Tanvi Sharma et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 4, Issue 2, June 2017, pp. 94-97

Intelligent Heart Disease Prediction System using Machine Learning: A Review

Tanvi Sharma, Sahil Verma, Kavita

Kurukshetra University, Kurukshetra (Haryana)

Abstract: **Heart disease is a major life threatening disease that can cause either death or a serious long term disability. However, there is lack of effective tools to discover hidden relationships and trends in e-health data. Medical diagnosis is a complicated task and plays a vital role in saving human lives so it needs to be executed accurately and efficiently. An appropriate and accurate computer based automated decision support system is required to reduce cost for achieving clinical tests. This paper provides an insight into machine learning techniques used in diagnosing various diseases. Various data mining classifiers have been discussed which has emerged in recent years for efficient and effective disease diagnosis.**

Keywords: **Machine Learning, Data Mining, Heart Disease, Diagnosis, Classification**

I. INTRODUCTION

The Heart is one of the most important organs in the human body. It is the center of the circulatory system. The heart functions as a pump that propels blood to different parts of the human body through a network of blood vessels, supplying a constant supply of oxygen as well as other vital nutritional components. If the heart ever stops functioning and ceases to pump blood, the body will shut down and within very less time a person will expire.

The usage of information technology in health care industry is increasing day by day to aid doctors in decision making activities. It helps doctors and physicians in disease management, medications and discovery of patterns and relationships among diagnosis data. Current approaches to predict cardiovascular risk fail to identify many people who would benefit from preventive treatment, while others receive unnecessary intervention. Machine-learning offers opportunity to improve accuracy by exploiting complex interactions between risk factors. We assessed whether machine-learning can improve cardiovascular risk prediction.

"Cardiovascular disease is the leading cause of illness and death worldwide," said Dr. Stephen Weng, of Nottingham University's National Institute for Health Research School [1]. "Our study shows that artificial intelligence could significantly help in the fight against it by improving the number of patients accurately identified as being at high risk and allowing for early intervention by doctors to prevent serious events like cardiac arrest and stroke." Based on their results, it is clear that artificial intelligence and machine learning techniques have a key role in fine-tuning risk management strategies for individual patients.

This paper is organized as follows: Section 2 gives the classification of machine learning prediction techniques. Section 3 describes the work in the literature regarding classification algorithms for medical diagnosis of cardiovascular and heart diseases. Section 4 concludes the survey.

II. MACHINE LEARNING PREDICTION TECHNIQUES

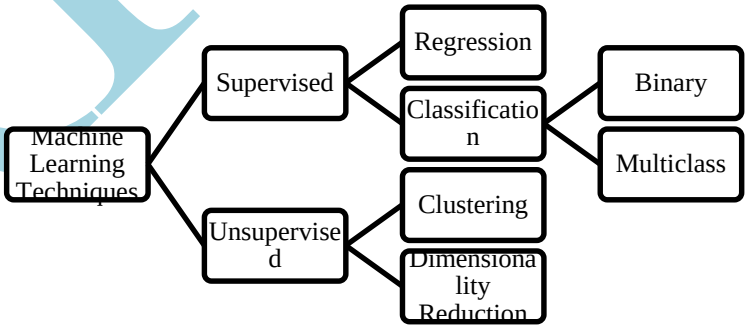

Figure 1: Taxonomy of machine learning techniques

These techniques as shown in Figure 1 can be predominantly be classified as:

1. Supervised learning: In supervised learning, answer is already known. The main idea is to design a model that can predict the answers for unknown instances. We give the machine the data, with inputs and outputs (the answers), and let it learn from the relationships between them. We can further divide supervised learning techniques into regression and classification based on the variable being predicted. If the predicted value is continuous, it is Regression problem. Otherwise, if the variable to be predicted is one of various independent categories, called classes, it is known as Classification problem. If there are two classes, then it is binary classification and if we have multiple classes then there

is multiclass classification problem. Some of the supervised algorithms are below:

- **Support Vector Machine:** It constructs a hyper plane as shown in Figure 2 i.e. a plane in an infinite dimension plane to classify the training data points into clearly demarcated classes [2]. The construction of an optimal decision plane for classification requires minimizing the error function. The shape of the error function becomes the foundation for further classification of these algorithms in the broad categories of linear, polynomial, sigmoid and radial SVMs. So putting in simple terms the philosophy of SVM is to obtain an optimal hyper plane for data points which are linearly separable. Support vectors actually refer to the data points that are closest to the demarcating surface which are hence tricky to classify. The metric that alludes to the optimality of a hyper plane is the margin around the hyper plane. So the problem transitions into that of an optimization one. As established the maximum margin classifier learnt and derived from the training data would lead us to optimal hyper plane. This is achieved by transforming the maximal margin classifier as the inner product (sum of multiplication of pair values) of two given data points rather than the data points. The general kernel function could then be defined as follows [3]:

$$f(x) = B_0 + \sum_i (a_i \times (x, x_i))$$

Here x is the new input vector and the coefficients B_0 and a_i must be estimated from training data.

- **Decision Tree:** Decision tree are supervised method used for the prediction of categorical as well as numerical value [4].They represents the data instances along with their class label in the form of a tree. A set of rules can be deduced from the tree which can be used to classify the unknown data record to its output value. A test on an attribute is performed on the internal node. The result of the test is depicted by the branch of tree and class label are present at the leaf node. In this technique the whole data set or the whole collection of sample points in split into two or more homogenous classes. The split is established from the parameter or the factor which is determined to be the best splitter or differentiator.

- **Naïve Bayes:** Rather than a single classifier it actually is a combination of multiple classifiers all working on the basic Naïve Bayes principle of independent features [5]. Hence each feature is assumed to be independent and autonomous contributing individually to the training data point's probability of belonging to a particular class.

As per the Bayes theorem [5],

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c)P(x_2|c) \times \ldots \times P(x_n|c)P(c)$$

Here
$P(c|x)$is the posterior probability of class given predictor
$P(c)$is the prior probability of class
$P(x|c)$is the likelihood probability of predictor given class
$P(X)$is the prior probability of predictor

- **Artificial Neural Network:** These are used to model/simulate the distribution, functions or mappings among variables as modules of a dynamic system associated with a learning rule or a learning algorithm. The modules here simulate neurons in nervous system and hence ANN collectively refers to the neuron simulators and their synapsis simulating interconnections between these modules in different layers [6]. The defining aspect of an ANN is the function implemented at each neuron and the learning algorithm for the dynamic weights assigned to the interconnections among neurons. What makes ANN stand apart is its ability to simulate human thought process coupled with continuous learning, growth and evolution. Also it is capable of handling large number of parameters and large set of data with noise and yet achieves high accuracy.

2. Unsupervised learning: In unsupervised learning, our aim is to find unknown trends. The data has no associated labels, but we want to organize the data into groups or clusters. Unsupervised learning techniques are further classified as Cluster Analysis and Dimensionality Reduction. In Cluster Analysis, data is grouped according to similarities or distances between them. In Dimension Reduction, duplicated or unnecessary variables are removed to produce a smaller subset of the original data.

III. LITERATURE REVIEW

Work done by various researchers in the field of heart disease diagnosis using machine learning techniques has been discussed in this section.

Das et al [7] introduced a neural network ensemble method based on SAS software 9.1.3 for diagnosing of the

heart disease. It combined posterior probabilities or the predicted values from multiple predecessor models. They obtained 89.01% classification accuracy on Cleveland heart disease dataset. Anbarasi et al [8] used three classifiers like Naive Bayes, classification by clustering and decision trees for heart disease prediction using 13 attributes first and then applied feature subselection using genetic algorithm and obtained almost similar accuracy. They found that decision tree data outperforms other two classifiers with accuracy 99.2% for binary classification. The accuracy of classification clustering is 88.3% and Naïve Bayes is 96.5%.

Vanisree et al. [9], proposed a Decision Support System for diagnosis of Congenital Heart Disease. The proposed system is based on multi layered feed forward neural network known as backpropagation neural network. Delta learning rule is applied for training the model. The dataset consists of 200 samples with 36 attributes each depicting signs, symptoms and the results of physical evaluation of a patient. The proposed system employs 80-20 rule for training and testing. An overall accuracy of 90% and mean square error of 0.016 is achieved.

Zhang et al. [10] proposed an efficient coronary heart disease prediction system using Support Vector Machine. In this, Principal Component Analysis (PCA) was used to extract the important features and different kernel functions were utilized as a classifier. The highest classification accuracy is achieved with Radial Basis Function (RBF). To find the optimal parameters values, Grid search method was employed and optimal values were found to be c=1 and g=0.0909. The highest classification accuracy reached is 88.6364%. It was used for prediction of two classes.

Vadicherla and Sonawane [11] suggested a sequential minimal optimization (SMO) technique of SVM for heart disease diagnosis system. The system is proposed for two classes. SMO helps in training of SVM by finding the optimal values of multipliers required during training phase. The result reveals that SMO shows good results even on large dataset and performance time is also improved.

Elshazly et al. [12] presented a classification approach called GA-SVM for lymph disease diagnosis in which genetic algorithm (GA) is used to reduce the number of features of the dataset from 18 features to 6 features. The experiments were performed with 10-fold cross validation. Different kernel functions were employed and for each function, performance was evaluated by measures like accuracy, sensitivity, area under curve (AUC), F-measure. The result indicates that GA-linear classifier achieved best results of 83.1% accuracy with 82.6% sensitivity, 82.7% F-measure and 84.9% AUC.

Masethe [13] performed a comparison of various data mining algorithms on WEKA tool for the prediction of heart attacks to find the best method of prediction. The algorithms used are J48, REPTREE, Naïve Bayes, Bayes net and CART with prediction accuracy as 99.07%, 99.07%, 97.22%, 98.14%, 99.07% respectively. It also works for binary classes and experiment was performed on WEKA tool. Prerana et al. [14] predicted the heart disease risk level of a patient using machine learning algorithms such as Naïve Bayes Classification and Probabilistic Analysis and Classification. They also created a centralized System for both doctors and patients to login and view the e-health data on cloud.

Dey et al. [15] analyzed SVM, Naive Bayes and Decision tree with and without using PCA for attribute selection to predict heart disease. The dataset contains both type of patients i.e. those who does and does not have heart disease so it is binary classification problem. They observed that SVM outperformed the other two. Weng et al. [16] compared four machine learning algorithms namely random forest, logistic regression, gradient boosting machines and neural networks to predict first cardiovascular event over 10-years. Grid search was used for parameter optimization.

IV. CONCLUSION

This survey provides the deep insight into machine learning techniques for classification of heart diseases. The role of classifier is crucial in healthcare industry so that the results can be used for predicting the treatment which can be provided to patients. The existing techniques are studied and compared for finding the efficient and accurate systems. Machine learning techniques significantly improves accuracy of cardiovascular risk prediction through which patients can be identified during an early stage of disease and can be benefitted by preventive treatment.

References

[1] https://www.digitaltrends.com/computing/artificial-intelligence-cardiovascular-disease/

[2] G. Kesavaraj and S. Sukumaran, “A study on classification techniques in data mining,” Computing, Communications and Networking Technologies 2013 Fourth International Conference (ICCCNT), pp. 1–7, IEEE, 2013.

[3] Suykens, Johan AK, and JoosVandewalle. "Least squares support vector machine classifiers." *Neural processing letters* vol. 9, 3, 1999, pp. 293-300.

[4] J.R. Quinlan, “Induction of Decision Trees,” *Expert System*, vol. 1, no. 1, pp. 81–106, 2007.

[5] D.S. Medhekar, M.P. Bote, and S. D. Deshmukh, “Heart Disease Prediction System using Naive Bayes,” International Journal of Enhanced Research in Science Technology and Engineering, vol. 2, no. 3, Elsevier 2013.

[6] Dreiseitl, S. and Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5), pp.352-359.

[7] Das, R., Turkoglu, I. and Sengur, A., 2009. Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, *36*(4), pp.7675-7680.

[8] Anbarasi, M., Anupriya, E. and Iyengar, N.C.S.N., 2010. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal*

of Engineering Science and Technology, *2*(10), pp.5370-5376.

[9] Vanisree K, Jyothi Singaraju, "Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks", International Journal of Computer Applications, vol. 19, no. 6, pp. 6-12, 2011.

[10] Y. Zhang et al., "Studies on application of Support Vector Machine in diagnose of coronary heart disease," Electromagnetic Field Problems and Applications 2012 Sixth International Conference (ICEF), Dalian, IEEE 2012.

[11] D. Vadicherla, and S. Sonawane, "Decision Support System for Heart Disease Based on Sequential Minimal Optimization in Support," *International Journal of Engineering Sciences and Emerging Technologies*, vol. 4, no. 2, pp. 19–26, 2013.

[12] H. I. Elshazly, A. M. Elkorany, and A. E. Hassanien, "Lymph diseases diagnosis approach based on support vector machines with different kernel functions," *Computer Engineering & Systems 9th International Conference (ICCES)*, Cairo, pp. 198–203, 2014.

[13] Masethe, H.D., Masethe, M.A.: Prediction of heart disease using classification algorithms. In: World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, San Francisco, USA, 22–24 Oct 2014.

[14] Prerana T H M, Shivaprakash N C, Swetha N, "Prediction of Heart Disease Using Machine Learning Algorithms-Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS", International Journal of Science and Engineering, Volume 3, Number 2 – 2015, pp. 90-99.

[15] Dey, A., Singh, J. and Singh, N., 2016. Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis. *Analysis*, *140*(2), pp. 27-31

[16] Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N., 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. *PloS one*, *12*(4), p.e0174944.

