.529	.79

a condition and prediction nodes. ADTree classifies an instance by analysing all the paths where decision nodes are true, and adding those nodes that are traversed[29].

2.1.3 RBF Network

An RBFN performs classification by measuring the input match to the training set examples. A prototype is stored in each neuron, which can be considered as an example in the training data. To classify a new input, each neuron measures the Euclidean distance among input and stored prototype. The items that have close values of the predictor variables are likely to be similar with the predicted target value of an item[30].

3. Performance Measures

The performance of a diagnostic model is assessed by measuring the discrepancy between the actual and predicted outcome. The model is considered good if this discrepancy is low. Accuracy and error rates are the popular metrics for measuring the performance of the diagnostic model.

Accuracy (Acc) -the percentage of correctly classified subjects by the model $Acc = \frac{TP + TN}{TP + \mathbf{TN} + FP + FN}$

Misclassification Error Rate – the percentage of incorrectly classified subjects by the model. MER =

$$\frac{FP + FN}{TP + FP + TN + FN}$$

Where True positive (TP) denotes the cases having CAD and our model detected as CAD effected patients. True negative (TN) denotes the cases who do not have CAD and our model detected as no CAD cases. False Positive (FP) are the cases who actually do not have CAD but the model detected as CAD patients and False Negative (FN) are those cases who actually are positive for CAD and our model detected as no CAD.

4. Results and Discussion

A total of 297 patients (96 females and 201 males) were included in the Cleaveland heart disease data set in which 80 subjects were female and 134 were males without disease. However, 16 females and 67 males had CAD detected by angiography. The mean age of studied population is 55 years ranging from 20 to 77 years. Prevalence of this disease in females is 16%and prevalence of disease among males is 66 %. Prevalence ratio, considering gender as a marker for CAD is 4.12%. The prevalence of coronary heart disease in males is 4.12 times larger than that in females. The odds of having CAD are 10 times higher among males than the odds among females. Instances with missing values were removed before model generation. Neural network and decision tree methods are used to construct prediction models by using clinical and laboratory parameters to predict CAD. Contingency matrix is generated to assess the performance of diagnostic models. Table 2 shows the contingency matrix generated using the probabilistic neural network model.

Table 2. Contingency matrix for dichotomous outcome

Predicted	Observed	
	CAD	No CAD
CAD	26	1
No CAD	1	70

All the features of the patient's population showed statistically significant variation between CAD and non CAD patients (Table 1). Mean value of serum cholesterol, resting blood pressure on admission and ST depression induced by exercise related to rest is higher in CAD patients as compared to healthy subjects. Tables 3 and 4indicates the performance measure of the diagnostic models where PNN model achieves the maximum prediction accuracy of 96 % and lowest misclassification rate of 4% respectively. RBF network achieves the prediction accuracy of 87.9% and misclassification rate of 12 % respectively which is higher as compared to Decision tree, i.e., 84.6 % and 15.3 %.

Table 3. Performance of PNN, RBF Network and ADTree

Model	Accuracy (%)	Error Rate (%)
PNN	96.0	4.0
RBF Network	87.9	12.0
ADTree	84.6	15.3

Table 4. Performance comparison with other research work

Reference	Accuracy (%)	Method	Authors
[31]	85.76	SVM	Bouali H, Akaichi J
[32]	81.41	Bagging	Tu MC et al.
[32]	78.91	C4.5	Tu MC et al.
[33]	82.22	MLP	Peter TJ, Somasundaram K.
[34]	86.8	MLP + Fuzzy	Kahramanli H, Allahverdi N
[35]	86 .0	Naïve Bayes	Palaniappan S, Awang R
[36]	92 .0	Naive Bayes	El Bialy R et al.
Our work	96		

Bouali et al.[31]applied supervised machine learning techniqeus namely Support Vector Machine, Baysian Network, ANN, Decision tree and Fuzzy pattern tree. SVM achieves the highest predction accuracy of 85.76% as compared to other classifiers and decision tree achieves the lowest prediction accruacy on Cleavland heart disease data set using ten fold cross validation. Tu Mc et al.[32] applied bagging and decision tree to construct diagnostic model and achieves the prediction accuray of 81.41 % for Baggaing and 78.91% for decision tree. Palniappan S et al[33] proposed intelligent prediction system for heart disease by applying data mining methods. Naive Bayes based technique achieves the highest prediction accuracy of 86%, Neural network 85.6 % and Decision tree 80 %.The performance of our model is also superior than work presented in[34-36].

5. Conclusion

Neural network based model achieves the higest prediction accuracy and lowest miss classification error rate as compared to other diagnostic models using benchmark Cleaveland Heart disease data. Patients clinical parameters can be easily collected from hospitals. Results are promising and reproducable and therefore the model can be considered as a significant tool in clinical practices.

6. References

1. Han J, Pei J, Kamber M. Data mining: concepts and techniques.3rd edn.Elsevier; 2011.
2. Gupta M, Dahiya D. Performance evaluation of classification algorithms on different data sets. Indian Journal of Science and Technology. 2016 Oct; 9(40):1–6.
3. Meenakshi M, Geetika G. Survey on classification methods using WEKA. International Journal of Computer Applications. 2014 Jan; 86(18):16–19.
4. Verma A, Gill A, Kaur I. Analysis and implementation of data mining algorithms for deploying ID3, CHAID and Naive Bayes for random dataset. Indian Journal of Science and Technology. 2016Oct; 9(40):1–32.
5. Verma TR, DeeptiG. Implementation of clustering algorithms in rapidminer. IFRSA International Journal of Data Warehousing and Mining. 2014Feb; 4(1):59–61.
6. Verma A, Kaur I, Singh I. Comparative Analysis of Data Mining Tools and Techniques for Information Retrieval. Indian Journal of Science and Technology. 2016 Mar; 9(11).
7. Aggarwal N, Gaur D. Classification of crime data using rapid miner. International Journal of Applied Engineering Research. 2015; 10(35).
8. Chhikara RR, Sharma P, Singh L. A hybrid feature selection approach based on improved PSO and filter approaches for image steganalysis. International Journal of Machine Learning and Cybernetics. 2016 Dec; 7(6):1195–206.
9. Zhonglin T, Xueping N. Application of data mining in university research management system. Proceedings of 4th IEEE international conference on Computational and Information Sciences (ICCIS), China; 2012 Aug 17. p. 761–3.
10. Ghosh S, Nag A, Biswas D, Singh JP, Biswas S, Sarkar D, Sarkar PP. Weather data mining using artificial neural network. Proceedings of IEEE Conference on Recent Advances in Intelligent Computational Systems (RAICS), India; 2011 Sep. p. 192–5.
11. Zamani Z, Pourmand M, Saraee MH. Application of data mining in traffic management: case of city of Isfahan. Proceedings of IEE international conference on Electronic Computer Technology (ICECT); 2010. p. 102–6
12. Da Cunha C, Agard B, Kusiak A. Data mining for improvement of product quality. International Journal of Production Research. 2006 Sep 15;44(18–19):4027–41.
13. Ahmed SR. Applications of data mining in retail business. Proceedings of IEEE International Conference on Information Technology: Coding and Computing (ITCC), 2004 Apr 5–7, Nevada. 2004;2:455–9.
14. Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA. Association rules and data mining in hospital infection control and public health surveillance. Journal of the American Medical Informatics Association. 1998 Jul 1;5(4):373–81.
15. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. Nature Reviews Genetics. 2012 Jun 1;13(6):395–405.
16. Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences. 2013 Jul 31;25(2):127–36.
17. Verma L, Srivastava S, Negi PC. A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. Journal of Medical Systems. 2016 Jul 1;40(7):1–7.
18. Kumar V, Verma L. Binary classifiers for health care databases: A comparative study of data mining classification algorithms in the diagnosis of breast cancer. International Journal of Computer Science and Technology. 2010 Dec; 1(2):124–9.
19. Available from:http://www.who.int/mediacentre/factsheets/fs317/en/ Date accessed: 01/01/2016
20. Acharya UR, Sree SV, Krishnan MM, Krishnananda N, Ranjan S, Umesh P, Suri JS. Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images. Computer Methods and Programs in Biomedicine. 2013 Dec 31; 112(3):624–32.
21. Escolar E, Weigold G, Fuisz A, Weissman NJ. New imaging techniques for diagnosing coronary artery disease. Canadian Medical Association Journal. 2006 Feb; 174(4):487–95.
22. Giri D, Acharya UR, Martis RJ, Sree SV, Lim TC, Ahamed T, Suri JS. Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. Knowledge-based Systems. 2013 Jan 31;37:274–82.
23. Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, Bahadorian B, Sani ZA. A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine. 2013 Jul 31;111(1):52–61.
24. Kahramanli H, Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. Expert Systems with Applications. 2008 Aug 31;35(1):82–9.
25. Polat K, Şahan S, Güneş S. Automatic detection of heart disease using an Artificial Immune Recognition System (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. Expert Systems with Applications. 2007 Feb 28;32(2):625–31.
26. Das R, Turkoglu I, Sengur A. Diagnosis of valvular heart

disease through neural networks ensembles. Computer Methods and Programs in Biomedicine. 2009 Feb 28;93(2):185–91.
27. Specht DF. Probabilistic neural networks. Neural Networks. 1990 Jan 1;3(1):109–18.
28. Kusy M, Zajdel R. Probabilistic neural network training procedure based on Q (0)-learning algorithm in medical data classification. Applied Intelligence. 2014 Oct 1;41(3):837–54.
29. Freund Y, Mason L. The alternating decision tree learning algorithm. Inicml. 1999 Jun 27; 99;124–33.
30. Fu X, Wang L. Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2003 Jun;33(3):399–409.
31. Bouali H, Akaichi J. Comparative study of different classification techniques: Heart disease use case. Proceedings of 13th IEEE international conference on Machine Learning and Applications (ICMLA), Detroit; 2014 Dec 3. p. 482–6.
32. Tu MC, Shin D, Shin D. Effective diagnosis of heart disease through bagging approach. Proceedings of IEE2nd International Conference on Biomedical Engineering and Informatics, 2009 Oct 17, China; 2009. p. 1–4.
33. Peter TJ, Somasundaram K. An empirical study on prediction of heart disease using classification data mining techniques. Proceedings of IEEE international conference on Advances in Engineering, Science and Management (ICAESM), India; 2012 Mar 30. p. 514–18.
34. Kahramanli H, Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. ExpertSystems with Applications. 2008 Aug 31;35(1):82–9.
35. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. Proceedings of 12th IEEE/ACS International Conference on Computer Systems and Applications, Egypt; 2008 Mar. p. 108–15.
36. El Bialy R, Salama MA, Karam O. An ensemble model for heart disease data sets: a generalized model. Proceedings of the ACM 10th International Conference on Informatics and Systems, Egypt;2016 May. p. 191–6.