Proceedings of the 3rd International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/23/20231076

Data attribution and quantitative analysis for heart disease screening

Chaozhi Geng

University of Illinois Urbana-Champaign

gengforabr@gmail.com

Abstract. Heart disease is a globally common fatal diseases, and its appearance has always been of great concern. The goal of this paper was to analyse factors of variation in the appearance of heart disease in Cleveland area with a view to provide a scientific basis for the prevention and comprehension of heart disease. This paper utilized data which includes the prevalence of heart disease in Cleveland region, which contains several characteristic variables such as age, gender, type of chest pain, etc., as well as the target variable, the prevalence of heart disease. First, the data were subjected to data pre-processing, including missing value treatment, outlier treatment, and feature engineering. Then, Granger causality test could be utilized to find out the correlation among features and target in order to distinguish the features that have a significant impact on the prevalence of heart disease. The outcome of the experiment showed that there was a major causal relationship between features such as ST depression led by oldpeak, slope of the highest exercise ST segment, a blood disorder called thalassemia (thal), and heart disease incidence. And different degrees of blood diseases may increase the risk of the disease. These results have significant implications for the progression of targeted heart disease prevention strategies and health management policies. Further studies can incorporate more features and data to enhance model's precision and accuracy and interpretability, providing more valuable information for better understanding the pathogenesis of heart disease and predicting risk.

Keywords: decision tree, heart disease, granger causality test.

1. Introduction

Based on the most recent statistics provided by the World Health Organization (WHO), there is a strong correlation between heart disease and its prevalence. Because of its impact on human health, it has received extensive attention in medical research. Through the implementation of data mining, human can get known about hidden knowledge which includes in data. Valuable relationships that exist among the data can be revealed through Data mining, and these rules can be applied to make informed decisions [1]. This paper utilizes Granger causality and decision tree to simulate the whole process of factoring heart disease.

The goal of this paper was to analyze the features of occurrence in heart disease incidence in Cleveland and to explore the main characteristics and factors that influence heart disease incidence in the region. By processing and analyzing the available data on heart disease incidence, this paper hope to identify the characteristics that are closely associated with heart disease incidence and explore in depth the causal relationship between them. This will help reveal the patterns and mechanisms of

DOI: 10.54254/2753-8818/23/20231076

changes in heart disease incidence rates in Cleveland, and provide a scientific basis for the development of targeted prevention and intervention strategies. Through in-depth analysis of the incidence of heart disease in Cleveland, people can better understand the risk factors of heart disease, predict the risk and take appropriate measures, and ultimately contribute to the improvement of heart health and the reduction of heart disease incidence.

2. Related work

Peter C. Austin and his co-authors explore cardiac dysfunctions in their study. The medical professionals involved in the study categorized patients into two groups: those "with" the disease and those "without" it. Their findings indicate that employing a decision tree in data mining yields superior outcomes compared to a regression model [2].

Jasmine Nahar and her colleagues conducted another study to investigate the correlation between heart disease and sex. The research reveals female has a lower risk of coronary heart disease compared to men. Both male and female can effectively alleviate chest pain by engaging in regular exercise [3].

In the research presented by K. Rajeswari and her co-authors, they investigate heart disease using a Neural Network. Getting into the heart of their investigation, their primary objective is to scrutinize the repercussions of feature selection on the neural network algorithm's efficacy in identifying patients afflicted with Ischemic heart disease. The paper uses a total of 12 features in their analysis. The study's results reveal that by utilizing all the features, the precision rate during the training mode reaches 89.4%, whereas during the test mode, it achieves a value of 82.2% [4].

Past studies have extensively investigated and studied the incidence and risk factors of heart disease. Factors such as age, gender, hypertension, hyperlipidemia and diabetes mellitus are widely recognized to be closely associated with the danger of progressing heart disease. In addition, lifestyle factors encompassing behaviors like smoking, alcohol consumption, dietary habits, and levels of physical activity are also recognized as important factors influencing heart disease. However, the factors contributing to the variation of heart disease incidence in the Cleveland area have not been fully explored and analyzed, and the wide variation in demographic characteristics and lifestyles in the area may have an impact on the findings, so further research is needed to gain a deeper understanding.

3. Method

3.1. Granger causality

Granger causality stands as a method wielded to discern causal relationships amid time series data, premised on the fundamental idea that if the past values of one time series exhibit the potential to predict the current values of another time series, people can assert that the former has effectively "caused" the latter [5].

The principle of implementation is based on the autoregressive model. Autoregressive models use past observations to predict current values. By comparing a model that encompasses the past values of only one time series with a model that includes the past values of both time series, people can test for causality. If the model's predictive performance improves with the addition of a second time series, this indicates that there may be a causal relationship between them [6].

The Granger causality test reveals causal relationships between time series data, showing how past values of one series can influence present values of another, but it cannot determine the specific causing relationship. It has important applications in forecasting and decision making [7].

The main steps implementing Granger causality is that if the past behavior of signal x aids in predicting signal y, then people can set that signal x leads to signal y. To put this idea into practice, people need to make predictions and assess their accuracy. This approach yields a practical yet less stringent interpretation of Granger causality. The concept of enhancing predictions is extended by considering conditional dependence or independence as a way to measure causality [8].

$$R_{\mathcal{F}}(B(n+1) \mid A^n) = \inf_{f \in \mathcal{F}} E\left[g\left(x_B(n+1) - f(x_A^n)\right)\right] \quad (1)$$

DOI: 10.54254/2753-8818/23/20231076

3.2. Decision tree

Decision tree is a commonly used machine learning algorithm for solving problems which mainly consist of classification and regression. Its idea is to simulate the human thinking process when making decisions, through a series of judgments and conditional branches to gradually classify or predict the target value of the sample [9].

Decision trees are a robust technique widely applied in diverse domains, including, image processing, and pattern recognition [10]. They are sequential models that efficiently and cohesively combine a series of basic tests. Creating conceptual rules for decision trees is far simpler compared to establishing the numerical weights in neural networks' connections between nodes [11]. The principle of realization comprises 4 steps: Build a decision tree, Feature selection, Stopping conditions, Handling missing values, Pruning [12].

Information gain, also referred to as mutual information, is a metric commonly employed for segmentation. It provides an intuitive measure of how much knowledge can be gained from knowing the value of a random variable [8]. Essentially, it represents the opposite of entropy, meaning that a higher information gain indicates better predictive power. The data gain G(S, A) is defined based on the concept of entropy, as shown in "equation (2)" [13].

$$Gain(\mathrm{S,A}) = \sum_{\mathrm{v}\in\mathrm{V(A)}} \frac{|\mathrm{S_v}|}{|\mathrm{S}|} Entropy(\mathrm{S_V}) \tag{2}$$

The merits of implementing decision trees are that they can handle various types of data (numerical and discrete), and are suitable for multiclassification and regression problems [9]. However, decision trees also have limitations, are prone to overfitting problems, and do not model certain complex relationships well [14].To address these issues, integrated learning methods such as random forests can be used to improve the performance of decision trees [15].

4. Experiment

The dataset this paper used is called the UCI Heart Disease Dataset. it was collected from four locations: the Advanced Clinic of the Hungarian Heart Institute in Budapest, the V.A. Medical Center in Long Beach, California, and the University Hospital in Zurich, Switzerland. Each database has the same instance format. Out of the 76 raw attributes present in the databases, only 14 are actively utilized. These 14 features has been described in Table 1.

Table 1. Features in dataset.

feature	value
age	All integer
sex	1: male, 0: female
Cp- types of chest pain	1: typical angina, 2: atypical angina, 3: non-angina, 4: asymptomatic
trestbps	Resting blood pressure on admission (Unit: mm/Hg)
chol	Serum cholesterol level (Unit: mg/dl)
fbs- fasting blood glucose (>120 mg/dl)	1: true, 0: false
restecg- resting normal	1: with ST-T wave abnormalities 2: showing possible or definite left ventricle
thalach	Maximum heart rate achieved
exang- exercise-induced angina	1: yes, 0: no
oldpeak	ST depression caused by exercise relative to rest
slope- slope of the highest motion ST segment	1: uphill, 2: flat, 3: downhill
ca- number of major blood vessels with fluorescent	0-3
thal- a blood disorder called thalassemia	0: normal, 1: fixed defect, 2: reversible defect
target: whether the patient has heart disease	1: presence, 0: absense

DOI: 10.54254/2753-8818/23/20231076

4.1. Analysis of granger causality

Firstly, this paper constructed a matrix containing all data features related to heart disease, excluding the target variable. Subsequently, the Select the best method was employed to iteratively identify the features most strongly correlated with the ultimate target variable. Then, this study set the maximum lag coefficient to 50 and used this feature matrix to repeatedly simulate the relationship with the target variable. Finally, this study determined Granger causality between two variables based on the obtained P-value results from each simulation.

4.2. Training of decision tree

Subsequently, within the decision tree algorithm, this paper established a mapping called "y_class," where zero means person do not have heart disease, and one means person do get heart disease. Then, this study performed decision training by combining the feature matrix and the target variable "target." The test size was set to 0.2. Following this, a decision tree classifier object was created, and the model was trained using the fit method. Utilizing the x_test dataset, this paper obtained predictions for the target variable, resulting in an optimal solution.

5. Result

5.1. Granger causality

This study determined the association between each feature and the target variable (presence or absence of heart disease) by examining the corresponding P-values obtained from each simulation of Granger causality. For each feature, this paper conducted 50 simulations and set the significance level (α) to 0.05. If the P-value exceeded 0.05, it indicated that the particular feature had limited relevance to the target variable.

However, with each iteration of the simulation, the P-values tended to increase. To address this, this study applied a filtering approach based on the proportion of results with P-values less than 0.05 among the 50 simulations to identify the most meaningful associations. As the result of Table 2, the count of p-value which is less than 0.05 has been represented, which means that the relevance of feature is closer to getting a heart disease when the count is bigger.

Table 2. Distribution of P-values.

	age	sex	cp	thalach	exang	oldpeak	slope	ca	thal
p-value <= 0.05	9/50	9/50	22/50	18/50	20/50	50/50	32/50	10/50	50/50

Thus, the Granger causality assists to rule out some of roughly related factors, and only takes 3 features into account, which are oldpeak, thal and slope.

5.2. Decision tree

To better obtain a quantitative categorization index to provide data to determine the basis for future self-diagnosis. To provide data support for the health service administration (CDC), focusing on the population, as well as to propose policies for the prevention of cardiac deaths. This study performed a quantitative analysis of the data after imputation using the decision tree method, and after a series of experiments, obtained a classification decision tree model with approximately 72% accuracy. And the decision rules of the decision tree model are visualized. As shown in Figure 1.

DOI: 10.54254/2753-8818/23/20231076

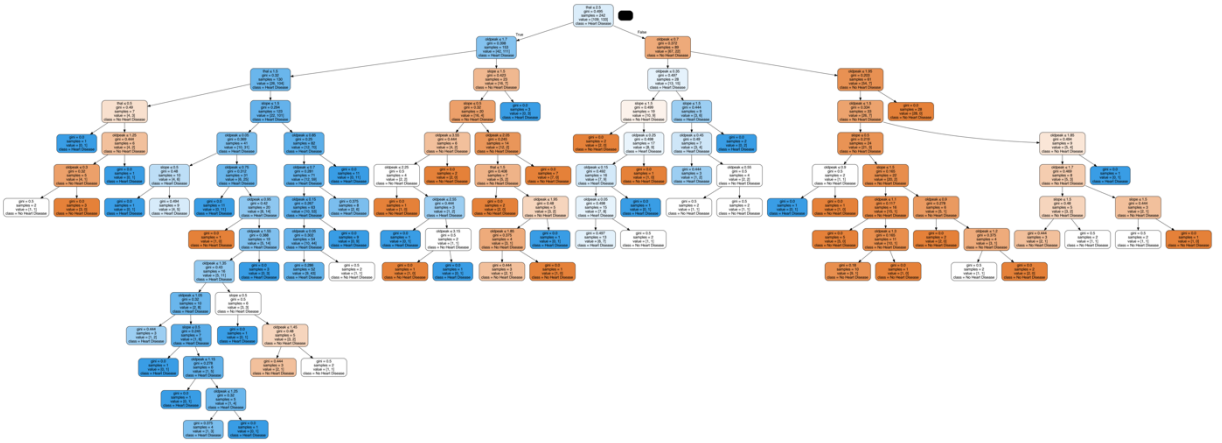

Figure 1. The visualization of decision tree.

References

[1] Abdar, Moloud, et al. "Comparing performance of data mining algorithms in prediction heart diseases." (2015).

[2] Peter C. Austin, Jack V. Tu, Jennifer E. Ho, Daniel Levy, Douglas S. Lee. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology, 2013; 66(4): 398-407.

[3] Jesmin N, Tasadduq I, Kevin ST, Yi-Ping Ph Ch. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Application, 2013; 40(4): 1086-1093.

[4] K.Rajeswari, V.Vaithiyanathan, T.R. Neelakantan. Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks. International Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012), Procedia Engineering, 2012; 41: 1818-1823.

[5] Freeman, John R. "Granger causality and the times series analysis of political relationships." American Journal of Political Science (1983): 327-358.

[6] Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." Shanghai archives of psychiatry 27.2 (2015): 130.

[7] Diks, Cees, and Valentyn Panchenko. "A new statistic and practical guidelines for nonparametric Granger causality testing." Journal of Economic Dynamics and Control 30.9-10 (2006): 1647-1669.

[8] Amblard, Pierre-Olivier, and Olivier JJ Michel. "The relation between Granger causality and directed information theory: A review." Entropy 15.1 (2012): 113-143.

[9] Magee, John F. Decision trees for decision making. Brighton, MA, USA: Harvard Business Review, 1964.

[10] Seth, Anil. "Granger causality." Scholarpedia 2.7 (2007): 1667.

[11] Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." Journal of Applied Science and Technology Trends 2.01 (2021): 20-28

[12] Kotsiantis, Sotiris B. "Decision trees: a recent overview." Artificial Intelligence Review 39 (2013): 261-283.

[13] Wehrl, Alfred. "General properties of entropy." Reviews of Modern Physics 50.2 (1978): 221.

[14] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics 21.3 (1991): 660-674.

[15] Quinlan, J. Ross. "Learning decision tree classifiers." ACM Computing Surveys (CSUR) 28.1 (1996): 71-72.