International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 5 Issue 9|| September 2016 || PP. 17-20

Classification of Health Care Data Using Machine Learning Technique

H.S. Hota[1], Seema Dewangan[2]
[1]*Bilaspur University, Bilaspur, India,*
[2]*Dr. C.V.Raman University, Kota, Bilaspur , India*

Abstract: *Diagnosis of any diseases using intelligent system produces high accuracy and treated as classification problem, also classification of health care data using machine learning techniques may be used as intelligent health care system (IHSM). Diagnosis of heart disease is a major issue and commonly found in human being due to changing life style and also need to be identified in advance. This paper explores various machine learning techniques to classify heart data downloaded from UCI repository site. After applying feature selection technique (FST), a decision tree based technique: CART is producing high accuracy with four features followed by other classification techniques.*

Keywords: *Machine Learning, Feature Selection Technique (FST), Decision tree (DT), Intelligent Health Care System (IHCS).*

I. Introduction And Framework

Diagnosis of any human disease must be accurate and efficient as it is directly related to human life. Physician mainly diagnoses human disease based on their experience and using pathological data. Involvement of Intelligent Health Care System (IHCS) in this process may improve the quality of decision making process. Health care system is now a day is a thrust research area; many intelligent health care systems have been developed to diagnose many human diseases.

Machine learning techniques are widely used to design and development of IHCS due to its ability to diagnose the problem in more efficient way. Many authors have developed many IHCSs to diagnose many human diseases using machine learning techniques. Authors have used K-means clustering, squared error criterion, genetic algorithm to diagnose hepatitis disease. Also neural network, decision tree, and naïve Bayes along with feature selection techniques were used to diagnose heart disease. A significant contribution is done by [6,10] , they have used many decision tree based techniques to diagnose many diseases. These techniques have also used by [4] for classification of health care data. An ensemble of ANN and SVM model is developed by D. Sharma and et al. [2] to diagnose dermatology disease and achieved 98.99% accuracy at testing stage. They have [3] also developed an ensemble model for life threatening diseases like Breast cancer and diabetes.

Literatures prove that machine learning techniques may be the best choice for health care diagnosis. This paper explores the use of machine learning techniques for classification of heart data downloaded from UCI repository site [8]. A best classification model obtained through experimental work is further used to apply feature selection technique (FST) to remove irrelevant features from heart data set. Figure 1 show a conceptual framework to apply the overall process. In order to train and test various classification models, data is randomly divided into two parts: Training and Testing where training data set is used to build the model and testing data set is supplied to induct decision tree or for building models, many decision tree techniques are applied and the best model is considered for further process, once satisfying results are obtained at training stage, testing data are supplied to validate the model. The performance of models is evaluated with the help of three error measures: accuracy, sensitivity and specificity as shown in equations 1, 2 and 3 respectively and at last ranking based feature selection technique was applied to remove irrelevant features from the data set to obtain best feature subset.

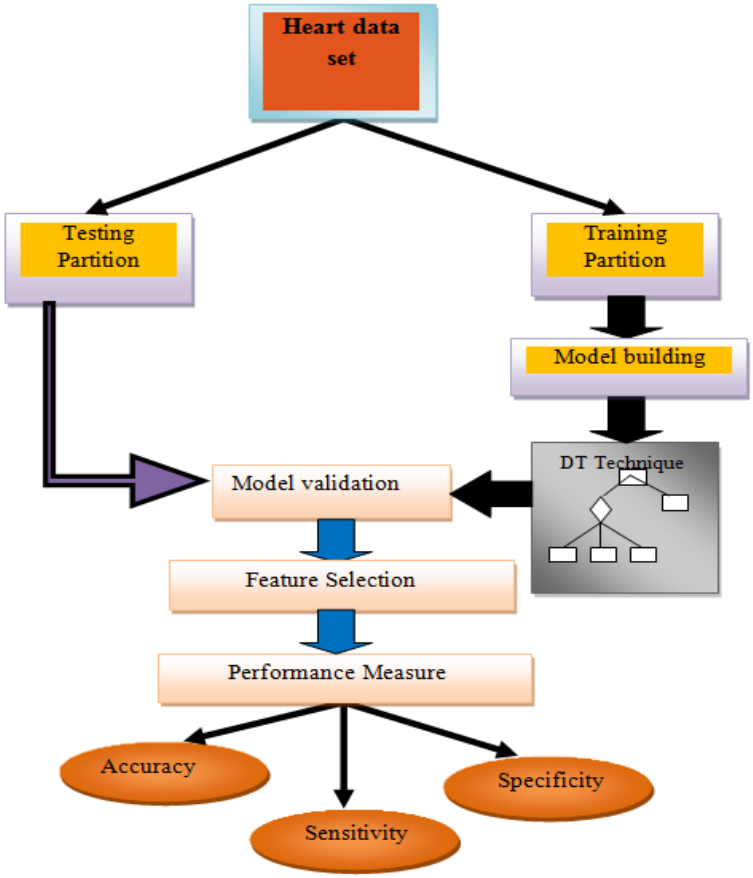

Figure1: Framework of proposed model

Accuracy used in this model to measure the quality of generated performance. Where TP (True Positive) is the positive cases which are classified correctly as positive, TN (True Negative) is the negative cases that are classified as negative, FP (False Positive) are cases with negative class classified with positive, and FN (False Negative) is the case with positive class classified as negative. The equation of accuracy is shown in equation (1).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots (1)$$

Sensitivity is the proportion of the cases with positive class that are classified as positive (true positive rate). Equation is as follow:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \ldots (2)$$

Specificity is the proportion of cases with negative rate class, classified as negative (true negative rate, shown as a percentage). Equation of specificity is shown in equation (3).

$$\text{Specificity} = \frac{TN}{TN+FP} \quad \ldots (3)$$

II. Research Data And Techniques

The propose research work is based on classification of heart data which is downloaded from UCI repository site [8], this dataset consists 303 samples with 14 features and two classes.

Techniques used to classify above data are mainly decision tree (DT) techniques which is [5] a popular and powerful classification technique where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. DT is inducted based on importance of features in a tree like structure. Many DTs used in the research work are as follows:

Classification and Regression Technique (CART): Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefine set of classes or group. CART [1] method uses recursive partitioning to split the training records into segments with similar output field values using Gini index.

BFTREE: Best First Tree (BFTREE) is an algorithm for traversing or searching tree or graph data structures. It starts at the tree root, sometime referred to as search key and explore the neighbor nodes first, before moving to the next level neighbors.

C4.5: C4.5 is also a DT based algorithm suggested by [7], C4.5 creates a decision tree based on a set of labeled input data. The decision tree generated by C4.5 can be used for classification, it is a successor of ID3 (Iterative Dichotomies).

Feature Selection Technique: Feature selection is a process by which, algorithm automatically search for the best subset of features from dataset. Ranking based feature selection techniques gives rank based on the feature data.

III. Experimental Work

For experimental work, the open source data mining tool WEKA [9] is used. WEKA support many machine learning tools and work can be simulated in easy and graphical way. After loading heart data [8] input and output features are selected and supplied to the decision tree model one-by-one, results are automatically derived and presented in form of various error measures. However we have considered three error measures as explained above, as these three measures clearly reflects efficiency of classification model as shown in Table 1. Among the three decision tree techniques, CART is producing remarkable results. The same is shown in Figure 2 in form of bar graph. A comparative performances of various DTs using FST are also shown in Table 2 with 14 and 4 features, table show that CART (84.82% accuracy, 87.1% sensitivity, 83.24% Specificity) is performing better than other two decision tree techniques with only 4 features as compare to C4.5, and BF Tree.

Table1: Comparative results of three decision tree techniques

Machine Learning Technique	Accuracy	Sensitivity	Specificity
CART	80.86	81.74	80.22
C4.5	77.56	77.8	77.4
BFTREE	77.9	77.95	77.84

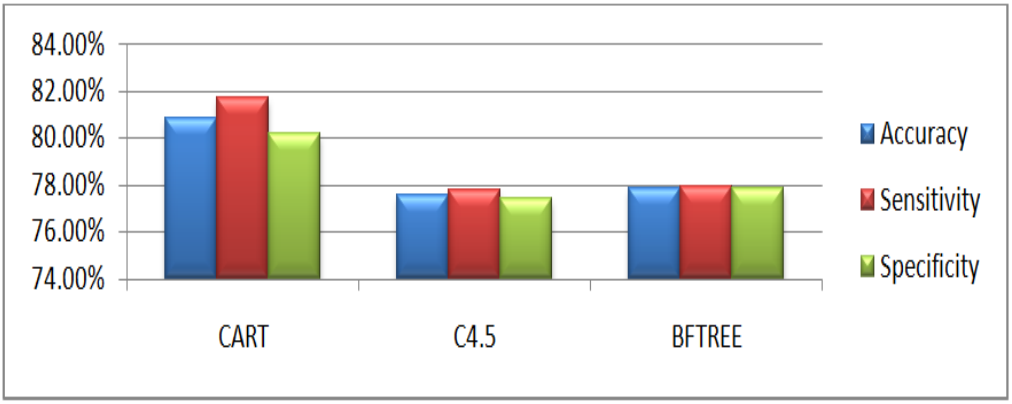

Figure 2: Comparative graphs of three different decision tree techniques

Table 2: Comparative performance after applying feature selection technique

Technique →	CART		C4.5		BF Tree	
No. of Feature →	14	4	14	4	14	4
Accuracy	80.86	84.82	77.56	81.52	77.9	83.17
Sensitivity	81.74	87.1	77.8	83.06	77.95	86.55
Specificity	80.22	83.24	77.4	80.45	77.84	80.98

IV. Conclusion

Intelligent Health Care System (IHCM) is the need of the today's medical world to diagnose critical diseases in more accurate way, this system may be helpful for the medical practitioners as well as for the medical students. In this paper, many machine learning techniques are explored along with rank based feature selection technique to classify heart data. Experimental results are obtained using WEKA which shows that CART is performing better than other two DTs even after applying FST as 84.82% accuracy with only four features.

References

[1]. Breimen R. and Anand, T. The process of knowledge discovery in databases: A human centered approach.In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds),Advances in Knowledge Discovery and Data Mining.Cambridge:MIT press,1996.

[2]. Dinesh K. Sharma and Hota, H.S., Data mining techniques for prediction of different categories of dermatology disease, Academy of Information and Management Sciences Journal (AIMSJ): A journal of Allied Academies, USA, 16(2), 2013, 103-116.

[3]. Dinesh K. Sharma, Hota H.S., Development of rule base system using intelligent techniques to diagnose life threatening diseases , Proceeding of review of business and technology research (RBTR), (1), 2013, 14-19.

[4]. Gupta S., Kumar, D., and Sharma, A., Performance analysis of various data mining classification techniques on health care data, International Journal of Computer Science and Information Technology, 3(4), 2011, 155-169.

[5]. J. Han K. Micheline, Data mining: Concepts and Techniques, Morgan Kaufmann Publisher, 2009.

[6]. Bendi V. R., Prasad, M. S. Babu and Venkateswarlu N. B. ,A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosi, International Journal of Database Management Systems (IJDMS), Vol.3, No.2, PP 101-114, 2011.

[7]. Quinlan J.R., C4.5: Programs for machine learning (1st edition), San Francisco, Morgan Kaufmann Publishers, (1993).

[8]. UCI Machine Learning Repository of machine learning database, University of California, school of Information and Computer Science, Irvine. C.A. http://www.ics.uci.edu/ last accessed on Dec Oct 2015.

[9]. WEKA (2016), A open source data mining tool, web source: http://www.cs.waikato.ac.nz/ml/weka last accessed on July,2016.

[10]. Bendi V.R., Prasad, M. S. Babu and Venkateswarlu, N. B., A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis, International Journal of Computer Science Issues, Vol.9. Issue 3 ,No. 2 ,PP 506-516, 2012