A Novel Comparison of Neural Network and Decision Tree as Classifiers using R.

Dr.J.Vijayarangam,
Assistant Professor
Department of Applied Mathematics
Sri Venkateshwara College of Engineering
Chennai, India
jvijayarangam@yahoo.com

Dr.S.Kamalakkannan
Associate Professor
Department of Information Technology
Vels Institute of Science, Technology & Advanced Studies
Chennai, India
kannan.scs@velsuniv.ac.in

Dr.J.Anita Smiles,
Assistant Professor
Department of Information Technology
VISTAS
Chennai, India
anitasmiles78@gmail.com

Abstract—Machine Learning tools are currently a hot one among the analysis tools and there are many tools in that tool kit for tasks like classification. So, it is natural to compare them for efficiency to find a suitable one for a particular problem. This is a novel comparison of neural network and decision tree in that this research work has built a decision tree, prune it based on best cp, build two neural networks-one by general method and the other one using the variables in the pruned tree. The dataset used for classification is the heart disease dataset and this research uses R for building the tools. Then the Precision, Recall, Accuracy and F ration of all three models are obtained and compared.

Keywords: Classification, machine learning, decision tree, neural network, heart disease, comparison.

I. INTRODUCTION

Machine learning based analysis is the hottest problem in the planet now and it is applied for every conceivable problems possible. Machine Learning tool box has many tools for various problems which could fall mostly into two broad sets, the classification and the regression problems. Even for a single problem like classification, there are many tools in machine learning like Decision trees, Neural networks, Random forests, logistic regression. So, it is really a straightforward issue to address the selection of a suitable Machine learning tool for a particular problem and their efficiency.[1] is literally a go to book nowadays to get introduced to machine learning basics.[2] is a recent article which compares various machine learning tools and deep learning methods for diabetes prediction and concludes that CNN based methods are the best of the lot.[3] is also a paper evaluating the classifiers Naïve Bayes, SVM and Decision tree on PIDD dataset.

[4] predicts the disease diabetes using classification and clustering techniques and also Naïve Bayes on medical data. [5] uses datasets collected from two different places and two different sensors to evaluate univariate and multivariate decision trees in classification. [6] uses Decision tree based classification of mass spectral data towards diagnosis of ovarian cancer.[7] gives a detailed discussing about ID3, a decision tree analysis by explaining in detail the way a decision tree is built using the Shannon's Entropy concepts and also discusses some applications of decision trees. [8] is a heart disease prediction paper using Fog computing based WSN's.

This paper compares the efficiency of two machine learning classification tools, the decision tree and neural network in classification problem of heart disease.

Decision trees are inverted tree like structures developed towards classification and regression. There are many ways of building a Decision trees. All of them have common basic notion of a root node, the first branching point followed by many further branches before the terminal leaf nodes which are used for classification or regression. At any stage of branching, a feature vector from the underlying dataset is used for branching and the selection of a feature vector as a branching variable is where the various methods are involved like the entropy based or index based ones. Once there is an ability to build a decision tree for a data, it will be traversed from root node to any of the leaf nodes to create a rule, which is the basis for classification. A sample decision tree is provided in figure 1.
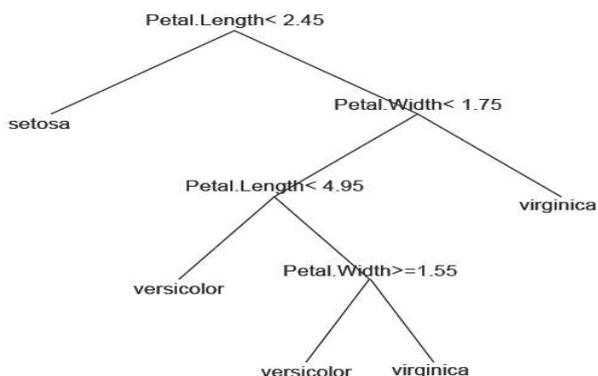
FIGURE 1- A SAMPLE DECISION TREE- FOR IRIS DATASET

FIGURE 2- A NEURAL NETWORK, HUMAN BIOLOGICAL(LEFT) AND THE ARTIFICIAL ONE(RIGHT).

Neural networks are models inspired by how a human brain works. Their structure also has resemblance to human brain , provided in figure 2. Technically speaking, there are three types of layers in any neural network, the input layer, the hidden layer where the processes and complexities are addressed and an output layer which is where we get the classification done. This is considered by many as a black-box tool as we are not in full control or knowledge of how and what happens in the processing of a neural network which is still now an area of research. The only thing anyone can be very clear about is regarding the input and output layers and the number of nodes in them. The number of hidden layers and the number of nodes in them which form the back bone of the whole process is what is being referred to as the black-box region as we mostly work in trial and error way. Regardless, this is considered by many as an exciting model to work with in classification problems.

The heart dataset is a data of persons from whom 13 aspects are recorded to explain whether a person has heart diseases or not . A sample of the data is provided in table 1.
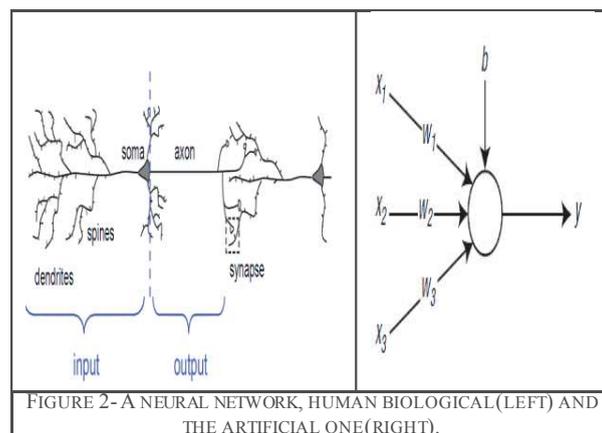
TABLE 1- HEART DATASET SAMPLE ROWS

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	Target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

The table has 13 feature vectors (columns 1-13) , age-gender-chest pain type-serum cholesterol-fasting blood sugar-max heart data...-and a class vector (col14-target) which is binary taking values 1 or 0 indicating whether the individual has heart disease or not.

II. METHODOLOGY

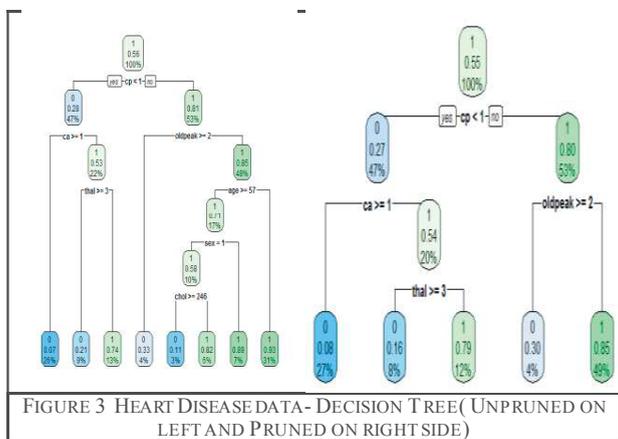For the heart data set , build a Decision tree using R.

Prune the Decision tree using the best cp.

Obtain a neural net using R and note down the error and the number of iterations to converge.

If the variables in the neural net matches those in the decision tree, then move on to analysis, else, obtain one more neural net using R specifically for the variables present in the pruned decision tree so that we are able to perform a reasonable comparison.

Calculate Precision, Recall, Accuracy and F ratio for all three models to compare.

III. RESULTS AND ANALYSIS

FIGURE 3 HEART DISEASE DATA- DECISION TREE(UNPRUNED ON LEFT AND PRUNED ON RIGHT SIDE)

The rpart package in R is employed to get the decision trees.

The above (figure3) are the unpruned and the pruned decision trees for the heart Dataset. We can clearly observe the change in the structure and the variables.

TABLE2- MEASURES OF THE DECISION TREES

	Unpruned		Pruned	
cp	xerror		cp	xerror
0.471	1		0.476636	1
0.072	0.6087		0.079439	0.52336
0.0217	0.471		0.037383	0.38318
0.0144	0.42754		0.01	0.41121
0.01	0.49275		.-	.-

Variable Importance	cp	thal	oldpeak	thalach	exang	ca	slope	age	chol	sex	fbs
Unpruned	37.6836	21.222	20.292	19.479	16.753	13.051	8.933	8.286	6.922	5.343	2.599
Pruned	33.443	20.225	18.36	15.829	15.094	11.927	8.566	4.224	.-	3.689	1.462

The variable importance of the two trees and their cp with the xerror is recorded in the table2.
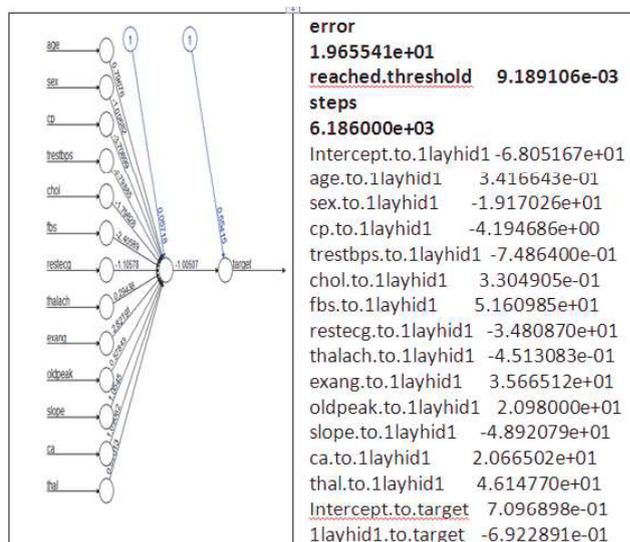

FIGURE 4 NEURAL NET FOR HEART DATASET AND ITS DETAILS

Figure 4 is the neural net, using the "neuranet" package in R, for the dataset and upon comparing with the decision tree obtained, figure 5 is the Neural network using the decision tree variables. We can very clearly observe the bolded values which are significantly better for the one of DT based.
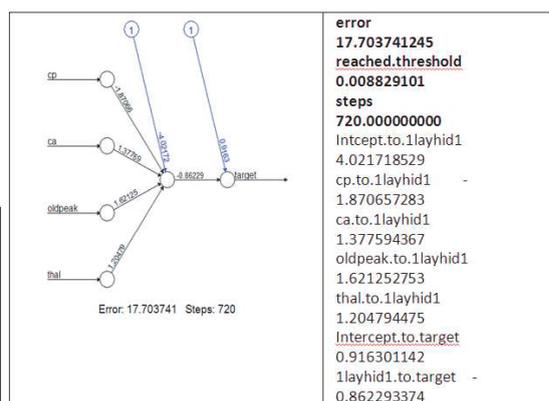

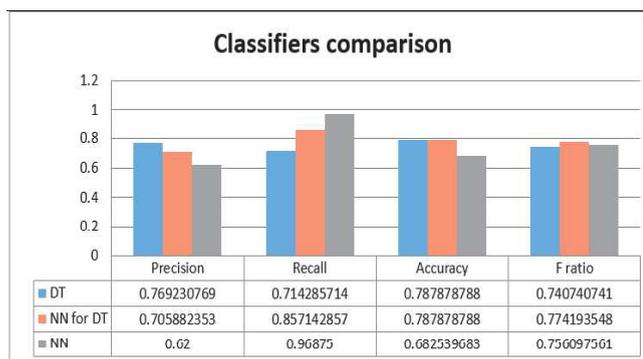FIGURE 5-NEURAL NET USING DECISION TREE VARIABLES AND ITS DETAILS

FIGURE 6- COMPARISON OF THE THREE CLASSIFIERS

From figure 6, we could clearly observe that the precision is low for a general NN as compared to the NN based on the DT as well as the DT whereas its recall is higher than both of them. The Accuracy of the general NN is lower than the other two but the F ratio is of no help to us in this comparison effort.

IV. CONCLUSION

This is a novel comparison of classifiers . We are using a refined form of a classifier(the neural network) where the refinement factor is in fact from another classifier. We could deduce that the neural network based on the pruned decision tree is much better in the converging aspect as well as the measurement aspects. So, in a way we could see a hybrid mode of classification for a dataset, first using a decision tree and then using the neural network. This also could be viewed as a way of validating our classification using a decision tree. We also see a way of controlling the neural network building. So, there are many positives available from this novel way of comparing the classifiers.

REFERENCES

[1] Tom M.Mitchell, Machine Learning, McGrawHill,2017.

[2] Abhaya Kumar Sahoo, Chittaranjan Pradhan and Himansu Das, Performance Evaluation of Different Machine Learning Methods and Deep-Learning Based Convolutional Neural Network for Health Decision making , Nature inspired Computing for data science, Springer,Volume SCI 871, 2020

[3] Deepti Sisodia, Dilip Singh Sisodia, Prediction of Diabetes using Classification Algorithms, Procedia Computer Science 132 (2018) 1578−1585,

[4] Das, H., B. Naik, and H.S. Behera. 2018, Classification of diabetes mellitus disease (DMD): a data mining (DM) approach, In Progress in computing, analytics and networking, 539−549, Springer

[5] Mahesh Pal, Paul M. Mather, An assessment of the effectiveness of decision tree methods for land cover classification, Remote Sensing of Environment 86 (2003) 554−565

[6] Antonia Vlahou et al, Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data , Journal of Biomedicine and Biotechnology (2003) 308−314.

[7] S.Mathivilasini, KonduriSuchithra, Hemasankari, J.Vijayarangam(2020), ID3-Decision trees in Machine learning, Alochana Chakra Journal, Volume IX,Issue VI, 9125-9132.

[8] Shakya, Subarna, and P. P. Joby. "Heart Disease Prediction using Fog Computing based Wireless Body Sensor Networks (WSNs)." IRO Journal on Sustainable Wireless Systems 3, no. 1 (2021): 49-58

[9] IlyesJenhani et al.,Decision trees as possibilistic classifiers, International Journal of Approximate reasoning,volume 48,Issue 3, Aug2008,784-807.

[10] Hemashankari,S.Mathivilasini,S.Dilliarasu,J.Vijayaranagm, Decision Trees of Machine Learning applied For Adult dataset,MuktShabdJournal,Volume IX, Issue VI, JUNE/2020,4641-4646.

[11] Ming, H., Wenying, N. and Xu, L., (2009) "An improved decision tree classification algorithm based on ID3 and the application in score analysis", Chinese Control and Decision Conference (CCDC), pp1876-1879.

[12] Agrawal, R., Imielinski, T., and Swami, A. 1993. Database mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering, 5(6):914−925.

[13] C.Andrieu et al., An Introduction to MCMC for Machine Learning, Machine Learning,50,5-43(2003).

[14] KF Foster,R.Koprowski,JDSkufca, Machine Learning, medical diagnosis and biomedical engineering research-commentary, Biomedical engineering online,13, Article No 94(2014).

[15] Riedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. The Elements of Statistical Learning. Vol. 1. Springer Series in Statistics New York, NY, USA:

[16] Paola, J. D., & Schowengerdt, R. A. (1995), A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification, IEEE Transactions on Geoscience and Remote Sensing, 33, 981−996.

[17] Hepner, G. F., Logan, T., Ritter, N., & Bryant, N. (1980), Artificial neural network classification using a minimal training set: Comparison to conventional supervised classification,Photogrammetric Engineering and Remote Sensing, 56, 469−473

[18] Esposito, F., Malerba, D., & Semeraro, G. (1997). A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19, 476−491.

[19] Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford: Clarendon Press.

[20] Bennett, K. P., & Blue, J. A. (1998). A support vector machine approach to decision trees. Proceedings of the IEEE international joint conference on neural networks, Anchorage, Alaska (pp. 2396−2401).