Middle Black Sea Journal of Health Science /
Mid Blac Sea J Health Sci

December 2021; 7(3):360-365

RESEARCH ARTICLE

DOI: 10.19127/mbsjohs.1004917

Determination of Factors Related to Coronary Heart Diseases by Associative Classification Technique

Zeynep Kucukakcali[1](ID) Ipek Balikci Cicek[1](ID)

[1]Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya,

Received: 05 October 2021, Accepted: 13 December 2021, Published online: 31 Decemder 2021
© Ordu University Institute of Health Sciences, Turkey, 2021

Abstract

Objective: The goal of this study is to categorize CHD using the relational classification approach on a CHD dataset made up of open access patients with and without CHD, as well as to disclose the disease's relationship to the risk factors that cause CHD.
Methods: The associative classification model was applied to the open-access data set “CHD” in this study. The performance of the model was evaluated by accuracy, specificity, negative predictive value. According to the results of the associative classification model, the factors associated with the disease were determined by specific rules. groups, examined using Mann-Whitney U, Pearson Chi-square test, and Fisher's Exact test. p<0.05 values were considered statistically significant.
Results: For the associative classification model applied to the data set, the results of the performance metrics that specificity, accuracy, and negative predictive value were calculated as 0.995, 0.852, 0.854, respectively.
Conclusion: The conclusions of this investigation revealed that the study conducted on the CHD data set with the associative classification model yielded successful results. Since the results obtained from the associative classification model reveal certain rules, it is very easy for users to understand and the results can be easily interpreted. Thus, the findings obtained with this model can be used quite easily in preventive medicine practices.

Key words: CHD, classification, association rules, associative classification.

Suggested Citation Kucukakcali Z, Balikci Çicek I. Determination of Factors Related To Coronary Heart Diseases By Associative Classification Technique. Mid Blac Sea Journal of Health Science, 2021; 7(3):360-365

Address for correspondence/reprints:

İpek Balikci Cicek

Telephone number: +90 (422) 341 06 60-1337

E-mail: ipek.balikci@inonu.edu.tr

Introduction

In parallel with the developments in science, medicine, and technology in the world, life expectancy increases with the increase in the level of perception of disease/health. With the prolongation of life expectancy, the prevalence of chronic diseases is also increasing (1, 2). According to the 2008 data of the World Health Organization, 63% of deaths in the world are caused by chronic diseases. Among the causes of death due to chronic diseases, cardiovascular diseases (48%) take first place with the highest rate (3). Coronary heart diseases (CHD), which is among the cardiovascular diseases, is the most important cause of morbidity and mortality in the World (4). Coronary heart disease (CHD), which is the most common cause of death due to cardiovascular diseases, is a progressive, systemic and inflammatory disease in which atherosclerosis plays a role in its etiology and can cause clinical events ranging from asymptomatic to acute myocardial infarction and sudden death. Often atherosclerotic plaques cause narrowing of the coronary artery (5). It presents as acute myocardial infarction or sudden cardiac death without symptoms in a significant part of the patients. Therefore, it is very important to know, prevent, and diagnose the risk factors of CHD (6).

Simply put, data mining is the process of finding usable information concealed in large amounts of data (7). With tools from several disciplines such as artificial intelligence, machine learning, statistics, and optimization, data mining allows researchers to make effective and well-informed conclusions. It also allows for the discovery of hidden, implicit, beneficial links, patterns, relations, or trends that would be difficult to uncover using traditional method (8).

Under the associative analysis model, which is one of the data mining models, there is an association rules model. Because of their simplicity and utility, association rules are commonly employed in data mining. Association rules are used while doing this analysis to express the occurrence of events along with their probabilities. The association rules' aim is to give relationships and associations as rules (9, 10) .

When developing a model, the associative classification uses the logic of merging the classification and association rule models, which are two data mining methodologies. Classification models are generated using the set of rules obtained from association rule analysis in associative classification. The response/target variable being on the right side of the obtained rule makes it easier to understand and interpret in the associative classification approach (11).

This study, it is aimed to classify CHD by applying the relational classification method on the CHD dataset consisting of open access patients with and without CHD, and to reveal the disease relationship with the risk factors that cause CHD.

Methods

Dataset

The associative classification model, a data mining method that combines classification and association rules methodologies, was used in the study to analyze an open-access data set called "CHD." The open-access data set "Cardiovascular Study Dataset" was obtained from the address https://www.kaggle.com/christofel04/cardiovascular-study-dataset-predict-heart-disea.

Association Rules

One of the data mining method is association rules, which use probabilistic expressions to explain the presence of certain occurrences in the database (12). Association rules that are unsupervised data mining methods are used to find hidden links in huge data sets. Potential data relationships can be defined by association rules. The goal is to uncover the rules that govern the occurrence of occurrences that are likely to occur at the same time. A series of operations are applied in bulk to the records in the databases, and the rules explaining the link between the records are derived using this method (13).

Associative Classification

Associative classification is a novel supervised learning approach that seeks to predict scenarios that haven't been encountered before. An associative classification, in particular, is a method for creating classification models that employs rules derived from association rules. Associative classification combines classification and association rule mining to can produce give more accurate results than other data mining classification techniques. Only the class/response / dependent variable categories make up the right side of association rules in associative classification. The rules of the association are derived using if-then clauses, which are precursor-successor clauses. Therefore, the user will have an easier time understanding and interpreting the results. As a result of this circumstance, associative classification is more advantageous than traditional classification methods (11).

Associative classification uses and develops a variety of algorithms. In this study, the classification based on the association rules (CBA) method was applied.

Statistical analysis

Quantitative data are summarized by median (minimum-maximum) and qualitative variables are given by number and percentage. Normal distribution was evaluated with the Kolmogorov-Smirnov test. In terms of input variables, the existence of a statistically significant difference and relationship between the categories of the output variable, " 10-year risk of coronary heart disease (yes) " and " 10 year risk of coronary heart disease (no) " groups, was examined using Mann-Whitney U, Pearson Chi-square test, and Fisher's Exact test. p<0.05 values were considered statistically significant. In all analyzes, IBM SPSS Statistics 26.0 for the Windows package program was used.

Results

The table showing the distribution of the dependent variable in the data set used in this study is given below (Table 1).

Descriptive statistics of the independent variables in this study are given in Table 2. According to this table; There is a statistically significant difference between the groups of the dependent variable (TenYearCHD) in terms of age, tot Chol, sys BP, dia BP, BMI, glucose and cigs Per Day variables (p<0.05).

Table 1. Table showing the distribution of the dependent variable

No		Yes	
Count	Percentage (%)	Count	Percentage (%)
2879	84.9	511	15.1

However, there was no statistically significant difference between the groups of the dependent variable (TenYearCHD) in terms of the heart Rate variable (p>0.05).

Table 3 shows that; there is a statistically significant relationship between the sex, education, is smoking, BP Meds, prevalent Stroke, prevalent Hyp and diabetes variables and the dependent variable (TenYearCHD) groups (p<0.05).

Table 4 shows the classification matrix for the associative classification model that was used to classify the Cardiovascular Work Dataset in this study

Table 2. Descriptive statistics table of quantitative independent variables

Variables	TenYearCHD (have 10 year risk of coronary heart disease (CHD or not))		p-value[*]
	No	Yes	
	Median (min-max)	Median (min-max)	
Age	48 (32-70)	55 (35-70)	**<0.001**
Tot Chol	232 (113-696)	243 (107-600)	**<0.001**
Sys BP	127 (83.5-243)	139 (83.5-295)	**<0.001**
Dia BP	81 (50-142.5)	85 (48-135)	**<0.001**
BMI	25.23 (16.48-51.28)	26.19 (15.96-56.8)	**<0.001**
Glucose	78 (40-386)	80 (45-394)	**0.001**
Cigs Per Day	0 (0-70)	4 (0-60)	**0.001**
Heart Rate	75 (45-143)	75 (50-120)	0.358

*: Mann Whitney U test

Table 3. Descriptive statistics for qualitative independent variables

Variables	Categories of Variables	TenYearCHD		p-value
		No	Yes	
		Number (%)	Number (%)	
Sex	Female	1684 (58.5)	239 (46.8)	**<0.001**[*]
	Male	1195 (41.5)	272 (53.2)	
Education	1	1135 (40.5)	256 (51.4)	**<0.001**[*]
	2	872 (31.1)	118 (23.7)	
	3	479 (17.1)	70 (14.1)	
	4	319 (11.4)	54 (10.8)	
is smoking	No smoke	1467 (51.0)	236 (46.2)	**0.047**[*]
	Yes smoke	1412 (49.0)	275 (53.8)	
BP Meds	No Meds	2775 (97.6)	471 (93.5)	**<0.001**[*]
	Yes Meds	67 (2.4)	33 (6.5)	
Prevalent Stroke	No stroke	2867 (99.6)	501 (98.0)	**0.001**[**]
	Yes stroke	12 (0.4)	10 (2.0)	
Prevalent Hyp	No Hyp	2065 (71. 7)	256 (50.1)	**<0.001**[*]
	Yes Hyp	814 (28.3)	255 (49.9)	
Diabetes	No diabetes	2825 (98.1)	478 (93.5)	**<0.001**[*]
	Yes diabetes	54 (1.9)	33 (6.5)	

*: Pearson chi-square test, **: Fisher's Exact test

Table 4. The associative classification model's classification matrix

Prediction	Reference		
	No	Yes	Total
No	2471	421	2892
Yes	12	23	35
Total	2483	444	2927

Table 5 shows the results of the classification performance criterion for the associative classification model. The model's specificity was calculated to be 0.995, the accuracy to be 0.852, and the negative predictive value to be 0.854.

Table 5. The model's classification performance criteria's values

Metric	Value
Specificity	0.995
Accuracy	0.852
Negative predictive value	0.854

The classification algorithm's association rules are shown in Table 6. As expressed in Table 6, when age=[32,55.5), is smoking=no smoke, prevalent hyp=no hyp and glucose=[40,122) are considered, the probability of not having 10 year risk of coronary heart disease is 94.7%. Similarly, as age=[32,55.5), is smoking=no smoke, sys bp=[83.5,145) and BMI=[16,28.8) are taken into account, the probability of not having 10 year risk of coronary heart disease is 94.6%. In the same way, age=[32,55.5), is smoking=no smoke, tot chol=[113,256) and dia bp=[48,99.2) are regarded, the probability of not having 10 year risk of coronary heart disease is 94.6%. If age=[32,55.5), is smoking=no smoke, prevalent stroke=no stroke, prevalent hyp=no hyp are considered, the probability of not having 10 year risk of coronary heart disease is 94.6 %. Other rules derived from the classification based on association rules model can be interpreted in the same way as the previously described rules. (Table 6).

Table 6: The classification algorithm's association rules

Left-hand side rules	Right-hand side rules	Support	Confidence	Frequency
{Age=[32,55.5), is smoking=No smoke, Prevalent Hyp=No Hyp, Glucose=[40,122)}	{Tenyearchd=No}	0.224	0.947	657
{Age=[32,55.5), is smoking=No smoke, Sys Bp=[83.5,145), BMI=[16,28.8)}	{Tenyearchd=No}	0.205	0.946	599
{Age=[32,55.5), is smoking=No smoke, Tot Chol=[113,256), Dia Bp=[48,99.2)}	{Tenyearchd=No}	0.215	0.946	629
{Age=[32,55.5), is smoking=No smoke, Prevalent Stroke=No stroke, Prevalent Hyp=No Hyp}	{Tenyearchd=No}	0.225	0.946	660
{Age=[32,55.5), is smoking=No smoke, Prevalent Hyp=No Hyp, Diabetes=No diabetes}	{Tenyearchd=No}	0.224	0.945	656
{Age=[32,55.5), Sex=Female, Prevalent Hyp=No Hyp, Tot Chol=[113,256)}	{Tenyearchd=No}	0.231	0.943	675
{Age=[32,55.5), is smoking=No smoke, Bp Meds=No Meds, Sys Bp=[83.5,145)}	{Tenyearchd=No}	0.247	0.94	724
{Age=[32,55.5), is smoking=No smoke, Tot Chol=[113,256), Glucose=[40,122)}	{Tenyearchd=No}	0.228	0.94	668
{age=[32,55.5), Cigs Per Day=[0,17.5), Prevalent Hyp=No Hyp, Tot chol=[113,256)}	{Tenyearchd=No}	0.271	0.94	792

Discussion

Today, cardiovascular diseases are quite common and one of the leading causes of death. Although death from coronary heart disease, a cardiovascular disease in which atherosclerosis plays a role and can cause clinical events ranging from asymptomatic to acute myocardial infarction, has fallen significantly, it remains the single leading cause of death for adults worldwide. This evidence demonstrates the need to implement effective primary prevention approaches worldwide and to identify risk groups and potential areas of improvement, with the fact that mortality from CHD is expected to continue to rise in developing countries. For this reason, accurate and timely diagnosis of coronary heart disease is very important in terms of treatment and reducing mortality rates (14, 15). Therefore, it is very important to determine the factors associated with the disease.

Association rules are methods for analyzing the coexistence of events, and they are one of the descriptive models in data mining. These connections are based on the coexistence of data items and express the co-occurrence of occurrences as well as certain possibilities. One of the most basic approaches of machine learning is classification analysis, which is employed by a vast scientific community (16). Classification is a rule-based estimate procedure that assigns each observation in a dataset to one of several specified classes. Associative classification combines two common data mining approaches, association rules, and classification methods, to provide categorization. In associative classification, association rules methods have been effectively employed to construct proper classifiers in recent years (17,18). Furthermore, when applied to medical data sets, associative classification stands out as a novel approach that makes it easier for users to interpret (19,20).

An open-access CHD data set was used in this work to test an associative classification model. In this context, the associative classification model was used to estimate distinct factors (explanatory variables) that may be connected with CHD (the dependent variable), and rules were established. According to the findings, the accuracy, specificity, and negative predictive value derived from the associative classification model were 85.20 %, 99.50%, and 85.40 %, respectively.

As a result, the associative classification model utilized in the study with the CHD data set yielded successful findings. Furthermore, this model has yielded specific disease-related criteria that might be applied in preventative medicine practices.

Ethics Committee Approval: This study does not require ethical approval because the open-source data set is used

Peer-review: Externally peer-reviewed.

Author Contributions: Concept: I.B, Z.K Design: I.B, Z.K Literature Search: I.B, Z.K Data Collection and Processing: I.B, Z.K Analysis or Interpretation: I.B, Z.K Writing: Z.K

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study hasn’t received no financial support.

References

1. Gunes Z. Social Support and States of Hopelessness Perceived by Individuals with Chronic Diseases from the Family. Florence Nightingale Journal of Nursing. 2009;17(1):24-31.
2. Kilic M. Primary approach to the prevention of chronic diseases: Screening tests. Turkish Journal of Family Practice/ Turkish Journal of Family Medicine. 2011;15(2).
3. Organization WH. Noncommunicable diseases country profiles 2018. 2018.
4. Durusoy E, Yildirim T, Altun A. Outpatient follow-up of coronary artery disease. Trakya Univ Medical Faculty Journal. 2010;27(1):13-8.
5. Cihan S, Karabulut B, Arslan G, Cihan G. Examination of the risk of coronary artery disease with data mining methods. International Journal of Engineering Research and Development. 2018;10(1):85-93.
6. Oguz S, Cesur K, Koc S. Coronary Heart Disease Risk Factors in the Determination of Nursing Students. Turk Soc Cardiol Turkish Journal of Cardiovascular Nursing. 2011;2(2):18-21
7. Silahtaroglu G. Basic Data Mining with Concepts and Algorithms Papatya Publishing Education Inc. Istanbul, Turkey. 2008.
8. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record. 2002;31(1):76-7.
9. Chen Y-L, Chen J-M, Tung C-W. A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales. Decision support systems. 2006;42(3):1503-20.
10. Vinodh S, Prakash NH, Selvan KE. Evaluation of leanness using fuzzy association rules mining. The International Journal of Advanced Manufacturing Technology. 2011;57(1-4):343-52.
11. Thabtah FA. A review of associative classification mining. Knowledge Engineering Review. 2007;22(1):37-65.
12. Kumar AS, Wahidabanu R, editors. A frequent item graph approach for discovering frequent itemsets. 2008 International Conference on Advanced Computer Theory and Engineering; 2008: IEEE.
13. Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining1996: American Association for Artificial Intelligence.
14. Allender S, Peto V, Scarborough P, Boxer A, Rayner M. Coronary heart disease statistics. 2007.
15. Sanchis-Gomar F, Perez-Quilis C, Leischik R, Lucia A. Epidemiology of coronary heart disease and acute coronary syndrome. Annals of translational medicine. 2016;4(13).
16. Percin I, Yagin FH, Guldogan E, Yologlu S, editors. ARM: An Interactive Web Software for Association Rules Mining and an Application in Medicine. 2019 International Artificial Intelligence and Data Processing Symposium (IDAP); 2019: IEEE.

17. Thabtah F. A review of associative classification mining. The Knowledge Engineering Review. 2007;22(1):37-65.

18. Jabbar MA, Deekshatulu BL, Chandra P, editors. Heart disease prediction using lazy associative classification. 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s); 2013: IEEE.

19. Park H-Y, Lim D-J. A design failure pre-alarming system using score-and vote-based associative classification. Expert Systems with Applications. 2021;164:113950.

20. Yao X, Pei X, Yang Y, Zhang H, Xia M, Huang R, et al. Distribution of diabetic retinopathy in diabetes mellitus patients and its association rules with other eye diseases. Scientific Reports. 2021;11(1):1-10.