Effective Coronary Artery Disease Diagnosis using Hybrid Data Mining Model

Dr.S.Karthik[1] Dharini.S[2] Chitra.A[3] Eshwaramoorthy.M[4]
[1]Professor & Dean [2,3,4]Student
[1,2,3,4]Department of Computer Science & Engineering
[1,2,3,4]SNS College of Technology, Coimbatore

Abstract— *Coronary artery disease (CAD) is one of the most critical diseases which is statistically growing today. Angiography is used to examine CAD in order to see how blood flows through arteries after introduction of a radiopaque substance. Angiography is an expensive invasive technique and needs trained practioners. Noninvasive techniques like ECG and echocardiogram sometimes left with undiagnosed symptoms. Many alternative methods such as machine learning algorithms that could use non-invasive clinical data for the disease diagnosis are therefore suggested and are under research. In this study, we present a hybrid method consisting of clinical data collection, dimensionality reduction with correlation based feature subset selection (CFS) with particle swarm optimization (PSO) followed by K-means clustering. Model construction is done using supervised learning algorithms such as multi-layer perceptron (MLP), multinomial logistic regression (MLR), fuzzy unordered rule induction algorithm (FURIA) and C4.5. The experiments are conducted using Waikato Environment for Knowledge Analysis (WEKA) toolkit. The approach tested on Cleaveland heart disease data from University of California, Irvine (UCI) machine learning repository proves to be a promising tool for CAD diagnosis with improved prediction accuracy.*

Keywords— Coronary Artery Disease, Particle Swarm Optimization, Correlation Based Feature Selection, Supervised Learning Algorithm

I. INTRODUCTION

CAD is a type of Cardiovascular disease (CVD) caused by disorders of heart and blood vessels. Coronary artery disease is caused due to deposition of fat in arteries which results in cardiac arrest and heart attack. According to World Health Organisation (WHO), there is a considerable increase in the mortality rate due to CAD. To diagnose heart disease, there are methods like echocardiography and myocardial cintigraphy which prove to be expensive and not always usable. Exercise testing is commonly used but is not very specific and has low sensitivity. These methods may also not be able to detect disease in early stages. Currently, the most reliable method for detection of heart disease is angiography which may bear unwanted risks for patients being an invasive procedure. It also involves lot of time and technical expertise. Nowadays techniques like image processing, signal processing, statistical and machine learning techniques are increasingly used to assist medical diagnosis to minimize time and technical expertise. This combination of intelligent algorithms and machine learning techniques help physicians to deal with large volumes of data much easier. The use of data mining techniques thereby significantly improves the quality of medical diagnosis.

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation. Medical data mining is considered as gold standard due to its application in healthcare domain. It also effectively identifies hidden patterns within medical datasets.

Datasets are usually associated with high dimensional attributes. The datasets from clinical databases are prone to systematic and human errors. There is always a need for new type of data analysis and medical diagnosis tools due to noisy and high dimensional nature of data which greatly affects classification accuracy. A detailed study of different techniques is to be considered for accurate and efficient implementation of a new technique. Therefore, a new hybrid model is presented for medical predictions that use feature subset selection with optimization techniques to improve prediction accuracy. To improve the accuracy of classification, the technique is further combined with clustering and classification techniques. The proposed approach also reduces the computational complexity by effective dimensionality reduction techniques.

II. LITERATURE SURVEY

To overcome the limitations of expensive medical diagnosis, many researchers prompted for noninvasive, less complex, low cost, reproducible and objective diagnoses excluding angiography to predict CAD cases. These techniques can be used for screening large number of patients based on clinical data obtained at hospitals and can do automated detection of disease.

Cardiometabolic risk (CMR) estimation can significantly contribute to the early prevention of atherosclerosis and cardiovascular diseases. An intelligent software system can save time and money by conducting tests only on people with higher CMR. A MATLAB solution is presented based on ensemble of well-learned artificial neural networks and evolutionary algorithm. The system has few limitations – data cannot be collected at the same time and the same place, dataset can be very large and data of some other region are not applicable since every population has own features[2].

Feature selection plays a major role in the accuracy of classification and analysis of data. Feature subset selection is viewed as a problem of combination and optimization. A hybrid method is presented or optimal feature subset selection using endocrine based PSO and Artificial Bee Colony algorithm(ABC) guided by evolutionary algorithm[3].

The presence of noise and high dimensional features decrease classification accuracy. To reduce features from a massive medical data, an algorithm called swarm intelligence based ABC for feature selection is used and the

results performed better than reverse ranking and forward ranking method of feature subset selection[4].

One of the challenging problems is the accurate prediction of disease in the presence of large number of missing values. A novel hybrid prediction mode with missing value imputation (HPM-MI) is presented that uses simple k-means clustering with MLP. The data quality is significantly improved by the use of best imputation technique after quantitative analysis[5].

There is a wealth of hidden information present in the data generated by medical practioners. Determining patterns from hidden information can be important for making predictions. A study is presented on prediction of heart disease with classification algorithms such as J48, Naïve Bayes, REPTREE, CART and Bayes Net[6].

Artificial Neural Network (ANN) is used to establish complex relations between input and the output. The optimization of neural networks is done by Genetic algorithms with improved prediction accuracy[7].

Evolutionary algorithms are one in which random search process is guided by the goal of improving search results from one generation to the next, ultimatel leading to the optimal solution. A fuzzy expert system based on Imperialist Competitive Algorithm (ICA) and Improved ICA is presented with high convergence speed[8].

Feature selection based on Logistic Regression (LR) and Artificial Neural Network (ANN) using Cross Validation Sample (CVS) and Percentage Split as test options. More efficient results are obtained with reduced attributes[9].

Prediction of CVD risk factors based on heart rate variability (HRV) using techniques like SVM that represents a decision boundary using a subset of training examples, random forest and ANN. As compared to echographic parameters data mining based classifier show higher prediction accuracy[10].

It is essential to assess risk factors contributing more to CAD. A decision tree algorithm (C4.5) is employed to analyse clinical data using five splitting criteria based on gain ratio, chi-square statistics, gini index, likelihood ratio and information gain[11].

III. METHODOLOGY

The method involves clinical data collection followed by dimensionality reduction using CFS and PSO, clustering using K-means Clustering algorithm and model construction using MLP, MLG, FURIA, DT(C4.5). Results show improvement in prediction accuracy with effective classification techniques, noise and feature reduction in the dataset.

Clinical data is collected from UCI repository. A data set refers to collection of data in a database. Several characteristics define a data set's structure and properties. The UCI Machine Learning Repository is a collection of databases and theories that are used for analysis of machine learning algorithms. There are four data sets available for heart disease prediction – Cleveland, Switzerland, Hungarian, Long beach. Cleveland database is most commonly used by researcher till date.

The approach uses WEKA toolkit for conducting experiments. Weka is a collection of machine learning algorithms and contains tools for data pre-processing, classification, regression, clustering, association rules and visualization well suited for developing new machine learning schemes.

Clinical data from Cleveland database is used for testing. Cleveland database consists of 76 attributes and 303 instances. Highly ranked attributes are selected using appropriate attribute evaluator and search method. This method uses CFS Subset Evaluation and PSO search method for effective dimensionality reduction.

A. Correlation-based Feature Selection

It states that features are relevant if it is correlated with or predictive of the class, otherwise it is irrelevant. It involves correlating nominal features, feature-class correlation, feature-feature correlation may then be calculated. If the instance is same with respect to the value of second attribute, it is said to be a pure function and if the instances differ, it is said to be impure function. The attribute selector evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

B. Particle Swarm Optimization

PSO is a population based stochastic optimization technique developed by Dr. Eberhart and DR. Kennady. It iteratively tries to improve the candidate solution given the quality measure. An example scenario – A group of birds randomly search for food in an area and there is only one piece of food. The birds do not know where the food is but they know how far the food is in each iteration. The best strategy is to follow the bird which is nearest to the food. The algorithm thus keeps track of three global variables:

1) Target Value
2) Gobal best(gBest)
3) Stopping value

Each particle consists of

1) Possible solution
2) Velocity
3) Personal best(pBest)

C. K-Means Clustering

It is a clustering technique in which items are moved among sets of clusters until the desired cluster is reached. It uses Euclidean distance measure to compute distances between instances and clusters.

IV. MODEL CONSTRUCTION

A. Multi-Layer Perceptron (MLP)

ANN was inspired by attempts to simulate biological neural systems. Anologous to human brain structure, an ANN is composed of an interconnected assembly of nodes and directed links. MLP is a popular ANN model that maps sets of input data to appropriate outputs. It consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one.

B. Multinomial Logistic Regression Model (MLR)

It is used to model nominal outcome variables and allows for more than two categories. To evaluate probability of categorical membership MLR uses maximum likelihood estimation.

C. Fuzzy Unordered Rule Induction Algorithm (FURIA)

It makes use of efficient rule stretching method. The rules are obtained through replacing intervals with fuzzy intervals using trapezoidal membership function combined with sophisticated rule induction techniques.

D. C4.5

It constructs a decision tree using divide and conquer technique and works with top-down approach. The splitting criteria are based on gain ratio.

E. Evaluation Parameters

The models are evaluated using ten-fold cross validation method and the performance is calculated using confusion matrix. Accuracy and incorrectly classified instances are also calculated and compared with evaluation parameters such as true positive (tp), true negative(tn), false positive(fp) and false negative(fn) rate.

F. Limitations

The results show significant improvement in noise and feature reduction but when data instances are more accuracy is decreased.

V. CONCLUSION

The hybrid approach shows that by reducing the dimensionality of the data set with PSO complexity of the system is decreased. The results of clustering gave a desired set in arff file format by traning samples. The result window showed the centroid of each cluster which can be viewed through visualize cluster assignments option. MLP ranked top in model construction tested on various attributes. Experiment results show the superiority of the hybrid method with regard to prediction accuracy of CAD, we need only a few clinical data to apply this model. The accuracy can further be improved by effective classification and dimensionality reduction techniques when data instances are more.

REFERENCES

[1] Luxmi Verma,Sangeet Srivatsava,P.C Negi,'A hybrid data mining model to detect coronary artery disease cases using non-invasive clinical data', J Med Syst (2016) 40:178 DOI 10.1007/s10916-016-0536-z

[2] Kupusinac, A., Stokic, E., and Kovacevic, I., Hybrid EANN-EA system for the primary estimation of Cardiometabolic risk. J. Med. Syst. 40(6):1–9, 2016

[3] Lin,K.C., and Hsieh,Y.H., Classification of medical data sets using SVMs with hybrid evolutionary algorithms based on endocrine based particle swarm optimization and artificial bee Colony algorithms. J. Med. Syst. 39(10):1–9, 2015.

[4] Subanya, B., & Rajalaxmi, R. R., Feature selection using Artificial Bee Colony for cardiovascular disease classification. In Electronics and Communication Systems (ICECS) ,2014 International Conference on (pp. 1–6). IEEE, 2014.

[5] Purwar, A.,andSingh,S.K., Hybridpredictionmodel withmissing value imputation formedical data. ExpertSyst. Appl.42(13):5621– 5631, 2015. .

[6] 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.

[7] Amin,S.U.,Agarwal,K.,&Beg,R.,Geneticneuralnetwork based data mining in prediction of heart disease using risk factors. In Information & Communication Technologies (ICT), 2013 I.E. Conference on (pp. 1227–1231). IEEE, 2013.

[8] L.C. Jain et al., A new ICA based Algorithm for prediction of coronary heart disease, Advances in Intelligent Systems and Computing 309, DOI 10.1007/ 978-81-322-2009-1_47,2015

[9] Raghavendra.S, Indiramma., Classification and Prediction Model using Hybrid Technique for Medical Datasets, International Journal of Computer Applications(0975-8887) Volume 127-No.5, ctober 2015

[10] Melillo, P.,Izzo, R., Orrico,A., Scala, P., Attanasio, M., Mirra, M., and Pecchia, L., Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. PLoS One. 10(3):e0118504, 2015.

[11] Karaolis,M.A.,Moutiris,J.A.,Hadjipanayi,D.,andPattichi s,C., Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Trans. Inf. Technol. Biomed. 14(3):559–566, 2010.