Feature Selection and Hyper-Parameter Tuning for Optimizing Decision Tree Algorithm on Heart Disease Classification

Tsehay Admassu Assegie[1†], Sushma S.J[2††], Bhavya B.G[3†††] and Padmashree S[4††††]

[1†]Leturer, Aksum University, Department of Computer Science, Axum, Ethiopia
[2††]Associate Professor, Department of Electronics and Communication Engineering, GSSS Institute of Engineering and Technology for Women, MYSURU, India.
[3†††]Assistant Professor, Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, India
[4††††]Professor, Department of Electronics and Communication Engineering, GSSS Institute of Engineering and Technology for Women, MYSURU, India

Summary
In recent years, there are extensive researches on the applications of machine learning to the automation and decision support for medical experts during disease detection. However, the performance of machine learning still needs improvement so that machine learning model produces result that is more accurate and reliable for disease detection. Selecting the hyper-parameter that could produce the possible maximum classification accuracy on medical dataset is the most challenging task in developing decision support systems with machine learning algorithms for medical dataset classification. Moreover, selecting the features that best characterizes a disease is another challenge in developing machine-learning model with better classification accuracy. In this study, we have proposed an optimized decision tree model for heart disease classification by using heart disease dataset collected from kaggle data repository. The proposed model is evaluated and experimental test reveals that the performance of decision tree improves when an optimal number of features are used for training. Overall, the accuracy of the proposed decision tree model is 98.2% for heart disease classification.

Keywords:
Heart disease classification, Feature selection, Chi-squared test, Parameter tuning, optimizing decision tree.

1. Introduction

Heart disease is one of the deadliest disease in the world [1]. According to the study [1], one of the main reason that leads to heart disease is life style. Hence, heart disease is getting major focus in medical research. Early detection of heart disease and prediction of the likelihood that heart disease will affect a person is automated by using machine-learning algorithms. However, there is no machine-learning model that is in place for identification of heart disease with reliable accuracy.

A preliminary literature reviews [2-19] shows that one of the most difficult task in machine learning is determining the smallest set of features that correctly characterizes the training data. Statistical method such as correlation analysis is commonly employed to find the relationship between features and the target class. Correlational analysis helps in determining potential features with better score on training data. However, the correlational analysis has disadvantage of not detecting the relationship between features, particularly when the features have independent relationship. Overall, there are limited research work on feature selection and the use of statistical approaches such as χ 2 (chi-squared) test for determining user specified percentile of features and their effect on the accuracy score of machine learning algorithm. Therefore, this study investigates the answers to the following research questions:
1. How to find statistically important features for training decision tree on heart disease dataset to optimize the classification accuracy?
2. What is the optimal numbers of features that provide better accuracy for decision tree on heart disease classification?
3. What is the effect of parameter tuning on the performance of decision tree for heart disease classification?

2. Literature review

This section discusses some of the research works on heart disease classification with machine learning algorithms. In [8], the authors compared the performance of different machine learning algorithms, such as decision tree, K-nearest neighbor (KNN) and support vector machine for heart disease classification. The comparative result on the performance of the decision tree, support vector machine and KNN appears to prove that the KNN algorithm

Manuscript received February 5, 2024
Manuscript revised February 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.2.18

performed better for heart disease classification as compared to the decision tree and support vector machine. As showcased by the authors in their study, decision tree and support vector machine performed slightly the same for heart disease classification. The highest accuracy achieved by the KNN is 85%. A performance of 85% for heart disease classification with KNN is promising result. However, there is still a larger scope for improving the performance of machine learning algorithms for heart disease classification.

In [9], the authors proposed a convolutional neural network based heart disease classification model. The authors evaluated the model on test set and result appears to prove that the proposed model is effective for heart disease classification. However, there is much scope for improving the accuracy of the proposed model.

In another study [10] deep neural network, based chronic based heart disease diagnosis model is proposed. The proposed deep neural based heart disease diagnosis model is evaluated and result reveals that an accuracy of 83.67% is achieved with the model. The accuracy of the deep neural network appears to be acceptable for heart disease diagnosis although there is still larger scope for improving the accuracy to higher level for better classification result. A comparative study on the performance of logistic regression and artificial neural network is conducted in [11]. The result of performance comparison shows that logistic regression model performed better as compared to the artificial neural network. Moreover, performance analysis with accuracy metric shows that the predictive accuracy of logistic regression model is 87.6% for heart disease classification. In [12] applied K-nearest neighbor, artificial neural network, support vector machine, decision tree and logistic regression to heart disease dataset collected from e Cleveland heart disease data repository, and proposed a model to classify heart disease dataset. The experimental result on the performance test appears to prove that the highest classification accuracy is achieved with logistic regression (84% accuracy) as compared to the other algorithms employed in the study. The decision tree algorithm for heart disease dataset classification achieves the least accuracy (74%) as compared to the other algorithms.

3. Research method

This section discusses the dataset, the statistical methodology and the approaches for parameter tuning employed in this study. We have employed heart disease dataset collected from kaggle data repository for training a decision tree algorithm. The optimal feature selection employed the χ 2 (chi-squared) statistical methodology, for testing the number of features that could produce the highest possible accuracy for the decision tree algorithm for heart disease classification. For tuning the parameters, we employed a brute force approach; we conducted a test on decision tree algorithm with different parameters (Gini and Entropy) and then the effect on the performance of decision tree algorithm is analyzed for different criterion.

Algorithm for heart disease classification

Input: dataset (feature and the class label): **output**: class label.

Step 1. Split the dataset into training set and testing set {X_1←training, y_1←training and X_2←testing, y_2←testing}

Step 2. Transform the training set with χ^2 (chi-squared) to obtain optimal number features

Step 3. Train the decision tree on the transformed training set

Step 4. Fit the decision tree model on the X_1

Step 5. Predict X_2

Step 6. Evaluate the performance of the decision tree model based on the output in step 5.

3.1 Dataset description

The heart disease dataset used for training the decision tree consists of 1025 observations. In the dataset, 499 observations are heart disease negative and 526 are heart disease positive. The class distribution of heart disease positive (patient) and heart disease negative (healthy) observations is demonstrated in figure 1 and 13 features (described in table 1) characterize the heart disease dataset.

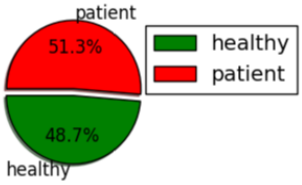

Fig. 1 Class distribution in the heart disease dataset.

As demonstrated in figure 1, the class distribution of the heart disease positive (51.3%) and heart disease negative (48.7%) classes is roughly balanced. Hence, there is no problem of imbalanced classification for the decision tree algorithm.

Table 1: heart disease dataset feature description

Feature	***Description***
Age	Age in years (Numeric, continuous value)
Sex	Sex of a person, Nominal (1=Male, 0=Female)
Chest pain (cp)	Nominal (1=typical angina, 2=atypical angina,3=non angina pain,
Total resting blood pressure	Numerical (continuous value in mm Hg on admission to the hospital)
Cholesterol (chol)	Serum cholesterol (in mg/dl)
Fasting blood sugar (fbs)	Nominal (fbs>120mg/dl (1=yes, 0=No)
Resting elector cardio graphic	Nominal (0=Normal, 1= Having ST-T wave abnormality)

Maximum heart rate achieved	Numerical (continuous value)
Exercise induced angina (exang)	Nominal (1=Yes, 0= No)
Depression induced by	Nominal (1=Yes, 0=No)
Slope of the peak exercise (slope)	Nominal (1=Up sloping, 2=Flat, 3= Down slopping)
Number of major vessels (ca)	Nominal (colored form 0 to 3)
Thalassemia (thal)	Nominal (0=Normal, 1= fixed defect, 2=reversible defect, 3-Irreversible
Target	Class label, Nominal (1=patient, 0=healthy)

3.1 Correlation model

To explore the relationship among the heart disease dataset features and the class label or target variable, we have employed Pearson correlation coefficient. Correlational analysis is important for identifying strongly correlated features to the heart disease class label or target variable. The correlation among heart disease dataset feature is demonstrated in figure 2.

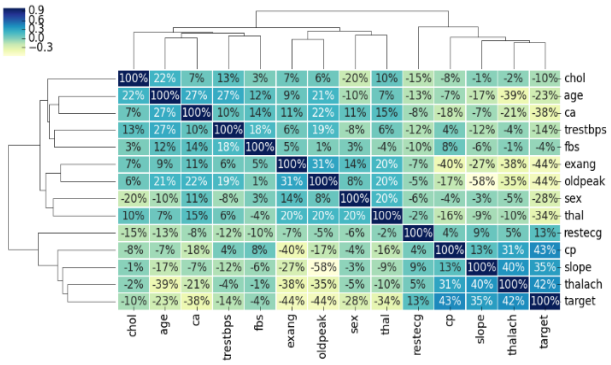

Fig. 2 Heart disease dataset feature correlation plot.

As demonstrate in figure 2, the target feature is strongly correlated with chest pain (cp), the maximum heart rate achieved (thalach) and slope of the peak exercise (slope). The heart disease dataset features such as exercise-induced angina (exang), number of major vessels (ca) and thalassemia (thal) are negatively correlated to the heart disease class label or target variable. Whether a person is suffering from heart disease or not can be predicted from the strongly correlated feature to the class label or target variable in heart disease dataset.

4. Result and discussions

The experiment focused entirely on the effectiveness of decision tree model for heart disease classification. An experimental is carried out to test the accuracy and confusion matrix of decision tree model on heart disease classification. Moreover, the performance of decision tree model is on heart disease classification with different parameters (GNI and Entropy) and varying size of dataset features. Overall, result shows that the proposed decision tree based heart disease classification model is effective in detecting the heart disease.

4.1 Correlation model

The graph shown in figure 3, shows how cross-validation accuracy changes with the number of training features selected with chi-square test.

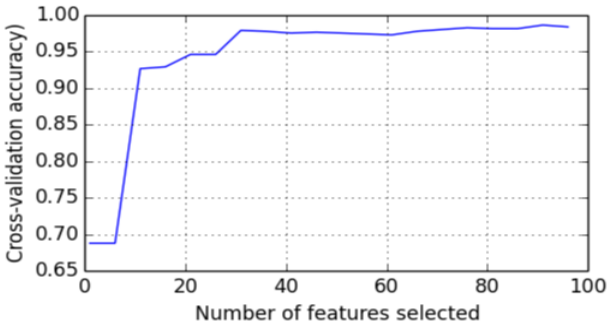

Fig. 3 Accuracy vs number of features.

We can see form figure 3, that the accuracy quickly improved when more features are used. In fact, the best accuracy is achieved with 95 of the original 1025 features (at the 20 percent percentile).

4.2 Confusion matrix

The confusion matrix is employed to test how effectively the proposed decision tree model predicts a given observation to the target class. The proposed model performs well on classification of heart disease as the number of miss-classification or classification errors are low as demonstrated in figure 4.

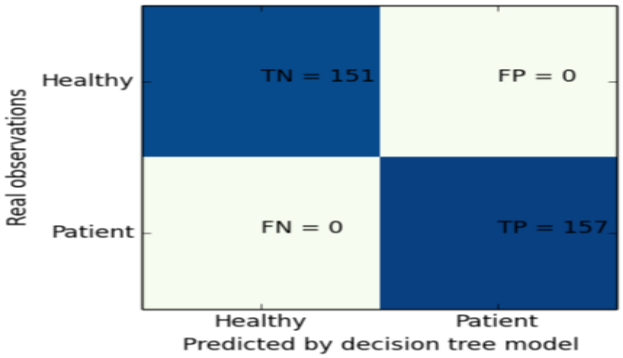

Fig. 4 Confusion matrix for the proposed model.

The predicted observations by the proposed model for 10 random experimental test are shown in figure 6. As shown in figure the model perfectly classified the first ten observations on random experimental test.

```
==========real heart disease dataset observations==========
True observations: ['not_patient' 'not_patient' 'not_patient' 'patient' 'not_patient'
'not_patient' 'not_patient' 'not_patient' 'not_patient' 'patient']
==========prediction by the proposed model==========
Predicted by model : ['not_patient' 'not_patient' 'not_patient' 'patient' 'not_patient'
'not_patient' 'not_patient' 'not_patient' 'not_patient' 'patient']
```

5. Conclusion

In this study, we employed grid search and chi squared ($\chi 2$) statistical approach for optimal feature selection. A decision tree algorithm is trained with the tuned hyper-parameter with grid search and the optimal number of feature is selected with chi-squared statistics. Moreover, the performance of decision tree model is evaluated with confusion matrix and classification accuracy. Experimental test result appears to prove that better classification accuracy is achieved with Gini parameter and 95 of the original features available in the heart disease dataset for training the decision tree algorithm. Overall, classification accuracy of 98.5% is achieved with the optimized decision tree model.

References

[1] Wan Hajarul Asikin Wan Zunaidi, RD Rohmat Saedudin, Zuraini Ali Shah, Shahreen Kasim, Choon Sen Seah, Maman Abdurohman, *Performances Analysis of Heart Disease Dataset using Different Data Mining Classifications*, International Journal on Advanced Science Engineering and Information Technology, 2018.

[2] Assegie Tsehay Admasssu, *A support vector machine based heart disease prediction*, Journal of Software Engineering & Intelligent Systems, 2019.

[3] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, *Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques*, IEEE, 2019.

[4] Moloud Abdar, Elham Nasarian, Vivi Nur Wijayaningrum, Xujuan Zhou, *Performance Improvement of Decision Trees for Diagnosis of Coronary Artery Disease Using Multi Filtering Approach*, IEEE,2019.

[5] Divya Krishnani, Anjali Kumari, Akash Dewangan, Aditya Singh, Nenavath Srinivas Naik, Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms, IEEE, 2019.

[6] Rahul Katarya, Polipireddy Srinivas, *Predicting Heart Disease at Early Stages using Machine Learning: A Survey*, IEEE, 2020.

[7] Isra'a Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azze, *A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods*, International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 12, December 2016.

[8] Noor Basha, Gopal Krishna C, Ashok Kumar P S, Venkatesh P, Early Detection of Heart Syndrome Using Machine Learning Technique, International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques, IEEE, 2019.

[9] Nikhil Gawande, Alka Barhatte, Heart Diseases Classification using Convolutional Neural Network, Proceedings of the 2nd International Conference on Communication and Electronics Systems, IEEE, 2017.

[10] Kathleen H. Miaoa, Julia H. Miao, Coronary Heart Disease Diagnosis using Deep Neural Networks, International Journal of Advanced Computer Science and Applications, Vol. 9, No. 10, 2018.

[11] Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande, Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease, International Journal of Machine Learning and Computing, Vol. 5, No. 5, October 2015.

[12] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms, Hindawi Mobile Information Systems Volume 2018.

[13] Aufzalina Mohd Yusof, Nor Azura Md. Ghani, Khairul Asri Mohd Ghani, Khairul Izan Mohd Ghani, A predictive model for prediction of heart surgery procedure, Indonesian Journal of Electrical Engineering and Computer Science Vol. 15, No. 3, September 2019.

[14] R. Chitra and Dr.V. Seenivasagam, Heart Disease Prediction System Using Supervised Learning Classifier, Bonfring International Journal of Software Engineering and Soft Computing, Vol. 3, No. 1, March 2013.

[15] V. Krishnaiah, G. Narsimha, N. Subhash Chandra, Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review, International

Journal of Computer Applications (0975 – 8887) Volume 136 – No.2, February 2016.

[16] R. Subha, K. Anandakumar, A. Bharathi, Study on Cardiovascular Disease Classification Using Machine Learning Approaches, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 6, 2016.

[17] Assegie, T.A, An optimized K-Nearest Neighbor based breast cancer detection, Journal of Robotics and Control (JRC) Volume 2, Issue 3, May 2020.

[18] Assegie, T.A, Sushma S.J, A Support Vector Machine and Decision Tree Based Breast Cancer Prediction, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-3, February, 2020.

[19] Assegie, T.A, Nair, P.S, *The Performance of Different Machine Learning Models On Diabetes Prediction*, International Journal Of Scientific & Technology Research Volume 9, Issue 01, January 2020.